Validation of biophysical models: issues and methodologies. A review

Gianni Bellocchi, Mike Rivington, Marcello Donatelli, Keith Matthews

Agron. Sustain. Dev. 30 (2010) 109–130
© INRA, EDP Sciences, 2009
DOI: 10.1051/agro/2009001

Available online at:
www.agronomy-journal.org

Review article

Validation of biophysical models: issues and methodologies. A review

Gianni BELLOCCHI[1*,**], Mike RIVINGTON[2], Marcello DONATELLI[1***], Keith MATTHEWS[2]

[1] Agriculture Research Council, via di Corticella 133, 40128 Bologna, Italy
[2] Macaulay Institute, Craigiebuckler AB15 8QH, Aberdeen, UK

(Accepted 20 January 2009)

Abstract – The potential of mathematical models is widely acknowledged for examining components and interactions of natural systems, estimating the changes and uncertainties on outcomes, and fostering communication between scientists with different backgrounds and between scientists, managers and the community. For favourable reception of models, a systematic accrual of a good knowledge base is crucial for both science and decision-making. As the roles of models grow in importance, there is an increase in the need for appropriate methods with which to test their quality and performance. For biophysical models, the heterogeneity of data and the range of factors influencing usefulness of their outputs often make it difficult for full analysis and assessment. As a result, modelling studies in the domain of natural sciences often lack elements of good modelling practice related to model validation, that is correspondence of models to its intended purpose. Here we review validation issues and methods currently available for assessing the quality of biophysical models. The review covers issues of validation purpose, the robustness of model results, data quality, model prediction and model complexity. The importance of assessing input data quality and interpretation of phenomena is also addressed. Details are then provided on the range of measures commonly used for validation. Requirements for a methodology for assessment during the entire model-cycle are synthesised. Examples are used from a variety of modelling studies which mainly include agronomic modelling, e.g. crop growth and development, climatic modelling, e.g. climate scenarios, and hydrological modelling, e.g. soil hydrology, but the principles are essentially applicable to any area. It is shown that conducting detailed validation requires multi-faceted knowledge, and poses substantial scientific and technical challenges. Special emphasis is placed on using combined multiple statistics to expand our horizons in validation whilst also tailoring the validation requirements to the specific objectives of the application.

accuracy / modelling / multiple statistics / validation

Contents

[*] Corresponding author: g.bellocchi@isci.it

[**] Current affiliation: European Commission Joint Research Centre, Institute for Health and Consumer Protection, Biotechnology and GMOs Unit, via E. Fermi 2749, 21027 Ispra (VA), Italy.

[***] Currently seconded to the European Commission Joint Research Centre, Institute for the Protection and Security of the Citizen, Agriculture Unit, AGRI4CAST Action, via E. Fermi 2749, 21027, Ispra (VA), Italy.

Article published by EDP Sciences

Abbreviations

AIC	Akaike's information criterion
BIC	Bayesian information criterion
CD	Coefficient of determination
C_p	Mallows' statistic
CRM	Coefficient of residual mass
d	Willmott's index of agreement
D	Kolmogorov-Smirnov's statistic
E	Mean relative error
EF	Modelling efficiency
EF1	Modified modelling efficiency
Fa_2	Factor of two
FB	Fractional bias
E_f	Fractional gross error
LC	Lack of correlation
LCS	Lack of positive correlation weighted by the standard deviations
LOFIT	Lack of statistical fit
MAE	Mean absolute error
MaxE	Maximum error
MB	Mean bias
MBE	Mean bias error
MdAE	Median absolute error
MG	Geometric mean bias
MSE	Mean square error
NMSE	Normalized mean square error
NU	Non-unity slope
PI	Range-based pattern Index
PI-F	F-based pattern index
PI_{doy}	Range-based pattern index versus day of year
PI_{Tmin}	Range-based pattern index versus minimum air temperature
P(t)	Student's t-test probability
r	Pearson's correlation coefficient
r^2	Least-square regression coefficient of determination
REF	Relative modelling efficiency
RMA	Reduced major axis
RMdAE	Relative median absolute error
RMSE	Root mean square error
ROC	Receiver-operator characteristic curve
RRMSE	Relative root mean square error
RMSV	Root mean square variation
SB	Simulation bias
SDSD	Square differences of the standard deviation
U	Theil's inequality coefficient
U_B	Systematic error proportion of Theil's inequality coefficient
U_S	Variance proportion of Theil's inequality coefficient
U_C	Covariance proportion of Theil's inequality coefficient
VG	Geometric mean variance

1. INTRODUCTION

The mathematical modelling of natural processes has undergone a large development during the last decades and, due to the complexity of the processes involved, this development is expected to pursue for a long time. The development of quantitative models to support the description of natural and semi-natural systems and decision-making in natural resource management is indeed considered to be of high priority. This is because models have a multitude of uses for scientists, managers and policy-makers investigating and governing natural processes. A major strength of models is in exploring interactions and feedback (e.g. Wainwright and Mulligan, 2004), helping to identify uncertainties and areas were we lack knowledge. They are also important supportive tools for the communication of complex issues to stakeholders of a non-scientific background. It is therefore important to demonstrate that a model has been tested using the most appropriate methods in order to achieve credibility with users of the estimates the model makes.

A mathematical model is, by definition, an approximate reconstruction of actual phenomena and an integration of natural processes into mathematical formulae. For agricultural and ecological systems and resource use (climate, land, vegetation…), a multitude of different theories coexist, not only amongst disciplines (plant physiology, hydrology, climatology…), but also within disciplines (Argent, 2004; Beven, 2007; Arnold et al., 2008). Ecological, soil, meteorological and hydrological conditions in the actual systems are indeed the product of multiple concurrent processes, where multiple factors interact at different scales, each examined by different disciplines (Parker et al., 2003). This produces an abundance of theories and alternative explanations and, consequently, alternative models. One of the challenges is the process of bringing data and models together. It is required that numerical models should be preceded by thorough evaluation before use in practical applications, because the approximations used for the synthesis of a model often lead to discrepancies and deviations of the model results from nature.

Model evaluation is an essential step in the modelling process because it indicates if the implementation of the calculations involved reproduces the conceptual model of the system to be simulated (model reliability) and the level of accuracy of the model in reproducing the actual system (model usefulness) (Huth and Holzworth, 2005). Model evaluation includes any action in which the quality of a mathematical model is established (e.g., Metselaar, 1999; Jakeman et al., 2006). The topic of model evaluation has long attracted considerable debate amongst members of the scientific community. Much debate

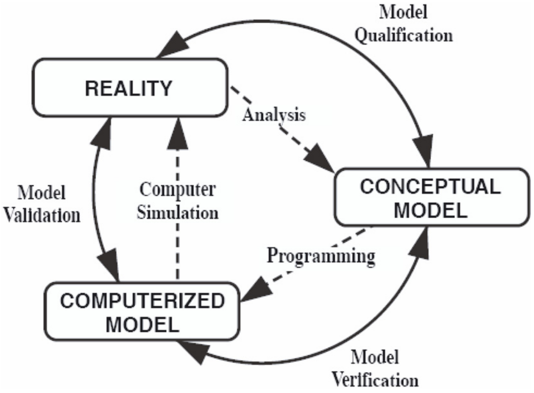


Figure 1. Phases of modelling and simulation and the role of validation and verification (after Schlesinger, 1979).

has stressed over the meaning of terms such as "testing", "validation", "verification" and "calibration" as part of the process collectively referred to as "evaluation" (Prisley and Mortimer, 2004).

The procedures to perform the evaluation task are also not widely accepted (Cheng et al., 1991) and appear in several forms, depending on data availability, system characteristics and researchers' opinion (Hsu et al., 1999). Biophysical, process-based models (unlike statistical models) are made up of mixtures of rate equations, comprise approaches with different levels of empiricism, aim at simulating systems which show a non-linear behaviour and often require numerical rather than analytical solutions. Figure 1 identifies two types of models: a conceptual model and a computerized model. The conceptual model is composed of information (input data, parameters and equations) that describes the physical system or process of interest. The computer program includes technical issues and possible errors. In practice, the computer program is tested, rather than the mathematical model representing the system (Leffelaar et al., 2003). Subjects such as system representation (model structure) and program verification play a role besides numerical evaluation (structural assessment and program testing, not discussed in this paper) in assessing model accuracy (Donatelli et al., 2002b). Shaeffer (1980) developed a methodological approach to evaluate models that consisted of six tasks: (a) model examination, (b) algorithm examination, (c) data evaluation, (d) sensitivity analysis, (e) validation studies, and (f) code comparison studies.

This paper focuses on task (e) of Shaeffer's methodology, where the term validation (Sage, 1987) is used for the process of comparing model outputs with measurements, although terminology is not standardized. A comprehensive list of publications regarding model validation was compiled by Hamilton (1991) but the connotation "valid" is rather controversial. Validation implies that the model is correct, whereas models (in the same way as hypotheses) can only be falsified rather than proven (Bair, 1994; Oreskes, 1998). This is why the meaning of the terms is not entirely consistent across fields with some eschewing the use of the term validation (e.g., Anderson and Bates, 2001), others noting the problems implied by the term while acknowledging it is widespread use (e.g., Oreskes and Belitz, 2001), and other distinguishing numerous kinds of validation including operational, conceptual, data, and even processes (e.g., Rykiel Jr., 1996). It is also acknowledged in this paper (and conveniently adopted) that the term validation is widely reported in the literature and generally used as a universal term to denote model assessment.

Definitions of validation in relation to computer software and modelling have changed little over the years. A summary of definitions is reported in Table I, which though not absolute, are becoming more definite over time. It is apparent that the most recent definitions tend to mirror the use of the concept in 1960s–1970s, whereas 1980s' definitions were more computer-oriented. The most recent definitions are by and large adopted by this paper. Whilst such relatively simple definitions of all the issues pertaining to validation can have their problems, they allow model community to communicate adequately enough in order to leave the semantic debate behind. It is worth noting that the understanding of validation is context dependant.

The overall model validation covers different areas, but the essence of it consists in defining criteria that will be taken into consideration in the choice of an "acceptable" model, and then testing the model performance according to those criteria. To assess the agreement between model results and observed data points, graphical plots are often made and judged qualitatively. It is acknowledged (e.g., Kuhnert et al., 2005) that if model output can be presented in appropriate formats, direct visual comparisons of models with data can yield significant insights about model performance. Statistical analysis by indices and test statistics play an important role to make comparisons reproducible, objective and quantitative. In general, however, the methodological basis for the validation of models to find the most suitable for specific applications is rudimentary, due to a lack of standardized terminology and procedures. Whilst statistical tools are easily applied for testing some empirical models, they might be of limited use with mechanistic (process-based) models whose primary aim is to represent system behaviour based on underlying principles (Berk et al., 2002). As modelling studies become more complex, models are used by parties less familiar with their peculiarities (e.g. structure, assumptions, data requirements and operability), and who may also lack understanding of the complexity of the entity being represented by the model. Some authors (e.g. Robinson and Ek, 2000) take the view that validation is the responsibility of the model user but improved accessibility to models by independent assessors and the ease with which they can be linked may increase their improper use. Hence, validation must not be seen as a one-off event or a "once-and-for-all" activity (Janssen and Heuberger, 1995), but as an on-going process to check for model compatibility to current evidence and variations (e.g. in spatial, climatic and hydrological conditions). Moreover, according to Sinclair and Seligman (2000), demonstration that model output more or less fits a set of data is a necessary but not sufficient indication of validity because model validity is rather the capability to analyzing, clarifying, and solving empirical and conceptual problems. Empirical problems in a domain are, in

Table I. Alternative definitions of model (software) validation.

Definition	Source
"It is a valid and sound model if it accomplishes what is expected of it"	Forrester (1961)
"The adequacy of the model as a mimic of the system which it is intended to represent"	Mihram (1972)
"Substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model"	Schlesinger (1979)
"Comparison of model results with numerical data independently derived from experience or observations of the environment"	American Society for Testing and Materials (1984)
"Validation means building the right system"	O'Keefe et al. (1987)
"The validation of a model package refers to the overall process of defining the range of circumstances or situations for which the package's behaviour and predictions are satisfactory"	Versar Inc. (1988)
"To determine that it [the software] performs its intended functions correctly, to ensure that it performs no unintended functions, and to measure its quality and reliability"	Wallace and Fujii (1989)
"The determination of the correctness of a model with respect to the user's needs and requirements"	National Acid Precipitation Assessment Program (1990)
"The process of testing a computer program and evaluating the results to ensure compliance with specific requirements"	Institute of Electrical and Electronics Engineers (1991)
"The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model"	American Institute of Aeronautics and Astronautics (1998)
"A process of adding strength to our belief in the predictiveness of a model by repeatedly showing that it is not blatantly wrong in specific applications"	Marcus and Elias (1998)
"Having a conclusion correctly derived from premises"	Sterman (2000)
"The process of assessing the prediction ability"	Li et al. (2003)
"Substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model"	Refsgaard and Henriksen (2004)
"Examining whether the system achieved the project's stated purpose related to helping the user(s) reach a decision(s)"	Sojda (2004)
"To gauge how well a model performs against observed field and laboratory measurements"	Huth and Holzworth (2005)
"A procedure consisting in comparing the model output with field or laboratory data to prove the model efficiency"	Dias and Lopes (2006)
"Substantiating that the behavior of the model "mimics" the behaviour of the system with sufficient accuracy so that it is impossible to distinguish the behaviors of both systems in the experimental frames"	Aumann (2008)
"A procedure consisting in verifying if the model is able to reproduce data, independently of those involved in its calibration"	Cardoso and Lopes (2008)
"The assessment of the performance of a model against an independently collected dataset"	Matthews et al. (2008)

general, about the observable world in need of explanation because not adequately solved by a model, solved in different ways by rival models, or solved/unsolved depending on the model. Conceptual problems arise when the concepts within a model appear to be logically inconsistent, vague and unclear, or circularly defined, and when the definition of some phenomenon in a model is hard to harmonize with an ordinary language or definition (e.g. Parker, 2001). This raises the issue of widening beyond numerical testing by also including stakeholder evaluation and expert interpretation through soft systems approaches (Matthews et al., 2008). Working all this out would extend much further than the scope of this paper that is principally meant to recognise the limitations of numerical testing in achieving salience, legitimacy and credibility of models. Issues are initially discussed on the difficulties encountered when performing validation tests. Secondly, a review is given on how models are currently evaluated using qualitative and quantitative statistical techniques. Details are

Table II. Key validation issues and relative modelling features.

