Environmental Modelling and Software 159 (2023) 105559

Contents lists available at ScienceDirect

Environmental Modelling and Software

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

How to keep it adequate: A protocol for ensuring validity in agent-based simulation

Christian Troost[a,*], Robert Huber[b], Andrew R. Bell[c], Hedwig van Delden[d], Tatiana Filatova[e], Quang Bao Le[f], Melvin Lippe[g], Leila Niamir[h], J. Gareth Polhill[i], Zhanli Sun[j], Thomas Berger[a,k]

[a] *Hans-Ruthenberg-Institute, Universität Hohenheim, Wollgrasweg 43, 70597, Stuttgart, Germany*
[b] *Agricultural Economics and Policy Group ETH Zürich, Switzerland*
[c] *Earth & Environment Department, Boston University, Boston, MA, USA*
[d] *Research Institute for Knowledge Systems (RIKS), Maastricht, the Netherlands*
[e] *Multi Actor Systems Department, Faculty of Technology Policy and Management, TU Delft, the Netherlands*
[f] *International Center for Agricultural Research in the Dry Areas (ICARDA), Cairo, Egypt*
[g] *Thünen Institute of Forestry, Hamburg, Germany*
[h] *Energy, Climate, and Environment Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria*
[i] *The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK*
[j] *Leibniz Institute of Agricultural Development in Transition Economies (IAMO), Germany*
[k] *Hans-Ruthenberg-Institute, Universität Hohenheim, Stuttgart, Germany*

ARTICLE INFO

Keywords:
Model validation
Model inference
Calibration
Generalisation
Regime shift

ABSTRACT

There has so far been no shared understanding of validity in agent-based simulation. We here conceptualise validation as systematically substantiating the premises on which conclusions from simulation analysis for a particular modelling context are built. Given such a systematic perspective, validity of agent-based models cannot be ensured if validation is merely understood as an isolated step in the modelling process. Rather, valid conclusions from simulation analysis require context-adequate method choices at all steps of the simulation analysis including model construction, model and parameter inference, uncertainty analysis and simulation. We present a twelve-step protocol to highlight the (often hidden) premises for methodological choices and their link to the modelling context. It is designed to aid modelers in understanding their context and in choosing and documenting context-adequate and mutually consistent methods throughout the modelling process. Its purpose is to assist reviewers and the community as a whole in assessing and discussing context-adequacy.

1. Introduction

The increasing application of agent-based simulation models (ABM) for policy analysis in environmental and land system sciences, among other fields, has been accompanied by persistent calls to improve and formalise methods for their validation (Heppenstall et al., 2021; Elsawah et al., 2020; An et al., 2020; Niamir et al., 2020b; Brown et al., 2017; Filatova, 2015; Filatova et al., 2013; Heckbert et al., 2010; Marshall and Galea, 2015; Rand and Rust, 2011; Siebers et al., 2010; Midgley et al., 2007). These calls are motivated by the concern that an ABM must prove its ability to provide useful and reliable insight for solving real-world problems if it is intended to be more than a theoretically-appealing academic thought-instrument.

If we look at discussions of validation in simulation modelling in general, then traditionally *empirical validation*, i.e. comparing model predictions to observations of the behaviour of a real-world system, was regarded as the ideal method for showing relevance and reliability of a model (Oreskes et al., 1994). It entails reproducible protocols and quantitative, replicable and transparently communicable results. However, along with any type of model inference from observed system behaviour (*behaviour-based inference*)[1], it relies on (statistical) assumptions about the data and modelled system and can be severely misleading if these assumptions are not fulfilled in a specific research context (e.g. Oreskes et al., 1994; Polhill and Salt 2017). *Structural*

* Corresponding author.
E-mail address: christian.troost@uni-hohenheim.de (C. Troost).

[1] *Behaviour-based inference* comprises modelling steps typcially called estimation, calibration, data-driven model selection, inverse modelling or empirical validation (see section 2.1).

https://doi.org/10.1016/j.envsoft.2022.105559
Received 22 June 2022; Received in revised form 30 September 2022; Accepted 14 October 2022
Available online 20 October 2022

validation, for contrast, aims to ensure correspondence of the structure, processes and mechanisms within the model with their real-world counterparts. It is often limited by incomplete structural system knowledge and typically less formalised. When it is conducted as empirical validation of model component behaviour (structural behaviour validation, microvalidation), it is subject to similar statistical prerequisites as empirical behaviour validation at the macro level.

Recognizing that neither empirical nor structural validation can ultimately prove absolute correspondence of a model to reality and that models are by definition abstractions from reality (Oreskes et al., 1994; Quine, 1951) has led the scientific community to replace the condition for model validity from 'corresponds to the real system' to 'is adequate for its intended purpose' (e.g. Forrester and Senge, 1980; Gass, 1983; McCarl and Apland, 1986; Oreskes et al., 1994; Barlas, 1996; Kydland and Prescott, 1996; Rykiel, 1996; Beck et al., 1997; Jakeman et al., 2006; Deichsel and Pyka, 2009; Augusiak et al., 2014; Edmonds et al., 2019). This means that the conditions for a valid, i.e. adequate, model and simulation analysis are context-dependent. They do not only depend on the characteristics of the system to be modelled, but also on the availability of data describing the system and its behaviour as well as the research question to be answered.

Discussing the validation of ABM against this background, one first notices that ABM are used for a large variety of purposes and contexts (Edmonds et al., 2019; Lippe et al., 2019; Schulze et al., 2017; Ligmann-Zielinska et al., 2020). As inherently structure-rich models, they are often (but not always) used in contexts where data-driven modelling approaches are not applicable and as a consequence many prerequisites for empirical validation are not fulfilled (Berger and Troost 2014). The importance of structural validation, uncertainty and sensitivity analysis for ABM used in these contexts has been widely recognised (Moss and Edmonds 2005; Brenner and Werker, 2007; Augusiak et al., 2014; Troost and Berger 2015a; Marshall and Galea, 2015; Polhill and Salt 2017) and has even led some to dismiss empirical inference and validation of ABM altogether (Verhoog et al., 2016). Nevertheless, methods for behaviour-based inference (incl. empirical validation) of ABM have been developed for specific disciplinary contexts: For example, indirect inference of ABM in financial economics (Chen et al., 2012); pattern-oriented modelling as de-facto standard in ecological modelling (Grimm et al., 2005; Thiele et al., 2014); Approximate Bayesian Computation for inference of individual-based models (van der Vaart et al., 2015), micro-validation in energy economics (Niamir et al., 2020a), automatised calibration for innovation diffusion models (Jensen and Chappin, 2016) and real estate market interactions (Filatova 2015; Magliocca et al., 2016; de Koning and Filatova, 2020); or, robust inference of parameter distributions in agricultural economics (Arnold et al., 2015; Troost and Berger 2015a; Berger et al., 2017). Hence, agent-based modelling appears as a very diverse field, in which a multitude of methods for model construction, model inference, validation, evaluation and sensitivity analysis is being used and advocated. Unfortunately, the contexts in which specific methods are applicable are typically not explicitly discussed in general terms.

The ABM community has successfully addressed communication challenges caused by the diversity of modelling structures through adopting the ODD protocol (Grimm et al., 2010, 2020) for formal model documentation. The TRACE format (Schmolcke et al., 2010; Grimm et al., 2014) was suggested for documenting also hidden steps of the modelling process. However, a consensus or a formal protocol regarding which modeling methods to choose for a specific ABM application context that transcends disciplines has not yet been established, — not even within the more confined field of ABM in environmental and land system sciences (An et al., 2020; Polhill and Salt 2017; Filatova 2015).

This article[2] aims to fill this gap by formalising a framework for validation, i.e. a concept and guideline for ensuring and documenting the adequacy of an ABM and the simulation analysis for which it is used.

In the following section, we conceptualise validation as "challenging and substantiating the premises on which the conclusions from simulation analysis are built". We revisit premises of inference and validation typically used in simulation analysis in general and discuss to what extent they are tested, and to what extent they are actually presupposed by empirical and structural validation, behaviour-based model inference, uncertainty analysis and result interpretation.