Key validation issues	Major factors to investigate				
	Modelling objective	Model inputs	Model outputs	Model structure	Modelling conditions
Validation purpose	X		X		X
Robustness of results			X		X
Interpretation of phenomena		X	X	X	
Model comparison				X	
Model predictions	X		X		X
Model complexity		X	X	X	
Data accuracy		X	X		
Time histories			X		

then provided on recent developments in validation criteria, decomposition of statistics and on how to combine validation statistics into single indicators of model performance. Principles and rules of general application are set forth with cited examples from the literature on ecology, agronomy, soil and climate modelling.

2. ISSUES ON MODEL VALIDATION

Before discussing specific approaches to validation, it is important to recognize a number of questions that arise in trying to validate a model. The following is a discussion of the main issues and factors as summarized in Table II.

2.1. Validation purpose

There are many purposes for validation, including establishment of overall credibility in the model, assessment of how "right" or "wrong" a model is in a given application, along with production of evidence to support that a specific configuration of input data, parameter sets and model structure are appropriate for a particular application.

A model may be appropriate in one context and not in another. The validation tests employed, therefore, also have to reflect the differing contexts of model application. Similarly, there is a need for model developers to understand the things that would make a model valuable to a model user. Hence, validation purpose is somewhat related to the purpose for which the model was created and used. Feedback from validation should provide valuable information to both the developers on how their model may be improved, but also to end users who need to know how confident they can be in the quality of outputs. The type of validation to be executed depends on the objective for which the model was developed or purpose to which the output is to be used. It cannot be assumed that a model that is valid for one purpose is also valid for another (Sargent, 2001). This dependence on purpose may explain why common criteria, standard terminology and formalized protocols are missing, and why subjective judgement is included in the validation process (Hamilton, 1991; Landry and Oral, 1993; Rykiel Jr., 1996). In crop modelling, it emerged out of the work of various scientific research teams that early focus was on providing a simulation capability for scientists to use in distinct agricultural situations. In recent years, however, the models have been increasingly used for informing policy development and even for real-time information support for land-managers (e.g. Hochman et al., 2005). This change in model application has led to a change in the focus in model testing. Huth and Holzworth (2005) appeal to how a growing user-base for a model (including users making real-time decisions) can place a greater importance on the need for testing for model reliability (ability of the calculations involved to reproduce the conceptual model) than for model usefulness (ability to reflect the behaviour of actual systems).

2.2. Interpretation of phenomena

Different interpretations of the real world (Checkland, 1981) may present problems for validation. It is essential that model variables have the same definition as the actual data meant to be represented by the model itself. In simulating the developmental response of photosensitive plants, for instance, the ability to compute day-length is essential, but day-length can be defined in different ways depending upon the angle of the sun with the horizon (Forsythe et al., 1995). In modelling leaf area expansion it is not always clear if both simulation representation and data collection target the expanded part of a leaf (i.e. lamina expansion and growth) only, or account for the stem-like structure of a leaf that is attached to the stem (i.e. base and petiole expansion and growth) as well (e.g. Cornelissen et al., 2003). Measurements of plant development present a series of challenges as differences in assessing development can be due to the subjectivity of an observer or to a definition that is not unambiguously applied in the field (Hanft and Wych, 1982). Similarly, methods of determining kernel number, kernel mass, and yield can vary among researchers, which can add errors to comparisons between experimental results and simulated values (e.g. Anonymous, 2003). Such examples emphasise the importance of meta-data associated with original observations and development of model parameters (Medlyn and Jarvis, 1999).

Process-based models are also moving targets: if, for instance, plant model version 1 is considered to be incorrect, even a small change in a sub-model introduced to correct its functionality may produce a different interpretation on simulated processes (similar to the problem of "regression" in software development jargon). The reason for these unwanted changes lies in the lack of independence/wrong dependencies of parts of code, which is not completely avoidable. This aspect might go beyond a simple evaluation by once again comparing against previously acceptable results (Huth and Holzworth, 2005) and poses the need for formal model validation against observed data at each published stage of model development (Van Oijen, 2002).

2.3. Model comparison

Model comparison can be useful as a complement to model validation (Meehl et al., 2005). When two or more models are constructed for the same system or purpose, it is possible to make comparisons between them in order to select which is the best.

The Global Climate and Terrestrial Ecosystems group (GCTE) recognized in 1992 (GCTE, 1992) that there were at least 14 models of physiological processes that govern wheat growth and development. Landau et al. (1998, 1999, 2000) and Jamieson et al. (1999) reported on validation of three of such models against observed grain yields in the United Kingdom.

Diekkrüger et al. (1995) illustrated the simulation results from using simple as well as complex models against a common dataset covering the processes of water, nitrogen, plant growth and pesticide dynamics. In general, the models reproduced the measured dynamic only in part, with different response for different models. The study also made it clear that the experience of a scientist applying a model is as important as the difference between various model approaches.

When either field or reference modelled data are not available, attempts can be made to determine the proximity of one model to the other, also known as co-validation (Wright, 2001). Co-validation requires the assessment of the difference between models with respect to the values of their common output. The most likely case is that competing models do not share the same form. Typically, dissimilarly structured models not only have different inputs (including both variable and parameter sets), but they also have different levels of aggregation and capabilities in terms of modelling the actual system. In absence of actual data, the fact that several models show the same behaviour does not really give more confidence in each of them but only demonstrates that the models are capable of reproducing similar results for the observed system. This concept is often referred as equifinality (Beven, 1993; Franks et al., 1997; Beven and Freer, 2001; Medlyn et al., 2005). In general, potential "extra" capabilities of one model compared to another should not be used in co-validation. For example, nitrogen stress effects to plant growth should not be part of the comparison between two models where only one model includes nitrogen processes.

When statistical regression models (e.g. generalized linear models, generalized additive models) are compared, artificial data based on explicit theory can be used as "truth" (Austin et al., 2006).

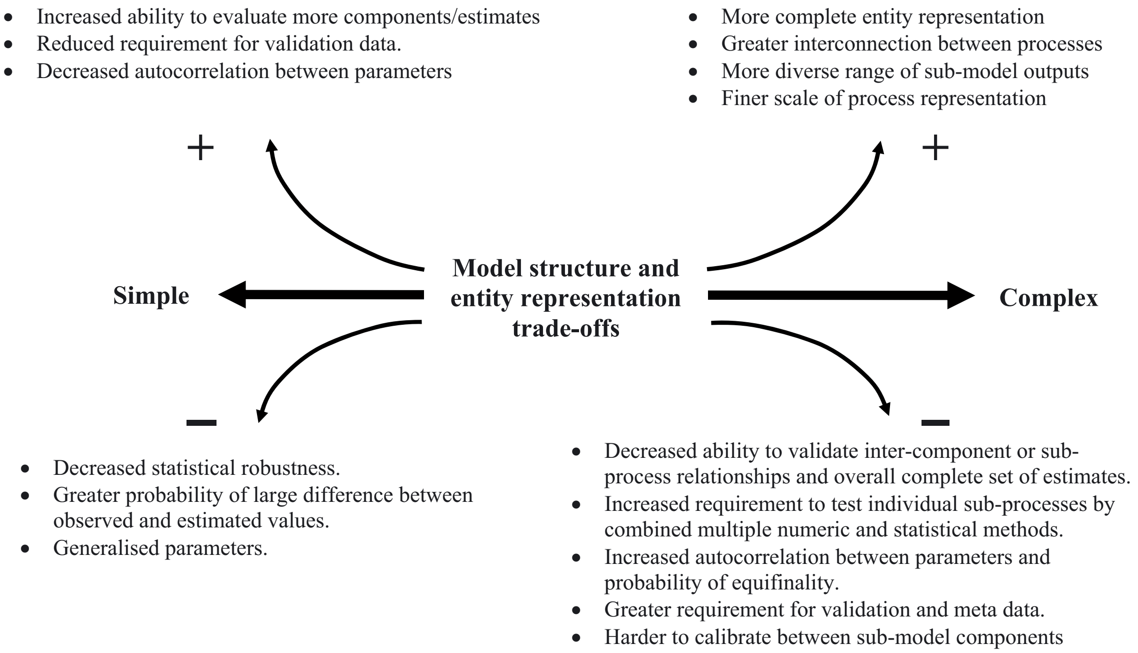
2.4. Model predictions

In model fitting, the model under investigation is evaluated for its adequacy in describing the observed data (Myung and Pitt, 2003). This is achieved by determining values for parameters in the model that will best describe the observations. Model fitting in this manner yields valuable estimates of quantities which might be used to differentiate between or explain process/system behaviours. However, papers on modelling often state that they aim to produce an instrument for prediction (Van Oijen, 2002). The issue of model prediction has been accompanied by some debate on the terminology to be used to describe model validation (see definitions by Marcus and Elias, 1998 and Li et al., 2003 in Tab. I). In this case, a fundamental issue is to quantify the degree to which a model captures an underlying reality and predicts future cases.

According to Van Oijen (2002), model-based predictions only contribute to science if the underlying model mechanisms are described, are innovative, are compared to other approaches, and if the predictions can be checked and used as a test of the model. Predictions pose special problems for testing, especially if prediction focuses on events in the far future. Predictive models can be accepted if they explain past events (ex-post validation). However, the probability of making reasonable projections decreases with the length of time looked forward. A continuous exchange of validation data among developers and test teams should either ensure a progressive validation of the models by time, or highlight the need for updated interpretations of the changed system.

The problem of global change has generated much interest in the development of predictive models for crops and ecosystems (e.g. Intergovernmental Panel on Climate Change, http://www.ipcc.ch) because model estimates are increasingly being used for decision support and strategic planning (e.g. Matthews et al., 1999; Rivington et al., 2007; Tingem et al., 2009). This requires that the quality of model estimates is assessed in advance, or that the decision support outcomes be made insensitive to the estimation uncertainty (Norton, 2003). Model quality may also show variability over geographical locations. For example, Moberg and Jones (2004), in testing hindcast estimates produced by the Hadley Centre Regional Climate Model (http://www.metoffice.com/research/hadleycentre) at 185 sites in Europe, found only some sites well represented. Responses like this restrict the geographical location to which the predicted climate change data could be used in model-based impact studies. Daily forecasts in hydrology largely use the previous week's real-time monitoring data as a training hindcast period, which permits validation and then project ahead for a period of two weeks (Quinn, 2008). In plant breeding, the issue of model-based prediction is dealt with when models are used as decision support tools to predict yield components of new cultivars and/or over new environments (Barbottin et al., 2006).

Like models which extrapolate in time, models used to extrapolate from small (e.g. leaf photosynthesis) to large spatial scales, such as regions, continents, or the global biosphere, are difficult to evaluate (e.g. Bolte et al., 2004; Chen and Coughenour, 2004). The heterogeneities in distributions of processes and non-linearity in their functional responses may make it infeasible to apply models representing these small scale processes over large areas and vice versa. There are several steps in performing spatial predictions where the variography, a well known geostatistical tool to analyse and to model anisotropic spatial correlations, is proposed to be used in the assessment of modelling results in addition to the traditional statistical analysis to demonstrate presence/absence of spatial structures in datasets (Kanevski et al., 2008).



- Increased ability to evaluate more components/estimates
- Reduced requirement for validation data.
- Decreased autocorrelation between parameters

- More complete entity representation
- Greater interconnection between processes
- More diverse range of sub-model outputs
- Finer scale of process representation

- Decreased statistical robustness.
- Greater probability of large difference between observed and estimated values.
- Generalised parameters.

- Decreased ability to validate inter-component or sub-process relationships and overall complete set of estimates.
- Increased requirement to test individual sub-processes by combined multiple numeric and statistical methods.
- Increased autocorrelation between parameters and probability of equifinality.
- Greater requirement for validation and meta data.
- Harder to calibrate between sub-model components

Figure 2. Trade-off between model complexity and validation.

2.5. Model complexity

Process-based models reflect the same complexity as the system being represented (Bolte et al., 2004) and there are authors (e.g. Pilgram et al., 2002) who emphasize the need for models to accurately reflect the dynamics of the actual system. The comparison of model results with measured data is usually done at the system level and is thus the result of both the feedbacks within and between each of the sub-models making up the whole. Since possible counter-interaction between model components may go unnoticed, ideal validation should take place both at the level of sub-models and of the complete model (Leffelaar, 1990). As modelling projects progress, the model grows and it becomes harder to validate every part of it in detail. However, the independent testing of sub-units increases the possibility of building a complex system model as the assembly of interacting processes, rather than degrading a process-based model into a fitting exercise.

Systems with small timescale, spatial extension and complexity can be isolated, thus models of these systems are more readily accessible to validation. In complex models, specific sub-processes cannot always be tested at the level of the model output (Klepper, 1989). For systems with increasing time and spatial scales with increasing complexity, validation of system models requires increasingly more effort (Fig. 2). This increase in effort can be illustrated if one considers, for instance, validation requirements in terms of data, manpower, and organizational continuity for a model of an annual crop, for a model describing a crop rotation experiment, and for a model to analyze sustainable agricultural practices. At some point the effort required to gather validation data may no longer be feasible (De Wit, 1993).