It becomes clear that, given the diversity of contexts in which ABM are applied, it is not useful to prescribe one statistical or structural validation procedure to all ABM. What is more: under a paradigm of adequacy and given the constraints on empirical validation, validity cannot be tested solely by examining the behaviour or structure of the model, once it has been constructed. Validation cannot consist solely in one confined, isolated step of the modelling process - typically located after calibration and before predictive simulations - as which it is commonly still understood. Instead, validation, if understood as systematically examining the adequacy of a model for its purpose, requires careful justification of context-adequate and mutually consistent choices at all stages of the simulation analysis — including the choice of model components and choice of methods for model inference (inverse modelling, calibration, estimation, empirical validation)— and a consistent tracing, documentation and interpretation of uncertainties through the modelling process to finally ensure the validity of the conclusions drawn from the analysis.

On this basis, in the third section, we develop a step-by-step protocol of guiding questions to help agent-based modellers "keep it adequate" (KIA) by (i) defining the modelling context, (ii) adequately selecting models and methods for model inference and uncertainty documentation, and (iii) adequately deriving and interpreting simulation results and their uncertainty.

The fourth section discusses and concludes how the KIA protocol can help the ABM community. It is intended to (a) guide modellers during the research process, (b) provide a template structure for transparently documenting the rationale for modelling choices, (c) serve as a checklist for reviewers and stakeholders (addressees of simulation results) when assessing the validity of a documented study and its conclusions, (d) foster efficient communication between authors and reviewers, and (e) help in structuring the scientific discussion on the merits of choices regarding model selection, inference and evaluation made during the modelling process.

2. Arguments for model validity and their premises

If there is one cross-disciplinary consensus in the scientific literature on model validation, it is that model validity cannot be established in general, but only with respect to a specific purpose for which the model is intended to be used. Model validity is the adequacy of a model for its intended purpose (e.g. Forrester and Senge, 1980; Gass, 1983; McCarl and Apland, 1986; Oreskes et al., 1994; Barlas, 1996; Kydland and Prescott, 1996; Rykiel, 1996; Beck et al., 1997; Jakeman et al., 2006; Deichsel and Pyka, 2009; Augusiak et al., 2014; Edmonds et al., 2019). The purpose of any scientific simulation analysis is to answer a research question. Scientific answers result as conclusions from scientific argumentation and are accepted if the conclusions can be validly derived from accepted premises (McCloskey, 1983; Hands, 2001). Scientific objectiveness is ensured by transparently subjecting all premises and deductions to critical scrutiny and peer review (Klappholz and Agassi, 1959; Caldwell 1991; Longino, 1992).

In its most generic form, scientific arguments that employ simulation modelling conform to the following logical proposition (Troost and Berger 2020):

Major premise A: "If a simulation s fulfils conditions U and Results in y for inputs x, we can conclude Z.": $\exists s: U(s) \wedge R(s, x, y) \Rightarrow Z$.

Minor premise B: "Our simulation t results in y for inputs x and fulfils

[2] This article resulted from community discussions initiated at workshop W9 of the 2020 IEMSS conference.

conditions U.”: R(t, x, y) ∧ U(t).

Conclusion: “We conclude Z.”: ∴ Z by A ∧ B and modus ponens.

Premise B is a conjunction of two premises. The first premise “$R(t, x, y)$: Our model results in y for inputs x” is supported by result analysis. Showing that the second premise (“$U(t)$: Our simulation analysis fulfils conditions U”) holds is what is typically understood as validation.

A typical example: We conclude (Z) “Climate change will increase poverty among farming households” if $R(t, x, y)$: “Simulated farm agent income is lower in climate change scenarios than in the baseline”. The necessary condition $U(s)$ is very often formulated as: “The model employed in our simulation analysis provides sufficiently reliable predictions of $y(x)$ in the real-world system.” Empirical output validation and structural validation test whether a simulation t fulfils this (or a very similar) formulation of $U(s)$ but they, in turn, rely on further necessary premises.

These premises will be discussed in the following two subsections. The third subsection emphasises the role of uncertainty analysis for sound and robust conclusions (showing sufficient reliability). In the fourth subsection, we highlight that simulation analysis may also rely on differently formulated conditions $U(s)$ that allow for more useful conclusions in some contexts.

2.1. Premises of behaviour-based inference including empirical validation

The key underlying premise of any form of inference from the comparison of model and observed system behaviour is: “Predictive performance of a model in observed situations can be generalised to the target situations (i.e., the system situations relevant for the research question)”. This premise is trivially fulfilled if the target situation is part of the observed situations (*in-sample setting*). However, very often the simulation purpose is to anticipate[3] system behaviour for target situations (life after climate changed, in our example) that have not been (fully) observed or, in other cases, to find a generalisable model that explains mechanisms governing system behaviour in many target situations (explanation) (Edmonds et al., 2019).

Direct generalisation of behaviour (i.e., observed x-y relationships between system input and output including the strength of this relationship [= predictive performance]) from observed to unobserved situations relies on the two premises that the observed sample is redundant enough to control for sampling error and the target situations are part of a statistical population for which the observed sample is representative (*representative sample setting*). These basic statistical preconditions of representativity and control for sampling error apply to any form of model inference from observed behaviour (*behaviour-based inference, inverse modelling*), whether parameter values (estimation, calibration) or model structures (data-driven model selection) are selected, or predictive accuracy is estimated and compared to some (implicit) benchmark (goodness-of-fit evaluation) or between training and test samples (cross-validation)[4]: In all cases, ignoring sampling error and non-representativity (bias) leads to the generalisation of spurious, unsystematic, confounded or unstable relationships (overfitting) that causes inaccurate and misleading out-of-sample predictions and makes the inference invalid (Browne 2000; Forster 2000; Hansen and Heckman 1996).

Sampling error is the unavoidable, unsystematic error caused by using a sample and not the full population. It can potentially be reduced by increased sampling rates (Williams et al., 2022). Non-representativity occurs due to a biased sample, which can be caused by different, sometimes subtle reasons, including attrition, self-selection, survivorship or failure bias, observer bias, and unobserved heterogeneity (Vandecasteele and Debels 2007; Gangl 2010; Gormley and Matsa 2014; Jager et al., 2020; Smith 2020). While some minor biases may be corrected by statistical means, structural breaks, non-stationarity or regime shifts – such as climate change – substantially alter statistical x-y relationships causing extreme sample bias: Observed and target situations are so fundamentally different that they must be considered different statistical (sub)populations (*non-representative sample setting*) and direct generalisation is not possible (Perron 2006; Andersen et al., 2009; Leamer 2010; Filatova et al., 2016; Verstegen et al., 2016).

In non-representative sample settings, anticipation of system behaviour for unobserved situations has to rely on structural knowledge about internal system processes (see next section). Nevertheless, a sample can still be useful for *indirect generalisation*: Structural knowledge often admits alternative model formulations or parameter values (*candidates*). Even if a sample is not representative of the target situations, it can help discriminate between candidates if it is representative and sufficiently redundant for selected situations in which the candidates imply clearly distinguishable behaviour. Generalisation to a target situation then relies exclusively on structural knowledge embodied in the chosen candidate, whereas observed behavioural data only contributes indirectly by selecting this candidate.[5] Importantly, the predictive accuracy measured in the sample cannot be straightforwardly generalised to the target situation in these cases as even systematic differences in prediction errors between sample and target situations cannot be ruled out.