The scientific literature is full of instances where the behaviour of complex systems is reduced to a set of basic emergent behaviours. The intrinsic complexity typically captured by many biophysical models may suggest a shift in emphasis from rigorous validation to a more exploratory approach characterizing the likelihood of distinct system behaviours rather than estimation of specific outputs (Bolte et al., 2004). The benefit of more flexible approaches to validation is clearly demonstrated in their ability to highlight model behaviour in extreme circumstances that might not appear in model validation datasets, but that certainly exist in the real world in which the model will be applied. This is the case, for instance, of high-magnitude events such as weather extremes that permit a form of model validation against past events, also called a "mental model" (Favis-Mortlock et al., 2001) based on the observer memory of landscape changes, i.e. massive soil degradation (with, possibly, the support of photographic material and annotated texts). Sensibility tests (Huth and Holzworth, 2005) refer to as the comparison of model output against more subjective, local experts feeling for what the model should do. Model responses to various stimuli are evaluated against the regularly observed system responses that are often captured in the notion of "common sense". Examples from local agronomists include, for instance, statements like "under those conditions the model should have an average yield of x t ha^{-1} with a range of y to z". These tests are a way of making sure the model performs in situations where observed data for validation are not available, and their use in testing the correctness of the underlying model design should not be undervalued. The "hard" data available for model testing is indeed only a small snapshot of system behaviour, and some authors (Seibert and McDonnell, 2002 for a catchment's model; Pastres et al., 2004 for sea grass model) showed that a better process representation can be obtained by using "soft" data. This might protect against being the model "right for the wrong reasons" due to adapting a model to a small set of observations. One might become "less right" but for the right reasons.

2.6. Data accuracy and quality

The accuracy of a model is determined on one hand by the authenticity of the algorithms describing the processes of the real world, while on the other hand by the quality of both its input data and data used to evaluate its outputs. Inaccuracies are common in both inputs and measured outputs. Model validation must therefore be accompanied by critical examination of the source and nature of the data used. Random errors may occur when single, individual samples do not take into account temporal variability; when samples taken at different points do not represent the actual area of interest; or when the inputs have been modified by unnoticed factors (environmental or human). Random errors of sample collection and processing may also occur when the sample is contaminated by other environmental media, when the sample is modified during transit and storage, or when the sample data are misrecorded. Systematic errors might occur if instruments are miscalibrated, measurements are taken at inappropriate locations or seasons, no measurements or estimates are made of relevant factors, or the data are not representative of the same output as the modelled one.

Complex models must generally be tested against a broad range of data of varying types, quality, and coverage. In reality, modellers are often confronted with a poor database because data monitoring is normally limited to a few points where samples are collected and analysed at some intermittent frequency. This may also change with the variables analysed, e.g. only scattered field measurements for some variables, more complete time series data or even maps for other variables, only qualitative assessments for others again. In hydrology, for instance, consistent field data of extreme events are poor (Westrich, 2008). Due to their sparse resolution, the data are frequently subject to large errors and this is en ever-increasing problem in moving from plots or small catchments to watersheds and regions.

Weather and soil variables are key inputs for biophysical models. Common problems encountered with weather inputs are: records taken too far from the experimental site, errors under reporting values, poor maintenance and imperfect sensitivity of sensors. Soil properties are often essential input data, but soil name and classification are commonly reported in the literature for field sites based solely on location and defined soil series from published data, rather than from field measurements. The soil may only marginally meet the classification criteria and therefore have different characteristics (pH, particle size distribution, organic matter content, etc.) than a soil meeting the central trend of the stated soil, illustrating the need for an appropriate sampling strategy to capture the spatial variability (Wright et al., 2003). Inaccuracies in both weather and soil inputs may turn into combined effects, either able to strengthen or weaken each other's effect during model simulation (Fodor and Kovács, 2003) and being a source of errors in model estimates (Hoogenboom, 2000; Heinmann et al., 2002; Rivington et al., 2003, 2006). This makes it difficult to disentangle the impacts on outputs of the combined effects of error source manifesting themselves as compensating errors. Emphasis must then be to determine the uncertainties that imperfect data introduce to model estimates.

As regards model outputs, comparison of model results with observations is complicated by the fact that biophysical data are affected by errors. It is difficult to appropriately evaluate model performance if the uncertainty in the data is high. Fitting a set of data can therefore only establish legitimacy for a model under the specific circumstances for which it was evaluated (Oreskes et al., 1994). The data resource is often segregated into two groups: a sample is used to estimate model parameters (model calibration: the adjustment of model parameters to improve the fit); an independent sample is used to assess the fit of the model to the data. Validating the model against data not used to construct the model is probably the best and simplest way to assess uncertainty and derive the reliability of model estimates (e.g. Ljung, 1999; Gardner and Urban, 2003). However, biophysical data often exhibit a large degree of variability, and it is common that there are important discrepancies between model estimates and actual data despite significant calibration efforts (e.g. Gobas et al., 1998). Dedicated techniques do exist to generate reduced bodies of data (e.g. Stone, 1974; Efron, 1986; Breiman and Spector, 1992). Cross-validation is a random data split into a number (commonly in between 5 and 10) of roughly equal-sized parts; in turn, each part is used as a test set and the other parts for fitting the model. The prediction error is computed each time in the test set, and estimate of prediction error is the average of individual prediction errors. This technique is inappropriate for small sample size because the variance of the error is likely to increase considerably when splitting the data. The bootstrap technique is a valuable alternative (Wallach and Goffinet, 1989). A bootstrap sample is a sample created from random drawings with replacement from the original dataset (there can be repeats in a sample, and samples that do not contain an original point). It is also common in data-rich situations and for model comparison to generate three datasets: one to fit models, second part for estimating prediction error, third part for assessing prediction error of final selected model. As pointed out by Sinclair and Seligman (2000), a test of validity by separating data into sets may reflect only the success in splitting the data so that each group represents the same population of data. The same authors remarked: "a test of model performance is better served scientifically when based on all available data covering a wide range of conditions".

Extension of data series is also critical. Pastres et al. (2004) found that whilst traditional testing of their seagrass model gave an adequate description of the available model calibration data, the model failed to capture the known trends in sea grass evolution over a longer time frame.

Measurement uncertainty has important implications in modelling applications. Analysis of uncertainty in measured data which drive model calibration and validation improves model application and enhances decisions based on modelling results. However, the impact of uncertainty in model calibration and validation data is discussed but rarely included in the assessment of model accuracy. In order to change this omission, several goodness-of-fit indicators were modified to incorporate measurement uncertainty into model calibration and validation (Harmel et al., 2007, 2008).

2.7. Robustness of model results

Model robustness is its reliability under different sets of experimental conditions. Lack of robustness in model results may reflect an absence of explicit reference to physical processes in the construction of mathematical relationships. Concern is voiced about the fact that many model assessments are frequently limited to a small number of output variables and sets of conditions (Woodward, 2001). Well-known validation methods, whilst effective for some things, are not necessarily effective at ensuring that a model faithfully captures the underlying process driving the system in question and thus will be applicable in any new situations where a model may be employed. Simple comparisons can be misleading, for example, when only the final "end of growing season" results from a crop model are compared, because reasonable estimations of overall growth can be achieved by different pathways, not all of which are logically acceptable (Sinclair and Seligman, 2000). Performance of a model should be checked not only in terms of the major outcome of the model, but also in the estimations for critical constituent components throughout the simulation. Such checks give protection against spurious conclusions and indicate whether the model is conceptually consistent and related to reality, or specific areas require improvement. The validation of a sugarcane model by Keating et al. (1999) as cited by Sinclair and Seligman (2000), gives an example of a sound test of robustness for the following reasons: (1) a large number (i.e. 19) of data sets for sugarcane growth were used, (2) a broad range of conditions with differing locations, irrigation treatments, and nitrogen fertility treatments were explored, (3) the seasonal evolution of individual components of the model (leaf area index, green biomass, millable stalk biomass, stalk sucrose, and nitrogen accumulation) was compared with observed results. Qualified features were the discussion of the situations where relatively large deviations existed between simulated and observed results, and the presentation of hypotheses to stimulate further research to improve the understanding contained in the current model. Such model validation serves as the basis for building credibility with end users, and greatly improves the probability that the model will be accepted.

2.8. Time histories

When simulating energy transfer or mass transformation in dynamic models, a time delay/anticipation frequently occurs if estimated versus measured values are compared. Peak synchronization between estimates and measurements most often will not occur. If synchronous comparison between estimates and measurements is applied, models which produce no response with respect to a specific process can yield better results, compared to models which show a time mismatch in the response (e.g. Vichi et al., 2004). In cases where poor parameterization produces a time shift of estimates, large residuals from few points may lead one to discount the model. Model re-parameterization can help correct time-related bias. If not, apparent misalignment of observed and modelled output may lead to a re-assessment of the daily time interval (common in many systems models) as the basis for comparing modelled and monitored event data (e.g. Yagow, 1997). Average values during multi-day periods (ten-day, month or year) can be used to calculate and compare occurrence of the event. This is particularly important where timing of events (i.e. crop phenology and management synchronization) estimated by a model are used to construct recommendations or optimized practises.

2.9. Summary

In this section, we summarized the vast international experience in the validation of biophysical models and ordered issues that we regard key to model validation into a list that may add value to the modelling work. Basically, we stressed that validation is purpose-dependent, based on equivalent definition of modelled and observed phenomena, to be substantiated over a variety of conditions (robustness), and possibly run at the level of individual processes in complex models. The discussion about data quality raises the need of a system for grading the relative quality of the input and the relative importance of the variables to be fit. Concerns were addressed regarding specific aspects such as predictions in the far future and synchronization of modelled and observed peak values. Moreover, model comparison was discussed as complementary to proper validation. Put into a logical structure, the ideas we have discussed are virtually applicable to validation of any model, and could equally be consistent with modelling in a variety of fields.

3. VALIDATION OF MODELS

A range of statistical measures and visual techniques can be used to assess goodness-of-fit of a given model and to compare the performance of a suite of models, as informed by the specific context of the problem (e.g., Berk et al., 2001). Recent review papers include: Mayer and Butler (1993), Janssen and Heuberger (1995), Smith et al. (1997), Martorana and Bellocchi (1999), Yang et al. (2000), Bellocchi (2004), Tedeschi (2006), Wallach (2006). Such papers provide and discuss a range of statistical measures and visual techniques that can be used to assess goodness-of-fit of a given model and to compare the performance of a suite of models. In this section we do not replicate a detailed examination of validation techniques, rather we consider validation approaches as developed by many authors and applied in numerous modelling studies in the domain of natural sciences. It is also beyond the scope of this paper to critically appraise in detail the usefulness of each method, as details are available in the review papers detailed above. The commonly used forms of numerical and statistical approaches are highlighted here with examples of criteria for determining model acceptability. The above cited papers (and the other papers in this section dealing with specific statistics) contain the equations of the statistics and methods used for model validation. So the equations are not reported here, but the rationale behind the choice of particular sets of statistics is given.

3.1. Validation measures

There are two main categories of goodness-of-fit measures for testing of one-at-a-time output, which are: (a) residual-based; and (b) association-based. Residual-based measures such as the mean bias error (MBE), and root mean square error (RMSE) provide quantitative estimates of the deviation of modelled outcomes from measurements. On the other hand, measures of statistical association such as the Pearson's correlation coefficient (r) provide quantitative estimates of the statistical co-variation between observed and estimated values (Addiscott and Whitmore, 1987). Statistical measures aim to characterize the usefulness of a model for a specific application and may lead users to decide whether or not to use a model for that particular application. Moreover, visual comparison of modelled and measured data, and experience-based judgement on the part of the modeller have been deemed important by researchers for assessing model validity and applicability in decision making.

Mean residuals:

Mean bias (MB), the mean difference between observed and model-estimated values, is likely to be the oldest statistic to assess model accuracy (Cochran and Cox, 1957). More common is the mean square error (MSE), or equivalently its square root, the root mean square error (RMSE, or derived statistics such as the relative root mean square error RRMSE). MSE is also the statistic whose value is usually minimized during the parameter calibration process (Soroshian et al., 1993; Makowski et al., 2006). Mean absolute error (MAE) measures the mean absolute difference between observed and estimated values (Mayer and Butler, 1993), and is also used as the mean absolute percent error. Fox (1981) proposed to use: (l) RMSE or MAE to quantify the average difference between estimates and measurements; (2) the mean bias error (MBE) to identify under- or over-estimates; and (3) the variance of the distribution of the differences to quantify error variability. Such proposals are reflected in the paper of Davies and McKay (1989) and Trnka et al. (2005) for validation of solar radiation estimates.

Modelling efficiency:

Willmott (1981) developed an index of agreement (d) to be used in addition to the previous measures. The modelling efficiency statistic (EF, Nash and Sutcliffe, 1970), interpreted as the proportion of variation explained by the model, has been extensively used in plant and hydrology models (e.g. Greenwood et al., 1985; Loague and Green, 1991), and can certainly be used in biological and ecological models. Mayer and Butler (1993), likewise, indicated both RMSE and the MAE as stable statistics, and recognized modelling efficiency (EF) as an important overall measure of fit. Smith et al. (1997) pointed out that EF and a coefficient of determination (CD) should be used together for a better interpretation of RMSE when standard error of the measurements is unavailable. Alternative forms of the efficiency measures are given in Krause et al. (2005).

Correlation:

The Pearson's correlation coefficient is largely used in validation. Fox (1981) and Willmott (1982) provided strong arguments against the use of this coefficient alone as a measure of performance. Its magnitude is indeed not consistently related to the accuracy of estimates, as correlation between dissimilar estimates and measurements can be high while, conversely, small differences between estimates and measurements may occur with low correlation values. Nonparametric correlation measures are also used for model validation such as concordance, Spearman and Kendall's coefficients (Press et al., 1992; Dhanoa et al., 1999; Agresti, 2002).