Preconditions for reliably discriminating between candidates are *structural* and *practical identifiability* (Bellman and Åström, 1970; Cobelli and DiStefano, 1980; Stigter et al., 2017; Guillaume et al., 2019): *Structural identifiability* means that different candidates are not *observationally equivalent*, i.e. do not imply the same system behaviour in the observed situations. Even a fully representative and redundant sample is not able to distinguish between models that predict the same output for the same input.[6] *Practical identifiability* means that the variation in the observational data in connection with statistical assumptions (e.g. on representativity and the form of model errors) is sufficient to unambiguously attribute effects to the individual parameters of a given model structure. Sampling error, confounded input variation (correlated variables, multicollinearity), unobserved heterogeneity, and omitted variable bias are key obstacles for unambiguous model selection and parameter estimation. More complex models require more data or more restrictive prior assumptions on parameters to be practically identifiable (Browne, 2000; Burnham and Anderson 2004; Polhill and Salt 2017). Two candidates that cannot be discriminated by given data are termed ‘equifinal’ (Beven and Freer 2001).

[3] This purpose can be called prediction, projection, scenario analysis, counterfactual simulation, forecast or just simulation depending on context (a more detailed discussion follows in section 3.1).

[4] The latter two are most often associated with the term “empirical validation”. Both are behaviour-based inference methods because they are used to select/accept a model by comparison to other models [sample averages in the simplest case, see section 3.2 and 3.3.1]. If not satisfied, the search typically continues until a better model is found. In terms such as ‘calibration & validation’, the second word typically refers to the second stage in a simple two-sample cross-validation. Within that cross-validation process the calibration and validation stage each have their separate roles, but together they constitute a method for model selection. This narrow meaning of validation is not to be confused with the comprehensive idea of validation as evaluating model adequacy for purpose advocated in this article (which involves the adequacy of a model selection/inference method).

[5] Similarly, indirect generalisation occurs if the output variable of interest has not been observed itself and a model is indirectly tested using another related output variable. Generalisation of the variable of interest then relies on the premise that the structural knowledge embodied in the model correctly relates the two variables.

[6] Structural identifiability in our understanding subsumes also the problems of endogeneity often encountered in econometrics.

2.2. Premises of structure-based model choice and structural validation

Structure-based simulation is essential to anticipate behaviour for target situations for which direct generalisation from observed data is not possible and to derive structural explanations of system behaviour. Structure-based simulation deduces system reaction from existing knowledge about system components and their interactions. It is sometimes argued that such a deductive process does not create new information. However, as Frisch (1933) argued, the key contribution of quantitative modelling is to analyse the interplay of processes and compare the magnitudes and directions of their individual effects in relation to each other in order to deduce the behaviour of the whole system. This anticipated or *emergent* behaviour is new information that was not obvious from looking at existing knowledge on individual processes in isolation.

The key premise of structure-based modelling and structural validation is: "A model that contains a sufficiently complete and accurate representation of the internal structure and processes of a system is expected to predict system behaviour well."

Structurally assessing the premise of *sufficient completeness* is often complicated by incomplete knowledge of the system and its potential reconfigurations. In addition, modellers are typically forced to strike a balance between completeness and efficiency — striving to include all relevant processes, while omitting unimportant ones that complicate the model construction (Forrester and Senge, 1980).

Assessing the premise of *sufficient accuracy* in the representation of individual processes is the subject of micro-validation (Moss and Edmonds 2005; Windrum et al., 2007; Midgley et al., 2007; Arnold et al., 2015; Ghaffarian et al., 2021). Some structural processes and their parameters may be directly observable and measurable. Others, however, may have been generalised from observed subsystem behaviour by behaviour-based inference, in which case the preconditions discussed in section 2.1 (sample representativity, identifiability and control of sampling error) apply: The inclusion of estimated model components into a composite model requires ensuring that the observations from which they have been generalised are representative for all contexts for which they are applied in the composite system.

2.3. Uncertainty analysis: the premises for robust conclusions

Given the statistical nature of model inference and the typically incomplete nature of structural knowledge discussed in the previous subsections, simulation analysis is practically always subject to uncertainty. Just showing that one particular model results in a specific output for a particular input is hence not convincing: It invites the immediate criticism that plausible alternative models might show different results. Rather, it must be shown that the final conclusions towards the research question are robust and not affected by uncertainty and bias (van Asselt, 2000; Walker et al., 2003; Saltelli et al., 2013; Fischhoff and Davis 2014; Berger and Troost 2014; Troost and Berger 2015a; Marchau et al., 2019).

This implies, firstly, that the type and degree of uncertainty and bias that are compatible with conclusion Z must be carefully specified in the major premise. Secondly, it is a necessary subpremise of $U(s)$ that implications of uncertainty in structural knowledge and uncertainty in model inference from data (and, in predictive analysis, uncertainty in the anticipated input for target situations) and their effects on results have been carefully assessed.

2.4. Alternative basic premises

Not every scientific argument using simulation analysis is based on the premise that the model provides reliable predictions of $y(x)$ in the real-world system. Edmonds et al. (2019) have noted that some types of analysis (e.g. theoretical exposition) do not require any immediate claims about the relation of the model to reality at all or put more emphasis in representing stakeholder's views of the system.

A subtler relation is discussed by Troost and Berger (2020, p. 6f.), who use the following hypothetical ABM application:

> "Economic policy analysis often works in a normative context: Policy makers need to justify actions with respect to established societal values, norms or ideologies. For example, they might work in a political setting, in which the state is supposed to safeguard minimum living incomes but only to interfere in economic processes if market participants are not at all able to help themselves.
>
> "Assume that in this context analysts build their ABM to simulate the adaptation of farmers to climatic change and model each farm agent decision as a rational optimisation problem with perfect anticipation of (projected) climatic impacts on production and market conditions. In addition, farm agents are embedded into a social network of mutual solidarity, in which agents less affected by climatic extreme events indiscriminately help the severely affected ones. Analysing their simulations, the analysts find that their optimising farm agents become food insecure under projected impacts. They conclude that if perfectly-foresighted, optimising agents in a perfectly functioning social solidarity network do not fare well, real-world farmers are even more unlikely to do so and should receive government help."

As Troost and Berger (2020) observe, the model would likely not pass conventional structural and empirical validation: Key modelled processes do not correspond to our best knowledge of their real-world counterparts. (In reality farmers do not behave as fully rational optimisers with perfect foresight and networks typically discriminate by family ties, ethnicity, etc.). The model will almost surely overestimate observed farm incomes in the past. Nevertheless, the conclusions would withstand such criticism, because accurately predicted farmer or network behaviour is not a relevant premise of the argument here.

In this case, the premise that would need to be challenged in validation is that the model calculates the best possible reaction in economic terms. Empirically this could be done, for example, by searching for observed cases for which the model predicts worse than observed outcomes. One might also identify other unexpected deviations, e.g. larger farm holdings having higher per-area incomes than smaller ones, which might be observed in the data but not in the model (or vice versa) and that are not expected to be caused by imperfect optimisation of real-world farmers alone. Nevertheless, even if the intention is not to show accurate prediction, premises on representativity, sampling error and identifiability also apply here. Structural validation could, for example, assess whether assumed constraints are overly pessimistic or alternative production, safety or income options that might become available with climate change have been omitted.

Troost and Berger (2020) further observe that if, for contrast, the analysts find that their computational agents fare well, it would be a logical fallacy to conclude that real-world agents will fare well based on the same premises. Such an argument would require different premises that are much more difficult to support using a model with a clear upward bias. Both cases use the same model in the same empirical context towards the same motivating research question. This illustrates that to judge a model's adequacy we require a very precise definition of its empirical context and the exact argumentative premise it is supposed to support.

3. A protocol for ensuring validity in agent-based simulation

Summarising section 2, validation means ensuring the adequacy of simulation analysis for answering a specific well-defined research question and such adequacy requires:

(I) laying out a logically valid argumentative structure on which potential conclusions from simulation towards the research question can be built;
(II) choosing model components and methods of inference and evaluation that (i) fit the requirements implied by this argumentative structure and (ii) rely only on preconditions regarding observation data, system properties and structural knowledge that are fulfilled in the given context;
(III) carefully assessing whether the simulation results and specifically their uncertainty and bias are consistent with the requirements of the argument.