Linear regression:

A linear regression between estimated and observed values is also commonly used. The hypothesis is that the regression passes through the origin and has a slope of unity (see Subsect. 3.3). The use of the r^2 regression statistic (least-squares coefficient of determination) for model performance is flawed, as it does not account for model bias (Mayer and Butler, 1993; Mitchell, 1997). Krause et al. (2005) proposed to use r^2 as a weighing factor of regression slope to quantify under- or over-estimates. An alternative (nonparametric) method to compute r^2 was proposed by Kvålseth (1985), resulting in a coefficient that is more resistant to outliers or extreme data points.

Combined graphical and statistical approaches:

The factor of two (Fa_2) is currently used to evaluate air dispersion models (e.g. Kumar, 2000); combined with the values of different indices, MBE, fractional bias (FB), normalized mean square error (NMSE), correlation coefficient (r), geometric mean bias (MG), and geometric mean variance (VG). FB in the form of absolute differences is presented as fractional gross error (E_f) in Seigneur et al. (2000).

Loague and Green (1991) suggested the use of both statistical and graphical measures for validation. Model performance can be compared using either summary statistics (mean, range, standard deviation) or using individual measured versus estimated pairs of data, which can also be displayed in both statistical and graphical forms. Assessment of data pairs usually proceeds with an analysis of the residual errors in the forms of maximum error (MaxE), RMSE, EF, CD and coefficient of residual mass (CRM) (James and Burges, 1982; Green and Stephenson, 1986). Suggested graphical displays include: (l) comparison of measurements and estimates; (2) comparison of ranges, medians and means; (3) comparison of matched estimated and measured time-series values and/or residuals; (4) comparison of cumulative values; and (5) cumulative frequency distributions.

Zacharias et al. (1996) presented robust quantitative techniques, from median-based nonparametric statistical methods (MdAE: median absolute error; RMdAE: relative median absolute error; REF: relative modelling efficiency), that can be used when the distribution of the observed data is non-Gaussian or when the sample size is not large enough to determine the underlying data distribution. Zacharias and Coakley (1993) categorized validation techniques into three main categories: summary statistics, hypothesis testing, and measures of goodness-of-fit, i.e. MaxE, a relative measure of the root mean square error (RRMSE), CD, EF and CRM. They listed examples of summary statistics as the mean, standard deviation, and those statistics commonly used with box or whisker plots (range, inter quartile range, and median).

Yang et al. (2000) examined the correlation across different statistics, allowing one to choose from each correlated group without loosing accuracy. They argued that the same conclusion can be achieved by using together either RMSE, modified modelling efficiency (EF1), paired t-test and E, or MAE, EF and E.

Model assessment by Mankin et al. (1977) and improved by Scholten and van der Tol (1998) was based on a comparison between model estimates and observations by using Venn diagrams and measures such as model adequacy (number of agreements between model and experiments / number of experiments) and reliability (number of agreements between model and experiments / number of model responses). These are helpful in discriminating between a better and worse model, and to define cases of useless or good models. Gardner and Urban (2003) illustrated a general approach to test model performance across a spectrum of methods via receiver-operator characteristic curves (ROC). Such approaches are based on the classification of results into discrete categories and the concept of a "confusion matrix" (Campbell, 1996), and imply defining false and true positives in the estimation of binary variables. Some statistics based on the confusion matrix are presented and discussed in Beguería (2006). Pontius Jr. and Schneider (2001) described how to use the ROC as a quantitative measurement to validate a land-cover change model. In Barbottin et al. (2008), a ROC curve analysis was carried out to estimate the frequencies of correct and incorrect indicator-based and model-based decisions, using the area under the ROC curve as summary of the overall accuracy of a model.

Patterns:

Change of patterns in the residuals can be assessed by testing the autocorrelation in the residuals (Vincent, 1998). Lin et al. (2002) developed model-checking techniques by taking the cumulative sums of residuals over certain coordinates to ascertain whether or not specific patterns exist in the residual plot. Donatelli et al. (2000, 2004a) proposed to quantify the presence of patterns of residuals versus independent variables (e.g., a model input or a variable not considered in the model), by computing pattern indices of two types: range-based (PI) and F-based (PI-F). Macro-patterns were revealed in model residuals by dividing the range of values of the external variable in two to five, fixed or varying sub-ranges. A pattern index in a percent relative-to-mean form was used as a validation measure by Bellocchi et al. (2003). In Trnka et al. (2006), estimated and observed herbage productions from permanent grassland were compared by using a vast array of pattern indices (against nitrogen fertilizer application rate, year, cut number, location, length of the growing season, date of the previous cut, number of snow days, two variants of accumulated air temperature, two variants of accumulated global solar radiation, and total precipitation during the period of sward growth) in conjunction with a set of performance statistics, i.e. MBE, RMSE, d, and Theil's inequality coefficient (U, Theil et al., 1970). The latter (ranging from 0 – perfect fit – to 1 – absence of any fit) penalizes large errors more than small ones and it also assesses a model's ability to duplicate a turning point or rapid changes in the data (Topp and Doyle, 2004). Pattern indices (versus month of year and minimum air temperature) in conjunction with error and correlation measures were also used by Diodato and Bellocchi (2007a) to assess the relative performance of three models of rainfall erosivity.

Correction factors:

Correction factors were developed by Harmel and Smith (2007) for the error term in some goodness-of-fit indicators (modelling efficiency, index of agreement, root mean square error, and mean absolute error) to incorporate the uncertainty of measured data into model validation, later improved by Harmel et al. (2008) to consider the effect of model uncertainty.

In model predictions, one approach to estimating the prediction measures is to adjust the naïve measures to get less biased estimates of the prediction measures. Estimators that are based on this approach include the Mallows' statistic (C_p, Mallows, 1973), the Akaike's information criterion (AIC, Akaike, 1974), and the Bayesian information criterion (BIC, Schwartz, 1978). All of these estimators, functions of the naïve measures, size of the dataset, and number of parameters in the model, generally lack of robustness (Li et al., 2003).

3.2. Disaggregating statistics

Once parameters of linear regression of model estimates versus actual data are estimated, the fitted line can be applied to generate new estimates of the variable under study. The difference between model-based and regression-based estimates defines the erratic portion of the error (and is the basis for computation of prediction error), while the systematic portion is described by the difference between regression-based estimates and actual data (the basis to assess the precision of the fitted linear regression, Tedeschi, 2006). Both model users and developers will focus on reducing the systematic error, the formers by model re-calibration, and the latters by better defining the basic equations. This is a basic concept by Aitken (1973) and Willmott (1981). More recently, Kobayashi and Salam (2000) developed the same concept to have residuals disaggregated into erratic and systematic components. They used the root mean square variation (RMSV) to indicate how much the model fails to estimate the variability of the measures around the mean, together with derived measures such as simulation bias (SB), square differences of the standard deviations (SDSD) and lack of positive correlation weighted by the standard deviations (LCS). These statistics are supportive in locating the causes of possible large deviations between estimates and measurements. The proportional contribution of systematic and erratic portions to the total error is helpful in determining areas in the model requiring further improvement. Further developing those findings, a different partitioning of mean square error into three additive components was given by Gauch and Fick (2003) and Gauch et al. (2003): they retained SB and derived the non-unity slope (NU), and the lack of correlation (LC). There is a unique quality to these approaches in the way the authors commented them in a letter exchange (Gauch et al., 2004; Kobayashi, 2004).

Trnka et al. (2006) called attention to disaggregating the Theil's coefficient (U). One of the main advantages of coefficient U is the possibility of calculating proportions of: estimated bias (U_B), indicating systematic errors; variance (U_S) that measures the ability of the model to replicate the degree of variability in the data; covariance (U_C), that is any remaining error, after accounting for the bias and variance effects. The ideal distribution of inequality over these three sources is for the bias and variance effects equal to zero, and the covariance equal to one.

3.3. Statistical hypothesis tests

Hypothesis testing is a formal approach to validation where either summary statistics or goodness-of-fit measures are tested against prescribed criteria (range of accuracy). In using statistical hypothesis testing to assess the validity of a model for its intended application, two hypotheses are formulated under the given set of experimental conditions: for the null hypothesis model is valid for an acceptable range of accuracy, and for the alternative hypothesis it is invalid under the same acceptable range of accuracy. Accepting the alternative hypothesis when the null hypothesis is true corresponds to the type-I statistical error, whose probability is also called model builder's risk. Accepting the null hypothesis when the alternative hypothesis is true matches the type-I statistical error, the second type of wrong decision with a probability called model user's risk (Balci and Sargent, 1982a).

Fundamental requirements:

Statistical tests assume that the outcomes based on the model are statistically accurate representations of what they purport to estimate. However, both systematic and random errors in biophysical studies may influence the accuracy of the estimators (as seen in Subsect. 2.7). Pennell et al. (1990) stated that graphical analyses allow for identification of trends in the data, systematic errors, and other potential sources of error, such as outliers. Marcus and Elias (1998) elaborated on five major areas of concern when applying formal statistical tests: observational data may not have been collected for the purposes of model validation; the sample size may be too small, allowing inadequate power to detect model deficiencies or to discriminate among competing models; the sample size may be so large that even useful models may be rejected by a statistical test for deviations that have little practical importance; measurement errors may bias the test statistics in the direction of attenuating the apparent goodness of the model; the temporal rhythms of the output variable may be influenced by systematic and random errors occurring in environmental factors.

Typical forms of the goodness-of-fit test are the following: Does the observed value minus model-estimated value equal zero (showing that the estimates are unbiased)? Does the ratio of observed value to model-estimated value equal one (in which case the estimates are relatively unbiased)? Numerous approaches involving statistical analysis have been used to evaluate model adequacy and several forms of statistical hypotheses are structured to show the level of confidence that the hypothesis is not rejected.

Using regressions:

The regression between observed and model-estimated values is commonly used because estimates of the intercept and the slope are good indicators of accuracy (the simultaneously closer to zero and unity, respectively, the higher the accuracy). Nonetheless, necessary assumptions have to be considered when performing a linear regression: the X-axis values are known without error; the Y-axis values have to be independent, random and with equal variance; residuals are independent and identically distributed. Some authors (e.g. Bland and Altman, 1995; Kleijnen et al., 1998) critically revisited the role of regression analysis in model validation, and suggested alternative approaches (difference between simulated and actual outputs against their sum, against their average, etc.) for achieving non-misleading regression testing. Fila et al. (2003a, b) suggested using the reduced major axis (RMA) method in place of the ordinary least squares method to estimate regression parameters. RMA has three desirable properties (Ricker, 1984): it is symmetric in X and Y (if the x and y axes are interchanged, the slope is replaced by its reciprocal and the line remains stationary about the data points); it is scale independent (the line does not change with a change of scale); it is robust to clusters of observations in the frequency distributions of data (the line usually describes the central trend even when the sample is not bivariate normal).

Thomann (1982) suggested using regression analysis and tests of slope (against one) and intercept (against zero) of a regression line in conjunction with other performance statistics. If the model is a good one, the regression will be a 45° line thorough the origin. Thus, the adequacy of the model can be determined by testing if intercept equals zero and slope equals one, separately using Student t-tests, or simultaneously using the F-test (Dent and Blackie, 1979; Mayer et al., 1994). These parametric tests for whether the regression line is significantly different from the 1:1 line assume that the data are normally distributed and independent, which is often not the case for data determined by non-linear biophysical processes.

Interpretation of statistics:

The test statistics may also have ambiguous interpretations. A statistical test of observed versus estimated values which fails to achieve the desired level of confidence does not necessarily indicate a problem in the model (usually the specification of one or more key parameters), because the problem may (or may also) reside in the observed data (usually the result of measurement error). With the t-tests for the intercept and slope, the more scatter in the data points, the greater is the standard error of the regression parameters, the smaller is the computed value for the test statistic and therefore, the harder it is to reject the null hypotheses which states the parameters are equal to zero and one respectively. Therefore, the test can fail to reject the null hypothesis either because the regression parameters are really not different from the values desired or there is much scatter around the line (Harrison, 1990). The F-based calculation that tests if the intercept and the slope coefficients simultaneously are respectively not different from zero and unity is affected by the same ambiguity as the t-tests. As shown by Analla (1998), the greater the estimation error the more difficult it is to reject the null hypothesis. Alternatively, the confidence interval should be used to investigate the range of the slope (Mitchell, 1997).

Empirical confidence intervals were proposed by Parrish and Smith (1990) as practical test for model validity founded on an overlap between the ranges of values computed on both model outputs and observations. Upper and lower limits of the range were computed as division and multiplication of the nominal model estimate by a chosen factor. The factor of two (Fa_2) is commonly used in air dispersion modelling, Kumar, 2000). Summaries of confidence limits on normalized mean square error, geometric mean variance and geometric mean bias were used by Patel and Kumar (1998) to select the best out of three air dispersion models.

When replicated experiments are available, the lack-of-statistical-fit (LOFIT) can be calculated and tested against an F-distribution (Whitmore, 1991). Assuming random experimental errors, the LOFIT distinguishes the mean deviation as a source of error from the failure of the model. Smith et al. (1997) assessed the statistical significance of difference-based indices assuming a deviation corresponding to a given confidence interval of the measurements. Fila et al. (2003a, b) proposed statistical methods to compare experiments against estimates when both are replicated. The use of bootstrapping techniques for validation of simulation models when parametric assumptions are violated was proposed by Kleijnen et al. (2001).

Non-parametric tests:

Contrary to parametric analyses, with nonparametric tests (i.e. variants of the Kolmogorov-Smirnov test as described in Stephens, 1974; the Wilcoxon-signed rank test, as described in Daniel, 1995) the assessment of adequacy of a model is related to its ability to yield the same ranking between observed and model-estimated values rather than model-estimates on observed values per se. Reckhow and Chapra (1983) listed measures of error, the t-test, the non-parametric Wilcoxon test, regression analysis, cross-correlation and box plots as appropriate statistical methods for deterministic models. Reckhow et al. (1990) also recommended various combinations of graphic (bivariate plots, histograms, and box plots) and statistical procedures based on the proposed analysis and intended use of modelling results. The χ^2 tests described by Agresti (2002) indicate whether the data points are homogenously distributed or if there is any tendency of over- or under-estimation.

The comparison of the distribution of the observed and model-estimated values has also been utilized to identify model adequacy for stochastic (Reynolds and Deaton, 1982) and deterministic models (Dent and Blackie, 1979). The common Kolmogorov-Smirnov's D test has been used to assess the probability that two data sets (observed and model-estimated values) have the same distribution. It consists to measure the overall difference of the area between two cumulative distribution functions (Press et al., 1992).

The use of multivariate statistics in model validation is not new (Balci and Sargent, 1982b). With multi-dimensional models, multivariate statistical analyses can be proficiently applied as in Mabille and Abecassis (2003), where a geometric model of wheat grain morphology was evaluated via principal component analysis and discriminating factorial analysis, and generating confidence limits in an elliptical plane.

3.4. Validation criteria

One of the difficulties when evaluating assessment metrics is determining what values indicate "good" or "bad" models. This section outlines the development of criteria (not statistical) that have been applied within a range of published studies. Clouse and Heatwole (1996) stated that "primary usefulness is in assessing which modelling scenarios are better predicted than other scenarios". However, other authors have taken a different approach by setting definitive criteria for several statistics.

Criteria used in James and Burges (1982) included CD and EF > 0.97 for good hydrological model performance. Dillaha (1990) stated that good hydrologic model assessment should estimate observed values within a factor of two, where parameters are measured on site, or where the model is calibrated and within a factor of 10 otherwise. Kumar (2000) used a criterion that air dispersion model estimates be within a factor of two, by looking at the percentage of estimates meeting such a criterion ($Fa_2 \geqslant 80\%$), combined with: $NMSE \leqslant 0.5$, $-0.5 \leqslant FB \leqslant +0.5$, $0.75 \leqslant MG \leqslant 1.25$, and $0.75 \leqslant VG \leqslant 1.25$.

The general categorization for the range of values of Pearson's correlation coefficient (r) by Hinkle et al. (1994) may indicate a straightforward (non-statistically based) way of interpreting the calculated correlation between estimates and measurements: 0.0 to 0.3, little (very weak) if any correlation; 0.3 to 0.5, low (weak) correlation; 0.5 to 0.7, moderate correlation; 0.7 to 0.9, high (strong) correlation; 0.9 to 1.0 very high (very strong) correlation.

In the Erosion Productivity Impact Calculator validation performed by Chung et al. (1999), the following criteria were chosen to assess if the model results were satisfactory: RMSE and MdAE < 50%, EF and REF > 0.3, $-0.2 < CRM +0.2$. Standards of < 20% for the percentage error and > 0.5 for r^2 were also set. In a following paper (Chung et al., 2000) the target criteria to judge if the model results were satisfactory were: $EF > 0.3$, $r^2 > 0.5$, and *P*-value of the paired t-test between the observed and simulated values > 0.025.

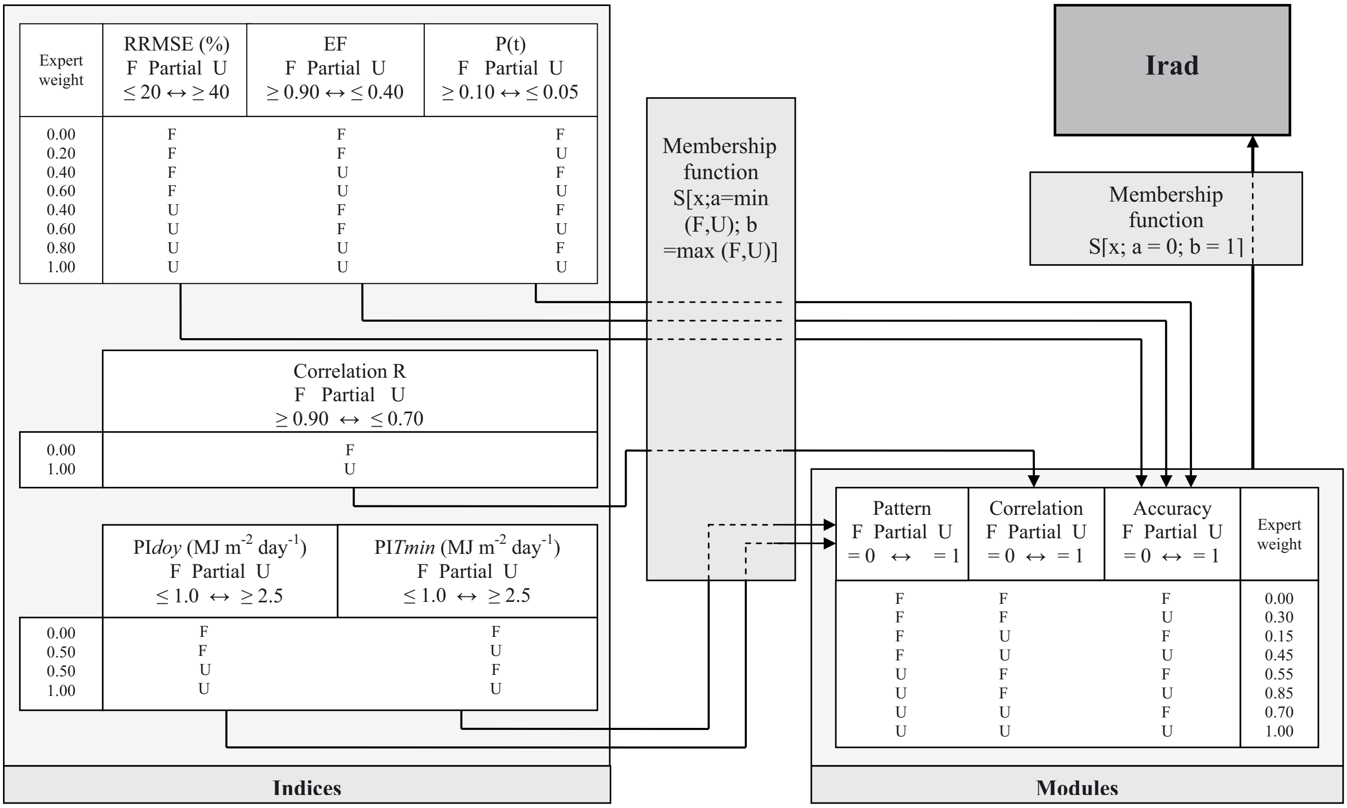


Figure 3. Structure of the fuzzy-based integrated index for solar radiation model assessment (F = favourable; U = unfavourable; S = membership function; a = minimum value of F; b = maximum value of U; after Rivington et al., 2005).

In Stöckle et al. (1999) RRMSE and Willmott's d index were taken together to evaluate the weather generator ClimGen. Upper and lower limits were suggested for a solid judgment on the model performance: good when RRMSE ⩽ 10% and d ⩾ 0.95, acceptable when 10% ⩽ RRMSE ⩽ 20% and d ⩾ 0.90, poor with other values. In a following paper, Stöckle et al. (2004) adopted similar standards (but in a more restrictive fashion) for eva1uating evapotranspiration estimates: d ⩾ 0.95 and RRMSE ⩽ 10%, very good; d ⩾ 0.95 and 15% ⩾ RRMSE > 10%, good; d ⩾ 0.95 and 20% ⩾ RRMSE > 15%, acceptable; d ⩾ 0.95 and 25% ⩾ RRMSE > 20%, marginal. Other combinations of d and RRMSE values indicated poor performance. In addition, all combinations with slope > 1.1 or < 0.9 and r^2 < 0.85 of the regression observed versus estimated values (forced through zero) were considered poor.

3.5. Combining multiple statistics

It emerges from the discussion above that each individual performance measure will only assess a single component of model behaviour and is not sufficient to judge model adequacy to a given purpose. It is possible that a model can be deemed unsuitable and rejected based on an assessment by one statistic assessing one form of model performance, whilst other attributes of the model may be desirable, as assessed by other measures. Similarly, a model may be seen as acceptable based on one performance statistic but still contains poor qualities not assessed by appropriate tests. The use of multiple metrics allows a greater range of model estimate behaviour to be tested, but still leaves the issue of how to achieve an overall model validation, i.e. weighing up the balance of positive and negative attributes.

The combination of multiple assessment metrics and the setting of criteria have evolved into formal structures, becoming attractive and being regarded as a positive step in achieving robust assessments. Bellocchi et al. (2002a) dealt with the need for integrated methods for model validation by introducing the concept of fuzzy multiple-metric assessment expert system to aggregate different metrics into one indicator for use in solar radiation model validation. The fuzzy system reflects the author's expert judgment about the quality of model performance. This approach enables the calculation of a single indicator made up of a number of individual metrics representing different indices or test statistics. Such an approach provides a comprehensive assessment, making it easier to identify best performing models.

The method exists as a flexible, open-ended structure (Fig. 3) in which a range of metrics can be aggregated into a single modular indicator, based on an expert weighting expression of the balance of importance of the individual indices and their aggregation into modules. The Bellocchi et al. (2002a) study used: relative root mean square error (RRMSE), modelling efficiency (EF), the probability of paired Student-*t* test (P(t)); the correlation coefficient of the estimates versus measurements (r) and the two pattern indices detailed in Donatelli et al. (2004a), one computed versus day of year (PI_{doy}), and the other versus minimum temperature (PI_{Tmin}). Values of each statistic are computed and then aggregated into modules: Accuracy (RRMSE, EF, P(t)); Pattern (PI_{doy}, PI_{Tmin}); and Correlation (r). A module is a validation measure calculated via a fuzzy-based procedure from one or more basic statistics. For each module, a dimensionless value between 0 (best model response) and 1 (worst model response) is calculated. The method adopts the Sugeno approach of fuzzy inference (Sugeno, 1985). Three membership classes can be defined for all indices, according to expert judgment, i.e. favourable, unfavourable and partial (or fuzzy) membership, using S-shaped curves as transition probabilities in the range favourable to unfavourable. A two-stage design of a fuzzy-based rules inferring system is applied, where firstly several metrics are aggregated into modules and then, using the same procedure, the modules are aggregated in a second level integrated index (again, ranging from 0 to 1), called *indicator*. The expert reasoning runs as follows: if all input variables are favourable, the value of the module is 0 (good agreement between estimates and measurements); if all indices are unfavourable, the value of the module is 1 (poor agreement), while all the other combinations assume intermediate values. The weights can be chosen based on the users own experience in handling each statistic. In Bellocchi et al. (2002a) a decreasing importance was assigned to the modules: Accuracy, Pattern and Correlation. Rivington et al. (2005), Diodato and Bellocchi (2007b, c) and Abraha and Savage (2008) demonstrated the value of employing this method, in conjunction with using graphical illustrations, to gain a fine level of detail about model quality. Similarly to this, Donatelli et al. (2004b) developed integrated indices for evaluating the estimates from pedotransfer functions.

A fuzzy-based methodology was also proposed (Donatelli et al., 2002a) to identify mismatches between estimated and measured time series by using a fuzzy-based approach. The mismatch in time series comparison of LEACHM soil nitrogen estimates was identified by means of an integrated index derived aggregating RMSE and PI_{doy} (values $\leqslant 2$ as favourable and $\geqslant 12$ as unfavourable for both), 0.8 and 0.2 being the relative weights, and calculated reiterating the computation over a 100-day shift of model estimates. The same approach was integrated with a complementary set of statistics from the environmental modelling literature (e.g. Environmental Protection Agency, 1991) to evaluate SUNDIAL soil mineral nitrogen estimates (Bellocchi et al., 2004). Bellocchi et al. (2002b) extended the original fuzzy-based multiple-metric assessment system approach to aggregating the RRMSE values (values $\leqslant 20\%$ as favourable and $\geqslant 40\%$ as unfavourable) computed over different outputs in cropping systems modelling under different sets of conditions, thus allowing a comprehensive assessment of the model's performance by means of one integrated index. They attributed major weight to above ground biomass (i.e., 2), where a minor incidence of soil variables (water content: 1, nitrate content: 0.5) was recognized.

Aggregating measures of performance have in common that the information contained in the errors is aggregated into a single numerical value. Herbst and Casper (2008) argued that essentially different model results can be obtained with close to identical performance measure values. Because of their low discriminatory power, performance measures might not be well suitable to give evidence of the difference or equivalence between alternative model realizations. As a step towards improved extraction of information from existing data they introduced an approach, the Self-Organizing Map (SOM), which uses a number of performance statistics and represents, for each of them, the colour-coded spectrum of model realizations obtained from Monte-Carlo simulations. It was applied to a distributed conceptual watershed model. SOM is a type of artificial neural network and unsupervised learning algorithm that is used for clustering, visualization and abstraction of multi-dimensional data. Such an algorithmic approach mainly targets the optimization and identification of parameters that mostly affect the model output (sensitivity analysis), and is not of direct interest for this review.

3.6. Summary

Because of the vast collection and diversity of the approaches to assess models, we sorted through and straightened out validation statistics showing how they were introduced and applied to biophysical modelling. The review of the measures of performance that are commonly used in model validation reveals the positions assumed over time in the modelling literature mostly emphasising that a single statistic will only reveal one aspect of model performance. As each approach has its advantages and drawbacks, they are rather complementary and are generally used in combination in model validation. What values for assessment metrics indicate satisfactory models remains a subjective issue and no definitive guidance exists because of heterogeneity of approaches and application domains. While agreeing with many authors that model validation has to be performed using a set of dissimilar validation statistics, we in particular advocate the use of combined multiple statistics where several measures for validation can be considered both separately (each individual metric) and collectively (integrated indicator). Test statistics may be problematic because they rely on assumptions that are difficult to check in biophysical systems. Decomposition of statistics in basic terms may disclose the characteristic and the actual structure of the error, but the combination of multiple metrics into synthetic indicators where subjective choices (expert decisions) are converted into explicit and transparent rules reveals a more comprehensive picture. The lack of precise and undisputable criteria to consider a specific metric as more effective than others, and the multiplicity of aspects to be accounted for a multi-perspective evaluation of model performance, logically leads to some use of composite metrics for model validation. A composite metric is not the only output of composition: the modeller can "drill down" to module values, and finally to basic metrics to better understand the synthetic result provided by the composite indicator. In such respect, composition of metrics should be considered a shift of paradigm from merely selecting the best out of a set of evaluation metrics.