Points (I)-(III) imply that adequacy is relative to a modelling context, which consists of the purpose (research question) and the available knowledge and data about the modelled system. Validity cannot be ensured by examining model structure and behaviour ex post only, ensuring it requires assessing the adequacy and mutual consistency of choices at all stages of the modelling process. Given the diversity of contexts in which ABM are applied, it will be impossible to identify one-fits-all model structures or statistical methods for all ABM and assessments of adequacy need to be able to cover a broad set of possible contexts.

Taking this into account, in the following sections, we propose a protocol (Fig. 1) of 12 steps covering and linking all stages of simulation analysis. The protocol helps characterise the modelling context (Part I, section 3.1), guides the choice of context-adequate methods based on this characterisation (Part II, section 3.2), and emphasises the documentation and consistent propagation of uncertainty through the modelling process, so that finally the robustness of conclusions can be comprehensively assessed (Part III, section 3.3). The protocol itself is provided in Tables 1–4 and 6, while the sections in the main text explain the rationale for each step. Where available, we list formal methods of analysis with useful references and highlight the premises for their applicability.

The concept of the protocol involves eleven dimensions (marked by letters a-k in Fig. 1) that characterise modelling contexts and determine adequate choices of models and methods. Six of these represent requirements of the research question (Fig. 1 a-f) that can be determined already at the beginning of the modelling process, while the other five (Fig. 1 g-k) require a more in-depth analysis of the relationship between research question and system knowledge and data during the modelling process. For a better overview, the numbering in Fig. 1 links the main stages of the modelling process (blue boxes) and context dimensions (grey boxes) to the associated steps of the protocol. The classification and propagation of uncertainty is indicated in red.

3.1. *Part I: defining the modelling context*

The first step is to characterise the modelling context: the precise research question and the knowledge and data that are available about the system being modelled (Table 1).

3.1.1. *Step 1: Precisely define the research question*

A research question typically arises from a larger debate, discourse, or decision problem: for example, a public, political or scientific debate, a participatory planning problem or an economic decision problem. A research question to be addressed by simulation analysis is supposed to contribute to this debate, even if answering it may not necessarily resolve the whole debate. Useful contributions can comprise very different questions (Edmonds et al., 2019; Epstein 2008): e.g. detailed, precise forecasts of future states of the world, statistical testing of explanatory models, but also exploring and stress-testing possible consequences of decision options (Berger and Troost 2014; Lempert 2019) or purely theoretical questions concerning hypothetical models themselves (theoretical exposition in the sense of Edmonds et al., 2019). It is paramount to be clear about what **precise question** the simulation analysis is supposed to answer and what precise argument it could contribute to the debate.

3.1.2. *Step 2: Characterise requirements implied by research question*

While typologies of model purposes exist (e.g. Edmonds et al., 2019; Epstein 2008), the understanding of commonly employed terms such as prediction, forecast, projection, exploration differs between scientific disciplines. Often, they are used inconsistently (Bray and von Storch, 2009), and all lack the necessary precision on some aspects relevant for methodological choices. Instead, Table 1 (a) defines six dimensions to precisely describe the requirements imposed by a research question: The most basic consideration is the *focus of interest:* Does this lie in anticipating system behaviour in specific situations[7] (output-focus) or in describing or understanding system structure[8] (structure-focus)? Carefully defining the *target situations* is a necessary precondition for judging the degree of generalisation in the next step. *Required resolution, required transparency* as well as *computational resource constraints* impose limits on *a priori* model selection. Judging the robustness of conclusions requires understanding the *required precision and accuracy* (tolerable uncertainty) in simulation outcomes. At this point, it is often not yet possible to formulate this quantitatively (e.g., 2% deviation is acceptable), and should be done in terms of consequences on conclusions (e.g., uncertainty should not affect ranking of policy alternatives by evaluation criteria). Together, these dimensions define requirements that the simulation analysis aims to fulfil. Whether this is actually possible can only be judged at the end of the modelling process (see section 3.3 and Table 7).

3.1.3. *Step 3: Identify knowledge and data about structure and behaviour of the modelled system*

In addition to the research question, the modelling context is defined by the available information about the simulated system in the form of structural and process knowledge, available observations of system behaviour (input-output trace data) as well as — in the case of an output-focus — the anticipated system input data for target situations. The next step is to identify which data, information and knowledge are available, can be obtained with reasonable effort or will remain unattainable for the analysis (e.g. input-output observations of far future system states) (Table 1b).

3.2. *Part II: Context-adequate model and parameter selection and uncertainty documentation*

Appropriate simulation models can be selected in two steps: In a first

[7] Such as in prediction, scenario analysis, counterfactual simulation, projection, or forecasts.

[8] Such as in explanation, causal identification, or description.

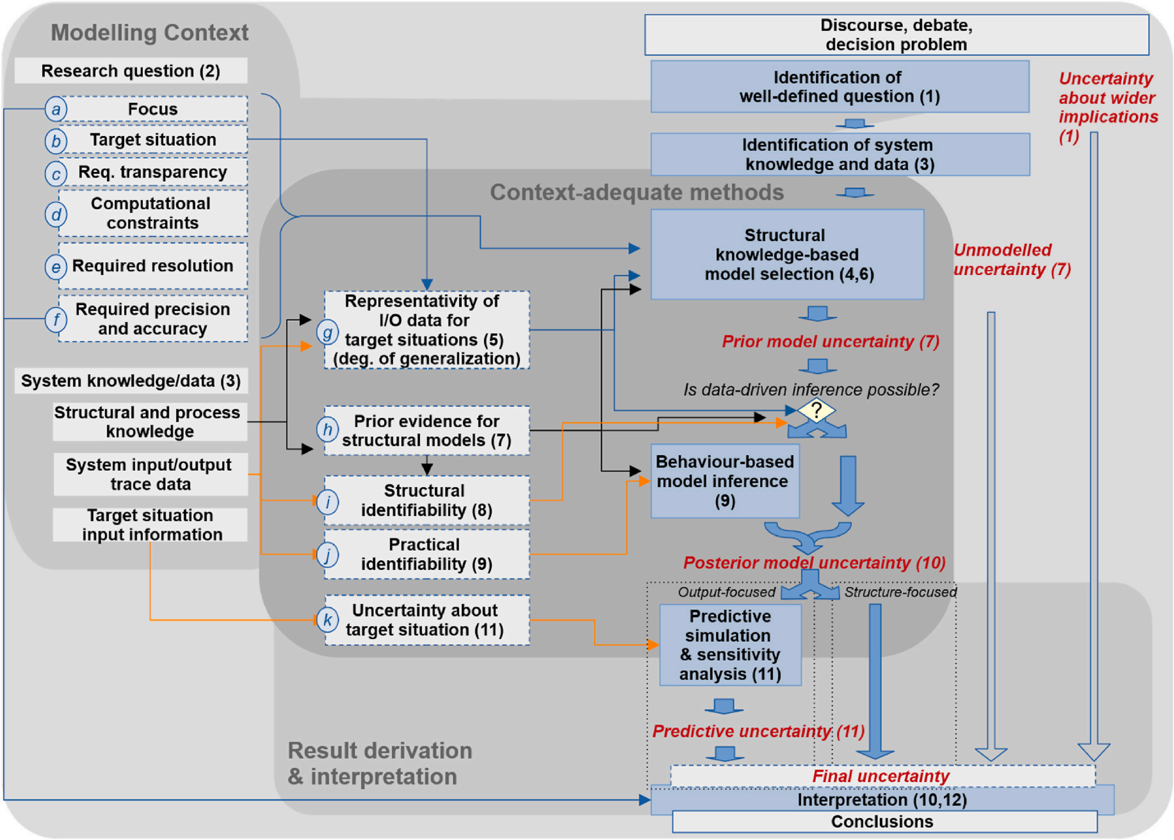
Modelling Context
Research question (2)
a Focus
b Target situation
c Req. transparency
d Computational constraints
e Required resolution
f Required precision and accuracy
System knowledge/data (3)
Structural and process knowledge
System input/output trace data
Target situation input information
Discourse, debate, decision problem
Identification of well-defined question (1)
Identification of system knowledge and data (3)
Uncertainty about wider implications (1)
Context-adequate methods
Structural knowledge-based model selection (4,6)
Unmodelled uncertainty (7)
g Representativity of I/O data for target situations (5) (deg. of generalization)
Prior model uncertainty (7)
Is data-driven inference possible?
?
h Prior evidence for structural models (7)
i Structural identifiability (8)
j Practical identifiability (9)
Behaviour-based model inference (9)
Posterior model uncertainty (10)
Output-focused
Structure-focused
k Uncertainty about target situation (11)
Predictive simulation & sensitivity analysis (11)
Predictive uncertainty (11)
Result derivation & interpretation
Final uncertainty
Interpretation (10,12)
Conclusions

Fig. 1. Tracing the influence of the modelling context on adequate decisions in and conclusions from simulation analysis. Conceptual basis and structural overview of the protocol. (Note: Numbers refer to steps in the protocol. Letters refer to the 11 characterizing dimensions of the modelling context. Blue boxes refer to stages of the modelling process. Uncertainty classification and propagation is printed in red. Arrows link context dimensions to the modelling stages in which they influence decisions. Colours of arrows help to visually trace crossing connections, but have no deeper significance.)

structural step, a set of candidate models and candidate parameter sets is constructed or identified whose theoretical characteristics comply with structural system knowledge and the requirements implied by the modelling context (Steps 4–6; Tables 2 and 3). A set of multiple candidates fulfilling the requirements represents the *prior model uncertainty*[9] (Steps 7–8; Tab 4). In a potential second step, behaviour-based inference can possibly be used to ascribe empirical likelihood to the candidates, rank them and narrow down the candidate set, reducing prior to *posterior model uncertainty* (Beck et al., 1997) (Step 9; Table 4).

Ideally, the two steps complement each other: The first step is key to ensure that only adequate candidates are considered in behaviour-based inference. Omitting this theory-based preselection can only be adequate if the simulation analysis is output-focused and the modelling context allows for the direct generalisation of statistical relationships (namely the expected predictive accuracy) to the target situations (representative and sufficiently redundant data) (Step 6i, Table 3). Only in this specific case, expected out-of-sample predictive accuracy and practical identifiability can be derived solely from the data and are sufficient criteria for model selection (Polhill and Salt 2017). Nevertheless, even for these direct generalisation cases, incorporating structural knowledge in chosen candidate models becomes more essential the scarcer the data: a defensible structure-based error model specification and pre-selection of candidate models increases practical identifiability (see e.g. Troost et al., 2022).

For the second step, it is key to ensure the adequacy of the inference process itself (Steps 7-9, Table 4). Do the necessary preconditions discussed in section 2.1 hold in the given modelling context? Is the specific method chosen appropriate for the context? Is uncertainty properly considered and documented? If not, behaviour-based model inference is clearly not adequate

3.2.1. Step 4: Representativity of data and degree of generalisation

The first step in model selection is to contrast the observed or observable data with the target situation of the research question to *determine the degree of generalisation and extrapolation implied* following the considerations on representativity and constancy (regime shifts, structural breaks, stationarity) discussed in section 2.1. This analysis requires a basic system conceptualisation (not yet a full conceptual model) that allows judging the system's degree of openness, internal stability, complexity, and stochasticity (Step 4, Table 2).

3.2.2. Step 5: ABM as composite models: structuring component context

While our protocol addresses modellers that are inclined to use an ABM, one question to ask in structural model choice is, of course, whether an ABM indeed suits the given modelling context or a different modelling approach is more promising. ABM are typically composite models (model systems), which are composed of lower-hierarchy models (components) that mirror relevant subsystems and processes. For example, they typically contain a model of individual agent behaviour based on the internal state of and external influence on the agent. This submodel for agent behaviour in turn may itself be a composite of lower-hierarchy components, e.g. for learning, demographics and economic

[9] While we use terminology (prior, posterior uncertainty) borrowed from Bayesian statistics here, this does not mean that this uncertainty can necessarily be cast into a formal prior probability distribution. More often than not, it cannot and it may well only be qualitative descriptions of uncertainty (cf. also Beck at al. 1997 for this general use).

Table 1
KIA Protocol, Part I: Guiding questions and categories for analysing and describing the modelling context and their relevance for the analysis.

Dimension	Questions	Categories	Relevance
Step 1: Define the research question			
Step 2: Analyse the research question and derive requirements for modelling			
Focus of interest	Are we interested in anticipating system behaviour for specific situations (*output-focus*)? Or are we interested in learning about inner system structure, processes, and relationships (*structure-focus*)?	*(i) output-focused* (for example, prediction, projection, scenario exploration) (ii)) *structure-focused* (for example, explanation, description, or causal inference)	*Output-focused analysis:* - if suitable representative data is available, predictions may be based on direct generalisation, in this case model structure can remain black box *(steps 4,5,6)* - uncertainty about inner system structure only relevant if it leads to uncertainties in prediction *(step 11)* *Structure-focused analysis* - transparency of model structure is essential - uncertainty about system structure relevant in itself, even if it does not lead to different predictions
Target situations	What are the target situations of our analysis? For which conditions do we expect our conclusions to hold? In output-focused analysis: for which situations do we want to predict outcomes? In structure-focused analysis: for which conditions do we expect our explanations to hold?	(i) only for *the model itself*, not necessarily in the real world (theoretical exposition) (ii) only for *a specific sample of observed data* (description, compression) (specify which) (iii) for a *well-circumscribed set of conditions/ target situations* (specify which) (iv) *universally: for any similar system in all possible situations*	- To be compared with the scope of observational data to determine the degree of generalisation *(step 4)* - To select an adequately comprehensive model in structural-knowledge-based model selection *(steps 5,6)*
Required interpretability of model structure	Does the inner structure of the model have to be interpretable, describable and communicable with 1:1 correspondence to real-world system elements (e.g. for stakeholder communication)? Or can it consist of trained statistical relationships or reduced form parameters that do not directly represent real-world system components (e.g. neural networks)? *(This is irrespective of transparent model documentation for peer review or structural requirements implied by out-of-sample predictions discussed in the following sections.)*	(i) communicable relationship to system components not needed (inner structure can be black box) (ii) requires communicable 1:1 relationship to real-world system components	If a model only needs to predict well, but not deliver a structural explanation, and if its predictive performance can be validly proved by sufficient representative observations, then transparency about the estimation process and predictive performance measurements are enough to prove its usefulness (*e.g.* machine learning models) while the model itself can remain opaque. In contrast, if it should serve for communication about system processes in stakeholder or educational contexts, its inner processes must be transparent and interpretable in real-world terms in any case. *For structure-based model selection (step 5, 6)*
Computational resource constraints	Does the model have to give an answer in a certain time frame, with a limited amount of resources?	Specify limits on time and resources for simulation if relevant	A sophisticated model with very high predictive accuracy and transparency is of little help if it cannot deliver the expected results with the available computational power in the necessary time frame. Different constraints may apply to model construction and estimation, on the one hand, and prediction, on the other hand. *For structure-based model selection (step 5, 6)*
Required resolution	a) Which spatial, temporal and socioeconomic resolution of simulation outputs is sufficiently disaggregated to answer the research question? Do statements about aggregates (e.g. watershed, regional totals or averages) suffice? b) Does the research question require exact statements about the exact behaviour of an individual entity (year Y, person X, location Z)? Or does it only require statements about the statistical distribution within a class of individuals?	Specify the *spatial, temporal, socioeconomic unit* of interest and the admissible level of statistical aggregation (totals or averages over individuals, statistical or probability distributions for classes of individuals or individual-specific values) at which reliable model outputs are required *(Note: This refers solely to the resolution at which reliable model predictions/outputs are required. It may later (step 6) turn out to be necessary to actually simulate at a different resolution for appropriate process representation, but this is not yet to be considered here.)*	- To be compared with the resolution of observational data to determine the degree of generalisation *(step 4),* - To be compared with effective resolution of candidate models in order to select an adequately detailed model and process representation in structural-knowledge- based model selection *(steps 5,6)*
Required precision and accuracy, acceptable bias	What degree of accuracy and certainty is required to derive a conclusive answer to the research question? *Relativity*: Do simulation results have to be accurate with respect to an *absolute* real-world reference (e.g. observed quantity, legally defined threshold, poverty line, etc.) or does it suffice if the *relative* position with respect to other model outcomes is simulated accurately (e.g. baseline vs policy intervention)? *Symmetry:* Will results be used for one-sided or two-sided comparisons? Is inaccuracy in one direction less tolerable than in another direction (*asymmetric*)? Is overestimation of the quantity as problematic as underestimation? *Conditionality*: Do statements about target	Specify required degree of accuracy and acceptable level of uncertainty for useful conclusions Examples: - Weather forecast: Predicted temperature should not miss actual one by more than ±2 K to allow a decision on what to wear (*absolute, symmetric, unconditional*) - Policy analysis: Simulation inaccuracy should not affect the preference order of suggested policies (*relative, symmetric, conditional*) - Vulnerability analysis: Simulated income should not systematically classify farm households as non-poor if they are poor (*absolute, asymmetric, conditional*)	- To be compared with theoretical model deviations in structural-knowledge-based model selection *(step 6)* - Relevant for the choice of loss function for direct generalisation cases *(step 9)* - To be compared with final posterior (predictive) uncertainty to judge the robustness of conclusions *(steps 10,12)*