4. CONCLUSION

This paper discusses issues concerned with model validation and reviews the most commonly used forms of estimate testing. Exposition of material and explanation of concepts presented in Section 2 ("Issues on model validation") reflect the authors' perception of issues that are fundamental to understanding the factors that are related to model validation. The examples provided throughout the text demonstrate how previous instances of model use (and success or failure associated with that use) are the growing knowledge bases acquired from using different models for various applications. The publications cited show how the scope and capabilities of validation approaches have evolved and improved with time. Though finding solution of how best to evaluate numerical values produced by models will remain an issue, a range of approaches do exist for improving the testing of model estimates. Our hope is that these approaches will continue to evolve.

Our historical reconstruction of the approaches serving the validation purposes, as presented in Section 3 ("Validation of models"), points towards three main outcomes achieved: disaggregation of validation statistics into basic components, introduction of validation criteria, and combination of statistics into synthetic indicators. Baseline thresholds of validation measures (extracted from the international literature and recapped in sub-section "Validation criteria") provide users with the modellers' perception of good/bad performance statistics. Such criteria are presented and discussed not only to make available reference values of possible use in future validation studies, but also they call on the need for using expert rules to guide the validation process. This review of the methods available for numerical testing has shown that greater value can be gained through combined use and rule-based aggregation of multiple approaches to achieve a more complete form of validation.

Advancements in these numerical testing methodologies for validation need to be put into structured frameworks comprised of processes such as sensitivity and uncertainty analyses (parameter and input variable appraisal), parameter optimization, model structure assessment (expert review), software testing, etc. As such, validation must be seen as an integral part of the overall model development and application process, whilst also encompassing the requirement for better data quality control and meta data recording. This places a greater emphasis on the need to include validation plans within model project proposals and a higher level of support for validation work by funding organisations. As shown, models may come in a variety of time and space resolutions and scales. Matching these scales and ensuring consistency in the overall model is not a trivial process and may be difficult to fully automate. Techniques to validate models need to be developed at the same pace with which the models themselves are created, improved and applied. Also, validation steps must be clearly stated, accessible, transparent, and understandable to non-modellers. As discussed in the context of the current knowledge, this can be achieved by means of reliability statistics, history of previous use, or personal preferences. However, details about validation techniques development go beyond the aim of this review, and a second paper on this broad topic may be arranged later as a natural evolution of what has already been presented.

Acknowledgements: This paper was developed with the support of the Italian Ministry of Agricultural Policies; project SIPEAA (http://www.sipeaa.it), paper No. 36. The authors also gratefully acknowledge the support of the Scottish Government Environment Rural Affairs Department. Thanks also to David Elston and David Walker of Biomathematics and Statistics Scotland (Macaulay Institute, Aberdeen, UK) for constructive criticism and suggestions.

REFERENCES

Abraha M.G., Savage M.J. (2008) Comparison of estimates of daily solar radiation from air temperature range for application in crop simulations, Agr. Forest Meteorol. 148, 401–416.

Addiscott T.M., Whitmore A.P. (1987) Computer simulation of changes in soil mineral nitrogen and crop nitrogen during autumn, winter and spring, J. Agr. Sci. 109, 141–157.

Agresti A. (2002) Categorical data analysis, (2nd ed.), Wiley, New York, NY, USA.

Aitken A.P. (1973) Assessing systematic errors in rainfall runoff models, J. Hydrol. 20, 121–136.

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

American Institute of Aeronautics and Astronautics (1998) Guide for the verification and validation of computational fluid dynamics, American Institute of Aeronautics and Astronautics, AIAA-G-077-1998, Reston, VA, USA.

American Society for Testing and Material (1984) Standard practice for evaluating environmental face models of chemicals, American Society for Testing and Material, Philadelphia, PA, USA, Standard E 978–984.

Analla M. (1998) Model validation through the linear regression fit to actual versus predicted values, Agr. Syst. 57, 115–119.

Anderson M.G., Bates P.D. (2001) Hydrological science: model credibility and scientific understanding, in: Anderson M.G., Bates P.D. (Eds.), Model validation: perspectives in hydrological science, John Wiley & Sons, New York, USA, Vol. 1, pp. 1–10.

Anonymous (2003) How to estimate grain harvest losses, Prairie Grains, Issue 54.

Argent R.M. (2004) An overview of model integration for environmental applications – components, frameworks and semantics, Environ. Modell. Softw. 19, 219–234.

Arnold T., Berger T., Uribe T. (2008) Step by step calibration of an integrated system for irrigation management, in: Quinn N.W.T. (Ed.) Integration of sensor networks and decision support tools for basin-scale, real-time water quality management, in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 1, pp. 584–591.

Aumann C.A. (2008) A methodology for building credible models for policy evaluation, in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 1, pp. 1025–1032.

Austin M.P., Belbin L., Meyers J.A., Doherty M.D., Luoto M. (2006) Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory, Ecol. Model. 199, 197–216.

Bair E.S. (1994) Model (in)validation – a view from courtroom, Ground Water 32, 530–531.

Balci O., Sargent R.G. (1982a) Some examples of simulation model validation using hypothesis testing, in: Highland H.J., Chao Y.W., Madrigal O. (Eds.), Proc. 14th Conference on Winter Simulation, December 6–8, San Diego, CA, USA, Vol. 2, pp. 621–629.

Balci O., Sargent R.G. (1982b) Validation of multi-variate response simulation models by using Hotelling's two-sample T^2 test, Simulation 39, 185–192.

Barbottin A., Le Bail M., Jeuffroy M.H. (2006) The Azodyn crop model as a decision support tool for choosing cultivars, Agron. Sustain. Dev. 26, 107–115.

Barbottin A., Makowski D., Le Bail M., Jeuffroy M.-H., Bouchard C., Barrier C. (2008) Comparison of models and indicators for categorizing soft wheat fields according to their grain protein contents, Eur. J. Agron. 29, 175–183.

Beguería S. (2006) Validation and evaluation of predictive models in hazard assessment and risk management, Nat. Hazards 37, 315–329.

Bellocchi G. (2004) Appendix A. Numerical indices and test statistics for model evaluation, in: Pachepsky Ya., Rawls W.J. (Eds.), Development of pedotransfer functions in soil hydrology, Elsevier, Amsterdam, The Netherlands, pp. 394–400.

Bellocchi G., Acutis M., Fila G., Donatelli M. (2002a) An indicator of solar radiation model performance based on a fuzzy expert system, Agron. J. 94, 1222–1233.

Bellocchi G., Donatelli M., Fila G. (2003) Calculating reference evapotranspiration and crop biomass using estimated radiation inputs, Ital. J. Agron. 7, 95–102.

Bellocchi G., Fila G., Donatelli M. (2002b) Integrated evaluation of cropping systems models by fuzzy-based procedure, in: Villalobos F.J., Testi L. (Eds.), Proc. 7th European Society for Agronomy Congress, 15–18 July, Cordoba, Spain, pp. 243–244.

Bellocchi G., Smith J., Donatelli M., Smith P. (2004) Improvements in time mismatch analysis of model estimates, in: Jacobsen S.E., Jensen C.R. Porter J.R. (Eds.), Proc. of 8th European Society for Agronomy Congress, 11–15 July, Copenhagen, Denmark, pp. 221–222.

Berk R.A., Bickel P., Campbell K. (2002) Workshop on statistical approaches for the evaluation of complex computer models, Stat. Sci. 17, 173–192.

Berk R.A., Fovell R.G., Schoenberg F., Weiss R.E. (2001) The use of statistical tools for evaluating computer simulations – an editorial essay, Climatic Change 51, 119–130.

Beven K.J. (1993) Prophecy, reality and uncertainty in distributed hydrological modelling, Adv. Water Resour. 16, 41–51.

Beven K.J. (2007) Towards integrated environmental models of everywhere: uncertainty, data and modelling as a learning process, Hydrol. Earth Syst. Sc. 11, 460–467.

Beven K.J., Freer J. (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol. 249, 11–29.

Bland J.M., Altman D.G. (1995) Comparing methods of measurement: why plotting against standard method is misleading, Lancet 346, 1085–1087.

Bolte J.P., Hulse D.W., Gregory S.V., Smith C. (2004) Modelling biocomplexity - actors, landscapes and alternative futures, in: Pahl-Woslt C., Schmidt S., Rizzoli A.E., Jakeman A.J. (Eds.), Complexity and integrated resources, Trans. 2nd Biennial Meeting of the International Environmental Modelling and Software Society, 14–17 June, Osnabrück, Germany, Vol. 1, pp. 1–10.

Breiman L., Spector P. (1992) Submodel selection and evaluation in regression: the X-random case, Int. Stat. Rev. 60, 291–319.

Campbell J.B. (1996) Introduction to remote sensing, 2nd ed., The Guilford Press, New York, NY, USA.

Cardoso A.C., Lopes J.F. (2008) 3D ecological modelling of the Aveiro coast (Portugal), in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 1, pp. 181–190.

Checkland P.B. (1981) Systems thinking, systems practice, John Wiley & Sons, London.

Chen D.X., Coughenour M.B. (2004) Photosynthesis, transpiration, and primary productivity: Scaling up from leaves to canopies and regions using process models and remotely sensed data, Global Biogeochem. Cy. 18, GB4033.

Cheng R.T., Burau J.R., Gartner J.W. (1991) Interfacing data analysis and numerical modelling for tidal hydrodynamic phenomena, in: Parker B.B. (Ed.), Tidal hydrodynamics, John Wiley & Sons, New York, NY, USA, pp. 201–219.

Chung S.W., Gasman P.W., Huggins D.R., Randall G.W. (2000) Evaluation of EPIC for three Minnesota cropping systems. Working paper 00-WP 240, Centre for Agricultural and Rural Development, Iowa State University, Ames, IO, USA.

Chung S.W., Gasman P.W., Kramer L.A., Williams J.R., Gu R. (1999) Validation of EPIC for two watersheds in Southwest Iowa, J. Environ. Qual. 28, 971–979.

Clouse R.W., Heatwole C.D. (1996) Evaluation of GLEAMS considering parameter uncertainty, ASAE paper No. 96-2023, St. Joseph, MI, USA.

Cochran W.G., Cox G.M. (1957) Experimental design, John Wiley & Sons, New York, NY, USA.

Cornelissen J.H.C., Lavorel S., Garnier E., Diaz S., Buchmann N., Gurwich D.E., Reich P.B., ter Steege H., Morgan H.D., van der Heijden M.G.A., Pausas J.G., Poorter H. (2003) A handbook of protocols for standardised and easy measurement of plant functional traits worldwide, Aust. J. Bot. 51, 335–380.

Daniel W.W. (1995) Biostatistics: a foundation for analysis in the health sciences, John Wiley & Sons Inc., New York, NY, USA.

Davies J.A., McKay D.C. (1989) Evaluation of selected models for estimating solar radiation on horizontal surfaces, Sol. Energy 43, 153–168.

De Wit C.T. (1993) Philosophy and terminology, in: Leffelaar P.A. (Ed.), On systems analysis and simulation of ecological processes – with examples in CSMP and FORTRAN. Kluver, Dordrecht, The Netherlands, pp. 3–9.

Dent J.B., Blackie M.J. (1979) Systems simulation in agriculture, Applied Science Publishers Ltd., London, United Kingdom.

Dhanoa M.S., Lister S.J., France J., Barnes R.L. (1999) Use of mean square prediction error analysis and reproducibility measures to study near infrared calibration equation performance, J. Near Infrared Spec. 7, 133–143.

Dias, J.M., Lopes, J.F. (2006) Implementation and assessment of hydrodynamic, salt and heat transport models: the case of Ria de Aveiro Lagoon (Portugal), Environ. Modell. Softw. 21, 1–15.

Diekkrüger B., Söndgerath D., Kersebaum K.C., McVoy C.V. (1995) Validity of agroecosystem models applied to the same data set, Ecol. Model. 81, 3–29.

Dillaha T.A. (1990) Role of best management practices in restoring the health of the Chesapeake Bay: Assessments of effectiveness, in: Haire M., Krome E.C. (Eds.), Perspectives on the Chesapeake Bay, 1990: Advances in estuarine sciences. CBP/TRS41/90. Chesapeake Bay Consortium. US EPA Chesapeake Bay Program. Annapolis, Maryland, USA, pp. 57–81.

Diodato N., Bellocchi G. (2007a) Estimating monthly (R)USLE climate input in a Mediterranean region using limited data, J. Hydrol. 345, 224–236.

Diodato N., Bellocchi G. (2007b) Modelling reference evapotranspiration over complex terrains from minimum climatological data, Water Resour. Res. 43, doi:10.1029/2006WR005405.

Diodato N., Bellocchi G. (2007c) Modelling solar radiation over complex terrains using monthly climatological data, Agr. Forest Meteorol. 144, 111–126.

Donatelli M., Acutis M., Bellocchi G. (2000) Two statistical indices to quantify patterns of errors produced by models, in: Christen O., Ordon F. (Eds.), Proc. 3rd International Crop Science Conference, 17–22 August, Hamburg, Germany, p. 186.

Donatelli M., Acutis M., Bellocchi G., Fila G. (2004a) New indices to quantify patterns of residuals produced by model estimates, Agron. J. 96, 631–645.

Donatelli M., Acutis M., Fila G., Bellocchi G. (2002a) A method to quantify time mismatch of model estimates, in: Villalobos F.J., Testi L. (Eds.), Proc. 7th European Society for Agronomy Congress, 15–18 July, Cordoba, Spain, pp. 269–270.