(*continued on next page*)

Table 1 (*continued*)

Dimension	Questions	Categories	Relevance
	situations have to be unconditional or can they be formulated conditional on yet uncertain target situation input (e.g. scenarios, states of nature)? *Precision*: Which variance or level of accuracy is acceptable? Which deviations are tolerable without affecting conclusions? Do we require quantified error probabilities?		
Step 3: Analyse the available knowledge, information and data about the system being modelled			
Structural and process knowledge	How open and stable is the system studied? How complex and how stochastic is its response?		To determine representativity of data and degree of generalisation *(step 4)* and select a model *(step 6)*
	How well are system structure and processes known? How strong is the prior evidence for a specific system structure and process model?		For structure-based model selection *(step 6)* and specification of prior uncertainty *(step 7)*
Obtainable system input-output observations	Which observational data is available that traces system behaviour by relating system input and system output? Which could be obtained within the allocated time and resource frame? Which domain does it cover?	An overview and characterisation of the potentially available data	Used for assessing degree of generalisation *(step 4),* structural identifiability *(step 8)* and behaviour-based inference/practical identifiability *(step 9)*
Input data for target situation *(for output-focused RQ)*	How well can boundary conditions and initial system state (model input data, X) be anticipated for target situations?		Used to select an appropriate method for predictive analysis in *step 11*

Table 2
KIA Protocol, Part IIA: Guiding questions for basic considerations before structure-based model selection.

Item	Guiding questions and actions	Outcome
Step 4: Determine representativity observed/observable system behaviour and degree of generalisation		
Representativity of observed system behaviour for target situations *Potential type of generalisation*	*Given system understanding and available data from step 3:* Does the RQ imply extrapolation/generalisation from observed system behaviour? Do we have to expect (potentially) structural breaks, regime shifts, non-stationarities between observed and target situations? Can the data be considered representative of all target situations implied by the research question given the characteristics of the system? Has the external influence on the system been observed in all relevant dimensions and across the relevant domain? Have low probability events (likely) been observed in the sample (Filatova et al., 2016)? Does the sample suffer from bias or confounding with unobserved heterogeneity? Can it be corrected by weighting, error clustering, etc.? (Vandecasteele and Debels 2007; Gangl 2010; Gormley and Matsa 2014; Jager et al., 2020; Smith 2020) Considering the above, is direct generalisation of statistical relationships from observations to target situations potentially possible?	A well-argued decision for either of - *No generalisation implied* (if target situations fully contained in observed data) - *Direct generalisation potentially possible* (if data sufficiently representative for target situations, possible biases correctable, no structural breaks/regime shifts etc.) - *Direct generalisation not possible* (if data not representative e.g. due to structural breaks, regime-shifts, etc.) *(used in step 5 and 6 for structure-based model selection and step 8 and 9 to decide on behaviour-based inference)*
Step 5. Structure the modelling tasks into model components		
Component definition	Structure the modelling task into components (and functional links between them) (e.g. agent behaviour, interactions, environmental response, market mechanisms) Identify and characterise the specific modelling context of each component by running through steps 1–4 for each component. Possibly: Start with a tentative structuring and iteratively run through steps 1–6 or even steps 1–10 until finding a satisfactory structuring.	A structuring of the simulation task into components and the characterisation of the modelling context for each component *(basis for structure-based selection of component models in step 6 and possibly behaviour-based inference of these components in steps 8 and 9)*

decisions (Schlüter et al., 2017). ABM also typically contain models of agent interactions, e.g. communication, markets, auction, collective action or network models (Schreinemachers and Berger 2011). In addition, many ABM in natural resource management link to biophysical components that model responses of natural systems (e. g. a crop field or watershed) to agent intervention (Arnold et al., 2015).[10]

System behaviour in an ABM emerges not only from the interactions between agents, but conceptually also from the interactions of individual model components. In general, such structure-rich composite models are typically used for structure-focused analysis or for output-focused analysis when direct generalisation from observed data is not possible (Nolan et al., 2009; Voinov and Shugart, 2013). In direct generalisation contexts, prediction is often achieved more efficiently with statistical or machine learning models (Polhill and Salt 2017).[11]

The adequacy of a composite model relies on (i) an assembly of components that together fulfil the relevant premises for the overall research question to be answered, (ii) a careful assessment of the

[10] Whether the overall composite model is labeled as ABM or the ABM is itself considered part of the integrated composite is irrelevant. The discussed considerations apply in both cases.

[11] This does not imply that ABM cannot be used for direct generalisation contexts. There may often just be more efficient approaches.

adequacy of each lower hierarchy component for its intended role in the composite, and (iii) a consistent consideration of the uncertainty in each component at the composite level (Arnold et al., 2015).

It is important to realise that each component has its specific own question to answer and has its own specific modelling context, which may differ considerably from the modelling context of the composite as a whole or that of other components. Even if the ABM is used in an overall modelling context that is not apt for direct statistical inference, this does not rule out that within-model contexts of lower hierarchy components exist in which representative samples allow for direct generalisation and e.g. the use of machine-learning methods for these components. For example, we may not yet have observed how a specific group of farmers behaves and fares in a warmer climate, so we cannot empirically measure the predictive performance of a composite model that simulates potential future farmer behaviour and welfare. We may, however, be able to include a plant growth component into this composite model that can be tested based on observations and experiments in a range of warmer and colder regions if we consider this range representative for potential future growth conditions (Troost et al., 2020).

An important step hence is to structure the overall modelling task into subcomponents and then ***recursively revisit the steps of the protocol also for each component individually*** (Step 5, Table 2**)**. This step may often not directly result in the final structure, but may involve various iterations through steps 4–10 until an adequate composite structure for the overall modelling context has been established, which may or may not involve an ABM.

3.2.3. *Step 6: Choosing structurally adequate candidate models and prior parameter ranges for each component*

The guiding questions in Table 3 (step 6, items i-ix) help to check potential model (component) candidates for context-adequacy from a structural point of view. The table also lists selected literature sources that provide formal tests or more in-depth discussions of each question.

For adequate structure-based model selection, it is useful to first sketch a comprehensive conceptual system model (Argent et al., 2016), even if not all system processes can or finally have to be included in the simulation model. This conceptual sketch can serve as a benchmark to check a candidate's *match of the domain of applicability* and *sufficient completeness* of processes for the target situations (Parker et al., 2008). It must be ensured that model structure and parameters fixed in the candidate are also expected to be *constant* (no change over time) and *invariant* (unaffected by policy, treatment, change to target situation) in the real-world system (Lucas 1976; Engle and Hendry 1993; Hendry 1996). *Relevant changes between situations must be captured as exogenous input* or result from internal feedback in the model. It is not always possible to explicitly simulate all potential real-world feedback in the model itself, but it should then at least be possible to capture potential feedback as changing boundary conditions that may then later be assessed in uncertainty analysis (Troost and Berger 2015b; Troost et al., 2022) (Table 3, ii-iv).