Donatelli M., Acutis M., Nemes A., Wösten H. (2004b) Integrated indices for pedotransfer function evaluation, in: Pachepsky Ya., Rawls W.J. (Eds.), Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam, The Netherlands, pp. 363–390.

Donatelli M., van Ittersum M.K., Bindi M., Porter J.R. (2002b) Modelling cropping systems – highlights of the symposium and preface to the special issues, Eur J. Agron. 18, 1–11.

Efron B. (1986) how biased is the apparent error rate of a prediction rule, J. Am. Stat. Assoc. 81, 461–470.

Environmental Protection Agency (1991) Guideline for regulatory application of the urban airshed model. U.S., Environmental Protection Agency Office of Air Quality Planning and Standards, Research Triangle Park, NC, 27711, EPA-450/4-91-013.

Favis-Mortlock D., Boardman J., MacMillan V. (2001) The limits of erosion modeling: why we should proceed with care, in: Harmon R.S., Doe W.W. III (Eds.), Landscape erosion and evolution modeling. Kluwer Academic/Plenum Publisher, New York, NY, USA, pp. 477–516.

Fila G., Bellocchi G., Acutis M., Donatelli M. (2003a) IRENE: a software to evaluate model performance, Eur. J. Agron. 18, 369–372.

Fila G., Bellocchi G., Donatelli M., Acutis M. (2003b) IRENE_DLL: A class library for evaluating numerical estimates, Agron. J. 95, 1330–1333.

Fodor N., Kovács G.J. (2003) Sensitivity of 4M model to the inaccuracy of weather and soil input data, Appl. Ecol. Environ. Res. 1, 75–85.

Forrester J.W. (1961) Industrial dynamics, Pegasus Communications, Waltham, MA, USA.

Forsythe W.C., Rykiel E.J. Jr., Stahl R.S., Wu H., Schoolfield R.M. (1995) A model comparison for daylength as a function of latitude and day of year, Ecol. Model. 80, 87–95.

Fox D.G. (1981) Judging air quality model performance: a summary of the AMS workshop on dispersion models performance, Bull. Am. Meteorol. Soc. 62, 599–609.

Franks SW., Beven K.J., Quinn P.F., Wright I.R. (1997) On the sensitivity of soil-vegetation-atmosphere transfer (SVAT) schemes: equifinality and the problem of robust calibration, Agr. Forest Meteorol. 86, 63–75.

Gardner R.H., Urban D.L. (2003) Model validation and testing: past lessons, present concerns, future prospects, in: Canham C.D., Cole J.J., Lauenroth W.K. (Eds.), Models in ecosystem science. Princeton University Press, Princeton, NJ, USA, pp. 184–203.

Gauch H.G. Jr., Fick W. (2003) In model evaluation, what is X and what is Y? in: 2003 annual meeting abstracts. [CD-ROM], ASA, CSSA and SSSA, Madison, WI, USA.

Gauch H.G. Jr., Gene Hwang J.T., Fick G.W. (2003) Model evaluation by comparison of model-based predictions and measured values, Agron. J. 95, 1442–1446.

Gauch H.G. Jr., Gene Hwang J.T., Fick G.W. (2004) Reply, Agron. J. 96, 1207–1208.

Global Climate and Terrestrial Ecosystems (1992) Effects of change on the wheat ecosystem, Workshop report, GCTE Focus 3 meeting, Saskatoon, SK, Canada, 22–24 July, University of Oxford, United Kingdom.

Gobas F.A.P.C., Pasternak J.P., Lien K., Duncan R.K. (1998) Development and field validation of a multimedia exposure assessment models for waste load allocation in aquatic ecosystems: application to 2,3,7,8-tetrachlorodibenzo-p-dioxin and 2,3,7,8-tetrachlorodibenzofuran in the Fraser River watershed, Environ. Sci. Technol. 32, 2442–2449.

Green I.R.A., Stephenson D. (1986) Criteria for comparison of single event models, J. Hydrol. Sci. 31, 395–411.

Greenwood D.J., Neeteson J.J., Draycott A. (1985) Response of potatoes to N fertilizer: dynamic model, Plant Soil 85, 185–203.

Hamilton M.A. (1991) Model validation: an annotated bibliography, Comm. Stat. Theor. M. 20, 2207–2266.

Hanft J.M., Wych R.D. (1982) Visual indicators of physiological maturity of hard red spring wheat, Crop Sci. 22, 584–588.

Harmel R.D., Smith P.K. (2007) Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modelling, J. Hydrol. 337, 326–336.

Harmel R.D., Smith D.R., King K.W., Slade R.M., Smith P. (2008) Data uncertainty estimation tool for hydrology and water quality (DUET-H/WQ): estimating measurement uncertainty for monitoring and modelling applications, in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 1, pp. 574–583.

Harrison S.R. (1990) Regression of a model on real-system output: an invalid test of model validity, Agr. Syst. 34, 183–190.

Heinmann A.B., Hoogenboom G., Chojnicki B. (2002) The impact of potential errors in rainfall observations on the simulation of crop growth, development and yield, Ecol. Model. 157, 1–21.

Herbst M., Casper M.C. (2008) Towards model evaluation using Self-Organizing Maps, in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 3, pp. 1055–1062.

Hinkle D., Wiersma W., Jurs S. (1994) Applied statistics for the behavioural sciences, 3rd ed., Houghton Mifflin Company, Boston, MT, USA.

Hochman Z., van Rees H., Carberry P.S., Holzworth D., Dalgliesh, N.P., Hunt J., Poulton P.L., Brennan, L.E., Darbas T., Fisher J., van Rees, S., Huth N.I. Peake A.S., McCown R.L. (2005) Can access to a cropping system simulator help farmers reduce risk in drought-prone environments? in: InterDrought-II, 2nd Int. Conf. Integrated Approaches to Sustain and Improve Plant Production Under Drought Stress 24–28 September, Rome, Italy.

Hoogenboom G. (2000) Contribution of agro-meteorology to the simulation of crop production and its applications, Agr. Forest Meteorol. 103, 137–157.

Hsu M.H., Kuo A.Y., Kuo J.T., Liu W.C. (1999) Procedure to calibrate and verify numerical models of estuarine hydrodynamics, J. Hydrol. Eng. 125, 166–182.

Huth N., Holzworth D. (2005) Common sense in model testing, in: Zerger A., Argent R.M. (Eds.), Proc. MODSIM 2005 International Congress on Modelling and Simulation: Advances and applications for management and decision making, 12–15 December, Melbourne, Australia, pp. 2804–2809.

Institute of Electrical and Electronics Engineers (1991) IEEE standard glossary of software engineering terminology, IEEE, IEEE Std 610.12-1990, New York, NY, USA.

Jakeman A.J., Letcher R.A., Norton J.P. (2006) Ten iterative steps in development and evaluation of environmental models, Environ. Modell. Softw. 21, 606–614.

James L.D., Burges S.J. (1982) Selection, calibration, and testing of hydrologic models, in: Haan C.T., Johnson H.P., Brakensiek D.L. (Eds.), Hydrologic modelling of small watersheds, American Society of Agricultural Engineers, St. Joseph, MI, USA, pp. 437–472.

Jamieson P.D., Porter J.R., Semenov M.A., Brooks R.J., Ewert F., Ritchie J.T. (1999) Comments on "Testing winter wheat simulation models predictions against observed UK grain yields" by Landau et al. (1998), Agr. Forest Meteorol. 96, 157–161.

Janssen P.H.M., Heuberger P.S.C. (1995) Calibration of process-oriented models, Ecol. Model. 83, 55–66.

Kanevski M., Pozdnoukhov A., Timonin V. (2008) Machine learning algorithms for geospatial data. Applications and software tools, in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 1, pp. 320–327.

Keating B.A., Robertson M.J., Muchow R.C., Huth N.L. (1999) Modelling sugarcane production systems. 1. Development and performance of the sugarcane module, Field Crops Res. 61, 253–271.

Kleijnen J.P.C., Bettonvil B., van Groenendaal W. (1998) Validation of trace-driven simulation models: a novel regression test, Manage. Sci. 44, 812–819.

Kleijnen J.P.C., Cheng R.C.H., Bettonvil B. (2001) Validation of trace-driven simulation models: bootstrapped tests, Manage. Sci. 47, 1533–1538.

Klepper O. (1989) A model of carbon flow in relation to macrobenthic food supply in the Oosterschelde estuary (S.W. Netherlands), PhD-Thesis, Wageningen Agricultural University, The Netherlands.

Kobayashi K. (2004) Comments on another way of partitioning mean squared deviation proposed by Gauch et al. (2003), Agron. J. 96, 1206–1207.

Kobayashi K., Salam M.U. (2000) Comparing simulated and measured values using mean squared deviation and its components, Agron. J. 92, 345–352.

Krause P., Boyle D.P., Bäse F. (2005) Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci. 5, 89–97.

Kuhnert M., Voinov A., Seppelt R. (2005) Comparing raster map comparison algorithms for spatial modeling and analysis, Photogramm. Eng. Rem. S. 71, 975–984.

Kumar A. (2000) Dispersion and risk modelling, Department of Civil Engineering, University of Toledo, OH, USA, CIVE-6630:995.

Kvalseth T.O. (1985) Cautionary note about R2, Am. Stat. 39, 279–285.

Landau S., Mitchell R.A.C., Barnett V., Colls J.J., Craigon J., Moore K.L., Payne R.W. (1998) Testing winter wheat simulation models' predictions against observed UK grain yields, Agr. Forest Meteorol. 89, 85–99.

Landau S., Mitchell R.A.C., Barnett V., Colls J.J., Craigon J., Payne R.W. (1999) Response to "Comments on 'Testing winter wheat simulation models predictions against observed UK grain yields by Landau et al. [Agr. For. Meteorol. 89 (1998) 85-99]' by Jamieson et al. [Agr. For. Meteorol., this issue]", Agr. Forest Meteorol. 96, 163–164.

Landau S., Mitchell R.A.C., Barnett V., Colls J.J., Craigon J., Payne R.W. (2000) A parsimonious, multiple-regression model of wheat yield response to environment, Agr. Forest Meteorol. 101, 151–166.

Landry M., Oral M. (1993) In search of a valid view of model validation for operations research, Eur. J. Oper. Res. 66, 161–167.

Leffelaar P.A. (1990) On scale problems in modelling: an example from soil ecology, in: Rabbinge R., Goudriaan J., van Keulen H., Penning de Vries F.W.T., van Laar H.H. (Eds.), Theoretical production ecology: reflections and prospects. Simulation Monographs 34, Pudoc, Wageningen, The Netherlands, pp. 57–73.

Leffelaar P.A., Meike H., Smith P., Wallach D. (2003) Modelling cropping systems – highlights of the symposium and preface to the special issues. 3. Session B. Model parameterisation and testing, Eur. J. Agron. 18, 189–191.

Li W., Arena V.C, Sussman N.B.. Mazumdar S. (2003) Model validation software for classification models using repeated partitioning: MVREP, Comput. Meth. Prog. Bio. 72, 81–87.

Lin D.Y., Wei L.J., Ying Z. (2002) Model-checking techniques based on cumulative residuals, Biometrics 58, 1–12.

Ljung L. (1999) System identification – Theory for the user, 2nd ed., Prentice Hall, Upper Saddle River, NJ, USA.

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

Mabille F., Abecassis J. (2003) Parametric modelling of wheat grain morphology: a new perspective, J. Cereal Sci. 37, 43–53.

Makowski D., Hillier J., Wallach D., Andrieu B., Jeuffroy M.-H. (2006) Parameter estimation for crop models, in: Wallach D., Makowski D., Jones J.W. (Eds.), Working with dynamic models. Evaluation, analysis, parameterization and applications, Elsevier, Amsterdam, pp. 101–150.

Mallows C. (1973) Some comments on Cp, Technometrics 15, 661–675.

Mankin J.B., O'Neill R.V., Shugart H.H., Rust B.W. (1977) The importance of validation in ecosystem analysis, in: Innis G.S. (Ed.), New directions in the analysis of ecological systems, Proc. Series Simulation Council 5(1), La Jolla, CA, USA, pp. 63–72.

Marcus A.H., Elias R.W. (1998) Some useful statistical methods for model validation, Environ. Health Persp. 106, 1541–1550.

Martorana F., Bellocchi G. (1999) A review of methodologies to evaluate agro-ecosystems simulation models, Ital. J. Agron. 3, 19–39.

Matthews K.B., Rivington M., Blackstock K., Buchan K., Miller D.G. (2008) Raising the bar – Is evaluating the outcomes of decision and information support tools a bridge too far? in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 1, pp. 948–955.

Matthews K.B., Rivington M., Buchan K., Miller D., Bellocchi G. (2008) Characterising and communicating the agro-meteorological implications of climate change scenarios to land management stakeholders, Climate Res. 37, 59–75.

Matthews K.B., Sibbald A.R., Craw S. (1999) Implementation of a spatial decision support system for rural land use planning: integrating GIS and environmental models with search and optimisation algorithms, Comput. Electron. Agr. 23, 9–26.

Mayer D.G., Butler D.G. (1993) Statistical validation, Ecol. Model. 68, 21–32.

Mayer D.G., Stuart M.A., Swain A.J. (1994) Regression of real-world data on model output: an appropriate overall test of validity, Agr. Syst. 45, 93–104.

Medlyn B.E., Jarvis P.G. (1999) Design and use of a database of model parameters from elevated CO_2 experiments, Ecol. Model. 124, 69–83.

Medlyn B.E., Robinson A.P., Clement R., McMurtrie E. (2005) On the validation of models of forest CO_2 exchange using eddy covariance data: some perils and pitfalls, Tree Physiol. 25, 839–857.

Meehl G.A., Covey C., McAvaney B., Latif M., Stouffer R.J. (2005) Overview of the coupled model intercomparison project, Bull. Am. Meteorol. Soc. 86, 89–93.

Metselaar K. (1999) Auditing predictive models: a case study in crop growth, PhD-Thesis, Wageningen Agricultural University, Wageningen.

Mihram G.A. (1972) Some practical aspects of the verification and validation of simulation models, Oper. Res. Quart. 23, 17–29.

Mitchell P.L. (1997) Misuse of regression for empirical validation of models, Agr. Syst. 54, 313–326.

Moberg A., Jones P.D. (2004) Regional climate model simulations of daily maximum and minimum near-surface temperatures across Europe compared with observed station data 1961-90, Clim. Dynam. 23, 695–715.

Myung J., Pitt M.A. (2003) Model fitting, in: Nadel L. (Ed.), The encyclopedia of cognitive science, Vol. 3, MacMillan, London, United Kingdom, pp. 47–51.

Nash J.E., Sutcliffe J.V. (1970) River flow forecasting through conceptual models, Part I - A discussion of principles, J. Hydrol. 10, 282–290.

National Acid Precipitation Assessment Program (1990) Evaluation of regional acidic deposition models and selected applications of RADM. Acid deposition: state of sciences and technology, The National Acid Precipitation Assessment Program, Washington, DC, USA, Vol. I, Report 5.

Norton J.P. (2003) Prediction for decision-making under uncertainty, in: Post D.A. (Ed.), Proc. MODSIM 2003 International Congress on Modelling and Simulation: Integrative modelling of biophysical, social and economic systems for resource management solutions, 14–17 July, Townsville, Australia, Vol. 4, pp. 1517–1522.

O'Keefe R.M., Balci O., Smith E.P. (1987) Validating expert system performance, IEEE Expert 2, pp. 81–90.

Oreskes N. (1998) Evaluation (not validation) of quantitative models, Environ. Health Persp. 106, 1453–1460.

Oreskes N., Belitz K. (2001) Philosophical issues in model assessment, in: Anderson M.G., Bates P.D. (Eds.), Model validation: perspectives in hydrological science, John Wiley & Sons, New York, NY, USA, Vol. 3, pp. 23–41.

Oreskes N., Shrader-Frechette K., Belitz K. (1994) Verification, validation and confirmation of numerical models in the earth sciences, Science 263, 641–646.

Parker V.T. (2001) Conceptual problems and scale limitations of defining ecological communities: a critique of the CI concept (Community of Individuals), Perspect. Plant Ecol. Evol. Syst. 4, 80–96.

Parker D., Manson S., Janssen M., Hoffman M., Deadman P. (2003) Multi-agents systems for the simulation of land-use and land-cover change: a review, Ann. Assoc. Am. Geogr. 93, 314–337.

Parrish R.S., Smith C.N. (1990) A method for testing whether model predictions fall within a prescribed factor of true values, with an application to pesticide leaching, Ecol. Model. 51, 59–72.

Pastres R., Brigolin D., Petrizzo A., Zucchetta M. (2004) Testing the robustness of primary production models in shallow coastal areas: a case study, Ecol. Model. 179, 221–233.

Patel V.C., Kumar A. (1998) Evaluation of three air dispersion models: ISCST2, ISCLT2, and Screen2 for mercury emissions in an urban area, Environ. Monit. Assess. 53, 259–277.

Pennell K.D., Homsby A.O., Jessup R.E., Rao K.S.C. (1990) Evaluation of five simulation models for predicting aldicarb and bromide behaviour under field conditions, Water Resour. Res. 26, 2679–2693.

Pilgram B., Judd K., Mees A. (2002) Modelling the dynamics of nonlinear time series using canonical variate analysis, Physica D 170, 103–117.

Pontius R.G. Jr. , Schneider L.C. (2001) Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA, Agr. Ecosyst. Environ. 85, 239–248.

Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. (1992) Numerical recipes in Fortran 77: the art of scientific computing, 2nd ed., Cambridge University Press, New York, NY, USA.

Prisley S.P., Mortimer M.J. (2004) A synthesis of literature on evaluation of models for policy applications, with implications for forest carbon accounting, Forest Ecol. Manag. 198, 89–103.

Quinn N.W.T. (2008) Integration of sensor networks and decision support tools for basin-scale, real-time water quality management, in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 1, pp. 44–53.

Reckhow K.H., Chapra S.C. (1983) Engineering approaches for lake management. Vol. 1: Data analysis and empirical modelling, Butterworth Publishers, Boston.

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

Refsgaard J.C., Henriksen H.J. (2004) Modelling guidelines-terminology and guiding principles, Adv. Water Resour. 27, 71–82.

Reynolds J.M.R., Deaton M.L. (1982) Comparisons of some tests for validation of stochastic simulation models, Commun. Stat. Simul. Comput. 11, 769–799.

Ricker W.E. (1984) Computation and uses of central trend lines, Can J. Zool. 62, 1897–1905.

Rivington M., Bellocchi G., Matthews K.B., Buchan K. (2005) Evaluation of three model estimations of solar radiation at 24 UK stations, Agr. Forest Meteorol. 135, 228–243.

Rivington M, Matthews K.B., Bellocchi G., Buchan K. (2006) Evaluating uncertainty introduced to process-based simulation model estimates by alternative sources of meteorological data, Agr. Syst. 88, 451–471.

Rivington M, Matthews K.B., Bellocchi G., Buchan K., Stöckle C.O., Donatelli M. (2007) An integrated assessment approach to conduct analyses of climate change impacts on whole-farm systems, Environ. Modell. Softw. 22, 202–210.

Rivington M., Matthews K.B., Buchan K. (2003) Quantifying the uncertainty in spatially-explicit land-use model predictions arising from the use of substituted climate data, in: Post D.A. (Ed.), Proc. MODSIM 2003 International Congress on Modelling and Simulation: Integrative modelling of biophysical, social and economic systems for resource management solutions, 14–17 July, Townsville, Australia, Vol. 4, pp. 1528–1533.

Robinson A.P., Ek A.R. (2000) The consequences of hierarchy for modelling in forest ecosystems, Can. J. Forest Res. 30, 1837–1846.

Rykiel Jr. E.J. (1996) Testing ecological models: the meaning of validation, Ecol. Model. 90, 229–244.

Sage A.P. (1987) Validation, in: Singh M.G. (Ed.), Systems analysis & control encyclopaedia: theory, technology, applications, Pergamon, Oxford, United Kingdom.

Sargent R.G. (2001) Verification, validation and accreditation of simulation models, in: Peters B.A., Smith J.S., Medeiros D.J., Rohrer M.W. (Eds.), Proc. 2001 Winter Simulation Conference, December 10–13, Arlington, VA, USA, pp. 106–114.

Schlesinger S. (1979) Terminology for model credibility, Simulation 32, 103–104.

Scholten H., van der Tol M.W.M. (1998) Quantitative validation of deterministic models: when is a model acceptable? in: Society for Computer Simulation (Ed.), Proceedings of the Summer Computer Simulation Conference, San Diego, CA, USA, pp. 404–409.

Schwartz G. (1978) Estimating the dimension of a model, Ann. Stat. 6, 461–464.

Seibert J., McDonnell J.J. (2002) On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multicriteria model calibration, Water Resour. Res. 38, 1241.

Seigneur C., Pun B., Pai P., Louis J.F., Solomon P., Emery C., Morris R., Zahniser M., Worsnop D., Koutrakis P., White W., Tombach I. (2000) Guidance for the performance evaluation of three-dimensional air quality modeling systems for particulate matter and visibility, J. Air Waste Manage. Assoc. 50, 588–599.

Shaeffer D.L. (1980) A model evaluation methodology applicable to environmental assessment models, Ecol. Model. 8, 275–295.

Sinclair T.R., Seligman N. (2000) Criteria for publishing papers on crop. modelling, Field Crop. Res. 68, 165–172.

Smith P., Smith J.U., Powlson D.S., McGill W.B., Arah J.R.M., Chertov O.G., Coleman K., Franko U., Frolking S., Jenkinson D.S., Jensen L.S., Kelly R.H., Klein-Gunnewiek H., Komarov A.S., Li C., Molina J.A.E., Mueller T., Parton W.J., Thomley J.H.M., Whitmore A.P. (1997) A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments, Geoderma 81, 153–225.

Sojda R.S. (2004) Empirical evaluation of decision support systems: concepts and an example for trumpeter swan management, in: Pahl-Woslt C., Schmidt S., Rizzoli A.E., Jakeman A.J. (Eds.), Complexity and integrated resources, Trans. 2nd Biennial Meeting of the International Environmental Modelling and Software Society, 14–17 June, Osnabrück, Germany, Vol. 2, pp. 649–655.

Soroshian S., Duan Q., Gupta V.K. (1993) Calibration of rainfall-runoff models: application of global optimization to Sacramento Soil Moisture Model, Water Resour. Res. 29, 1185–1194.

Stephens M.A. (1974) EDF statistics for goodness of fit and some comparisons, J. Am. Stat. Assoc. 69, 730–737.

Sterman J.D. (2000) Business dynamics: systems thinking and modeling for a complex world, Irwin McGraw-Hill, New York, NY, USA.

Stöckle C.O., Bellocchi G., Nelson R.L. (1999) Evaluation of the weather generator ClimGen for several world locations, in: Bindi M., Donatelli M., Porter J., van Ittersum M.K. (Eds.), Proc. 7th International Congress for Computer Technology in Agriculture, 15–18 November 1998, Florence, Italy, pp. 34–41.

Stöckle C.O., Kjelgaard J., Bellocchi G. (2004) Evaluation of estimated weather data for calculating Penman-Monteith reference crop evapotranspiration, Irrig. Sci. 1, 39–46.

Stone M. (1974) Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Ser. B-Stat. Methodol. 36, 111–147.

Sugeno M. (1985) An introductory survey of fuzzy control, Inf. Sci. 36, 59–83.

Tedeschi L.O. (2006) Assesment of the adequacy of mathematical models, Agr. Syst. 89, 225–247.

Theil H., Cramer J.S., Moerman H., Russchen A. (1970) Economic forecast and policy, 2nd ed., North-Holland Publishing Company, Amsterdam, The Netherlands.

Thomann R.V. (1982) Verification of water quality models, J. Env. Eng. Div. 108, 923–940.

Tingem M., Rivington M., Bellocchi G., Azam-Alia S., Colls J. (2009) Adaptation assessments for crop production in response to climate change in Cameroon, Agron. Sustain. Dev. 29, in press.

Topp C.F.E., Doyle C.J. (2004) Modelling the comparative productivity and profitability of grass and legume systems of silage production in northern Europe, Grass Forage Sci. 59, 274–292.

Trnka M., Eitzinger J., Gruszczynski G., Buchgraber K., Resch R., Schaumberger A. (2006) A simple statistical model for predicting herbage production from permanent grassland, Grass Forage Sci. 61, 253–271.

Trnka M., Zãlud Z., Eitzinger J., Dubrovský M. (2005) Global solar radiation in Central European lowlands estimated by various empirical formulae, Agr. Forest Meteorol. 131, 54–76.

Van Oijen M. (2002) On the use of specific publication criteria for papers on process-based modelling in plant science, Field Crop. Res. 74, 197–205.

Versar Inc. (1988) Current and suggested practices in the validation of exposure assessment models, Office of Health and Environmental Assessment, United States environmental Protection Agency, Washington DC, USA.

Vichi M., Ruardij P., Baretta J.W. (2004) Link or sink: a modelling interpretation of the open Baltic biogeochemistry, Biogeosciences 1, 79–100.

Vincent L.A. (1998) A technique for the identification of inhomogeneities in Canadian temperature series, J. Climate 11, 1094–1104.

Wainwright J., Mulligan M. (2004) Environmental modelling, Wiley & Sons, Chichester.

Wallace D.R., Fujii R.U. (1989) Software verification and validation: an overview, IEEE Software 6, 10–17.

Wallach D. (2006) Evaluating crop models, in: Wallach D., Makowski D., Jones J.W. (Eds.), Working with dynamic crop models, Elsevier, Amsterdam, The Netherlands, pp. 11–53.

Wallach D., Goffinet B. (1989) Mean square error of prediction in models for studying ecological and agronomic systems, Biometrics 43, 561–573.

Westrich B. (2008) Model based sediment quality management on river basin scale, in: Sànchez-Marrè M., Béjar J., Comas J., Rizzoli A.E., Guariso G. (Eds.), Integrating sciences and information technology for environmental assessment and decision making, Proc. 4th Biennial Meeting of the International Environmental Modelling and Software Society, 7–10 July, Barcelona, Spain, Vol. 1, pp. 639–646.

Whitmore A.P. (1991) A method for assessing the goodness of computer simulations of soil processes, J. Soil Sci. 42, 289–299.

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

Willmott C.J. (1982) Some comments on the evaluation of model performance, Bull. Am. Meteorol. Soc. 63, 1309–1313.

Woodward S.J.R. (2001) Validating a model that predicts daily growth and feed quality of New Zealand dairy pastures, Environ. Int. 27, 133–137.

Wright S.A. (2001) Covalidation of dissimilarly structured models, Dissertation, Air Force Institute of Technology, Dayton, OH, USA.

Wright G.G., Tapping J., Matthews K.B., Wright R. (2003) Combining metric aerial photography and near-infrared videography to define within-field soil sampling frameworks, GeoCarto International 18, 1–8.

Yagow E.R. (1997) Auxiliary procedures for the AGNPS model in urban fringe watersheds. PhD-Thesis, Virginia Polytechnic Institute, Blacksburg, VA, USA.

Yang J., Greenwood D.J., Rowell D.L., Wadsworth G.A., Burns I.G. (2000) Statistical methods for evaluating a crop nitrogen simulation model, N-ABLE, Agr. Syst. 64, 37–53.

Zacharias S., Coakley C.W. (1993) Comparison of quantitative techniques used for pesticide model validation, American Society of Agricultural Engineers. St. Joseph, MI, USA, ASAE Paper No. 93-2506.

Zacharias S., Heatwole C.D., Coakley C.W. (1996) Robust quantitative techniques for validating pesticide transport models, Trans. ASAE 39, 47–54.