Expected deviations, i.e. the part of the system behaviour that is not explained or predicted by the model from a theoretical point of view, should be consistent with the precision and accuracy required by the research question (Table 3, v). Research questions requiring accuracy with respect to an absolute reference necessitate not only a high degree of model completeness with respect to all systematic processes, but also with respect to probability distributions for unsystematic effects as well as reliable system input data for target situations. Research questions requiring accuracy only with respect to the relationships between simulated target situations demand model completeness only with respect to systematic differences. Simplifying assumptions (such as optimising agents in our example introduced in section 2.4) may lead to systematic over- or underestimation (*bias*). This is not problematic as long as major conclusions drawn from the simulation analysis will not depend on such simplification (robustness to the relaxation of simplifying assumptions, no model artefacts).[12],[13]

Logical consistency, correct technical implementation, and fit to the required resolution, transparency and resource constraints are obvious preconditions that must be assessed even if the component context allows for direct generalisation (Table 3, vi-ix).

3.2.4. *Steps 7 and 8: Documenting prior and input data uncertainty and assessing structural identifiability*

Structure-based model selection typically results in a number of plausible model structures and parameter values. This prior uncertainty should be documented (even if not all plausible alternatives can be implemented and tested) (Step 7, Table 4). The first step in determining whether behaviour-based inference can reduce this prior uncertainty then is to assess the structural identifiability of candidates in the observed range of data, i.e. check whether the behaviour of candidate models differs in the domain for which the data is representative (Step 8, Table 4). A variety of analytical and numerical approaches to assess structural identifiability are available (Guillaume et al., 2019; Chis et al., 2011) including numerical parameter screening methods from sensitivity analysis (Campolongo et al., 2007; Troost and Berger 2015a).

Not only uncertain parameters and structure in the model itself, but also uncertain auxiliary parameters or assumptions (e.g. error distributions for expected deviations and measurement error in input data, imputation to deal with incompleteness in the data,[14] alternative choices in data curation, preparation or aggregation) must be documented and considered when assessing identifiability. Structural identifiability may differ between parameters of the same model: Some parameters can be structurally identifiable in the available data (see Appendix A.1), while others are not and their uncertainty cannot be reduced by behaviour-based inference (e.g. Troost and Berger 2015a). Structural non-identifiability cannot be resolved by more of the same data, but requires either widening the range of situations observed or considering more dimensions of the data.

3.2.5. *Step 9: Choosing adequate methods for behaviour-based inference and measurement of predictive accuracy*

If structural identifiability is given or direct generalisation is possible, one can choose an adequate method for behaviour-based inference (Step 9, Table 4). If not, it is sometimes still useful to measure sample predictive accuracy of candidates and compare it against a null model to ensure the models do not completely go astray.

Behaviour-based inference requires choosing a loss function (a metric to weight deviations between observed and simulated behaviour) and an algorithm to characterise the distribution of the loss function over candidates (exploration/estimation of posterior parameter distribution) or find the candidate with the optimal loss function value (optimisation, calibration).

3.2.5.1. Adequate choice of loss function or likelihood. Loss functions (Step 9i, Table 4) are used to weight deviations between simulations and observations by severity. From a decision-theoretic point of view, loss functions should more strongly penalise those errors that would lead to

[12] The "Lucas critique" (Lucas 1976) is a famous example in economics for a challenge to modelling practice based on these grounds.

[13] Conclusions that are based on comparing model results to asymmetrical, one-sided thresholds even get stronger if the methodological approach is biased against them. Conversely, they are weakened by biases in their favour, especially if these cannot be precisely quantified and corrected. This principle mirrors the conservative rationale in statistical hypothesis testing: Type II errors, false-negatives, are preferred over type I errors, false-positives.

[14] A frequently encountered example in agricultural ABM would be a parameter used in imputing cash reserves of farm agents (which are typically unobserved or undisclosed) at simulation start from observed characteristics such as farm size, location, land use or livestock ownership.

Table 3
KIA Protocol, Part IIB: Checklist and formal methods for structure-based model selection and structural validation. The third column indicates selected literature sources for further reading that expand on the relevant theory or suggest formal tests for the assessment of the questions.

Step 6. Identify structurally adequate candidate models and (prior) parameter ranges for components by accepting or rejecting possible candidates based on the following checklist (using suitable formal methods listed in the third column if available)

Item	Guiding questions and actions	Formal methods
(i) Data or theory-driven approach?	Is the analysis output-focused (*step 2*) and no generalisation is implied or direct generalisation is possible (from *step 4*)? If yes, a data-driven model structure selection approach (or machine learning approach) can be chosen as long as it can also fulfil the transparency requirements (*step 2*) and sufficient data for practical identification is available. In this case, the checklist items marked with * can be skipped as reliance on statistical model structure selection methods such as cross-validation, AIC (*see step 9*) is sufficient. (One may still opt to go for a structure-based approach.)	
*(ii) Domain of applicability/ Structure and parameter constancy**	Do parameters and model structure represent relationships considered constant and stable across all relevant observed and target situations (*identified in step 2 and 3*)? Can all relevant differences between these situations either be formulated as external input or are endogenously simulated by the model? Can we expect the model to give correct results under extreme conditions?	- Domain of applicability/Identification of critical assumptions/ Parameter constancy and invariance (Hendry 1996; Alexandrov et al., 2011; Kloprogge et al., 2011; Fischhoff and Davis 2014; Rosenzweig and Udry 2016) - Extreme condition tests (Forrester and Senge 1980) Behaviour-sensitivity tests (Barlas 1996)
*(iii) Consistency with qualitative system knowledge**	Is the model formulation consistent with qualitative system knowledge to the extent required by the research question? Will it reflect any nonlinearity, non-additivity and asymptotic behaviour that we expect in the system? Does it remain realistic under extreme conditions?	- Structure-oriented testing/Behaviour-sensitivity tests (Barlas 1996) - Extreme condition tests (Forrester and Senge 1980) - Pattern-oriented modelling (Grimm and Railsback 2012) - Face validation, Stakeholder participation (Voinov and Bousquet, 2010; Voinov et al., 2016) - Turing tests (Barlas 1996; Rykiel 1996; Mössinger et al., 2022), Interactive modelling (Berger et al., 2010; Mössinger et al., 2022)
*(iv) Completeness/ Comprehensiveness**	Is the system representation embodied in the model comprehensive enough for the question? Can relevant system feedbacks be captured (at least in exogenous variables via uncertainty analysis)?	- Comprehensiveness in system representation: (Aumann 2007; Vester 2002) - Comparison with existing ontologies (Polhill and Salt 2017) or comprehensive conceptual frameworks (e.g. Le et al., 2012; Schlüter et al., 2017; Constantino et al., 2021) - Filtering by purpose and strong and weak patterns in behaviour (Grimm and Railsback 2012)
*(v) Expected deviations and robustness to simplifying assumptions**	Are the expected deviations (residuals, bias) of the candidate model *a priori* (from a theoretical perspective) consistent with the precision and accuracy (certainty, relativity, symmetry) required by the modelling context (*as identified in step 2*)?	
(vi) Match of effective resolution	What is the *effective* [a](temporal, spatial, thematic) *resolution* of the candidate? Does the *effective resolution* of the candidate model match the required resolution of the modelling context (*as identified in step 2*)?	- Aumann (2007); van Delden et al., (2011); Díaz-Pacheco et al., (2018); García-Álvarez et al., (2019).
(vii) Transparency and resource constraints	Does the candidate model match transparency, interpretability and resource use restrictions implied by the research question (*as identified in step 2*)?	
(viii) Logical consistency	Is the candidate model formulation in itself logically consistent?	- Face validation for logical errors; - Formal ontologies and ontology assessment tools (Polhill and Salt 2017)
(ix) Technical verification	Has the conceptual model been correctly implemented in computer code?	- Formal testing (see overview in Midgley et al., 2007); Unit testing (Onggo and Karatas 2016); Statistical debugging & trace validation (Gore et al., 2017); Model checking (Clarke et al., 2018)

[a] A spatial model may have a nominal map resolution of 1 ha grid cells, but the incorporated process understanding may reliably simulate only statistics over neighbourhoods of several cells (Pielke 1991; Laprise, 1992; Klaver et al., 2020). In this case, the effective resolution is the size of this neighbourhood. As an extreme example, consider the case when the spatial allocation in a nominally 1 ha grid model is purely based on land classes and all cells of the same class show the same behaviour (or just differ randomly following class-specific probabilities) without any further location or neighbourhood effects. The effective resolution is then ‘land class polygons’ and not ‘1 ha grid cells’. Similar considerations apply for temporal, thematic and ‘social’ resolution (e.g. individual, household, village, district).

Table 4
KIA Protocol, Part IIC: Guiding questions for model inference from observed system behaviour and documenting model uncertainty.

Item	Guiding questions and actions	Outcome
Step 7. Describe prior uncertainty comprehensively: List all candidate models, candidate parameters, error parameters and data uncertainty		
Documenting prior uncertainty	Which candidates for model structure and parameter values were identified in structure-based model selection (*step 6*)? Which parts of the candidates have to be considered uncertain and in principle adaptable/estimable using the data? Can this uncertainty be quantified as a prior probability distribution? Which additional uncertainty has to be considered and reflected as (potentially unstable) parameters during estimation (e.g. uncertainty in observations, imputation of data, alternative choices in data preparation, classification and aggregation, expected deviations)? Which potential candidates are ignored in the analysis (unmodelled uncertainty)?	List of model structures and parameter ranges used to represent model uncertainty in further analysis (and potentially estimated by behaviour-based inference) (*→used in step 8, 11*) List of auxiliary parameters used to represent data and data preparation uncertainty (*→ used in step 8, 11*) Ranges or, if available, prior probabilities for these models and parameters (*→step 8, 9, 11*) List of alternative models and parameter ranges theoretically suitable, but not explored in the analysis (*→ used in step 12*) List of critical assumptions for which no alternative assumptions will be tested during the further analysis (*→used in step 12*)
Step 8. Assess structural identifiability of candidate models in the population/domain represented by the observed sample *(possibly omit if data-driven model selection has been chosen in step 6 and appropriate methods for statistical model selection are used in step 9)*		
Structural identifiability	Is the difference between predictions of two candidates in the observed domain sufficient to distinguish them at a relevant order of magnitude? Are outcomes unique to a candidate or do different candidates produce the same outcome? *If not (not identifiable):* Can we employ additional relevant dimensions (variables) of the observed data? Can we subdivide the model into components/parameter groups that are identifiable? Can we reparameterise the model by aggregating unidentifiable ones to identifiable ones without violating structural knowledge on parameter stability? If yes, do and reassess identifiability.	List of parameters or model structures that cause detectable differences within the domain of the benchmark data available for model inference and are hence structurally identifiable (→ *step 9*) a) identified from a theoretical perspective (e.g. Guillaume et al., 2019; Chis et al., 2011) b) identified using specific sensitivity analysis to identify parameters that have an effect on those outcomes that can to be compared with observations (e.g. Campolongo et al., 2007; Troost and Berger 2015a)
Step 9: Choose and apply an adequate strategy for behaviour-based inference *(if direct generalisation or structural identifiability given, otherwise only for informal check of predictive accuracy)*		
(i) Choice of loss function/ acceptance criteria/predictive accuracy measure	*In direct generalisation cases and output-focus:* Which prediction errors would have the strongest effect on conclusions (*from step 2*)? Does the loss function appropriately reflect this? Does it focus on the relevant output variable? *In indirect generalisation or for structure-focus:* Does the loss function appropriately weight errors by the expected deviations of the model candidate (*from step 6*) including at least all variables considered for structural identifiability (*in step 8*)? Does it reflect expected bias, error patterns? Does it represent the systematic effects expected to be captured by the model? Does it appropriately consider the effective resolution of the model and data? (*all from step 6*) Consider formal likelihoods for well-specified models with tractable, well-defined error distributions. Consider robust loss functions for minor deviations from well-specified models and error distributions, e.g. outliers caused by unmodeled mechanisms Consider indirect likelihoods based on summary statistics or qualitative acceptance criteria if the exact form of the expected prediction error cannot be specified in a parametric form or outliers are likely.	A suitable loss function, likelihood or acceptance criterion which fits the context. For example: *Parametric likelihoods*: Schoups and Vrugt (2010); Hansen and Heckman (1996); Kukacka and Barunik (2017); Lux and Zwinkels (2018) *Indirect/Approximate likelihoods*: Chen et al., (2012); Beaumont (2010); Drovandi et al., (2015); Grazzini and Richiardi, 2015; Carrella et al., 2020 *Robust loss functions:* Willmott and Matsuura (2005); Troost and Berger (2015a) (ABM example) *Qualitative criteria*: Pattern-oriented modelling (Grimm and Railsback 2012, Gallagher et al., 2021); Binary acceptance (Spear and Hornberger 1980) *Landscape metrics* (as qualitative criteria or summary statistics in approximate likelihoods): e.g. Hagen-Zanker (2009); Chen (2011); Pontius and Millones (2011); Van Vliet et al., (2013); McGarigal (2014)
(ii) Benchmarking	Choose a proper benchmark/null model that reflects the best simple alternative model (e.g. sample average, random allocation, trend extrapolation) (Schaeffli and Gupta, 2007; Pontius and Millones 2011). Include it in the analysis either by explicit inclusion in the set of candidate models (Grimm and Railsback 2012) or implicitly by using it to calculate an absolute goodness-of-fit measure (model efficiency) from the loss function.	
(iii) Practical identifiability (a priori)	Can we at all expect the available data to be able to discriminate between the candidate model structures and parameter ranges? Are there enough degrees of freedom for the complexity of the model and assumed error terms? Does the data contain sufficient independent, unconfounded variation of input variables (absence of multicollinearity) so that main and interaction effects of input variables implied by candidate models can be disentangled (e.g. assess using variance inflation factors)? Is the whole domain well represented in the data or are we likely to have a strong influence of outliers?	A first quick assessment whether practical identifiability can at all be expected and it is worth to try model inference from the data.

(*continued on next page*)

Table 4 (*continued*)

Item	Guiding questions and actions	Outcome
(iv) Choice and application of an algorithm for behaviour-based inference	Does the chain of methods/algorithms chosen … a) … consider all (operational) alternative model formulations and parameter sets (*from step 6*)? b) … adequately consider prior evidence/probability of model structures and parameter values (*if available from step 7*)? c) … consider and deal with biases in *a priori* identifiability of models in a sample, e.g. using information criteria (AIC,BIC), k fold cross-validation? d) … quantify the effect of sampling error and the uncertainty in the inverse modelling process (e.g. in the form of confidence intervals, credible intervals, joint posterior parameter distributions, bootstrapping, cross-validation, by diagnostic tools such as VIF, Cook's distance, etc.)? e) … not rely on assumptions (e.g. certainty of model structure, well-specified likelihoods, practical identifiability) that are not fulfilled in the given context (cf. Table 5)?	Potentially: A strategy for the evaluation of posterior model uncertainty (potentially the identification of a best model), potentially combining various algorithms and diagnostic tools. Potentially: The result of applying this strategy to the candidate models and parameter values using the available system I/O observations Alternatively: the decision to not pursue behaviour-based inference and continue without being able to reduce prior uncertainty Potentially: The expected predictive accuracy of the candidate models in predicting situations for which the available I/O data is representative (possibly put in relation to the expected predictive accuracy of a simple benchmark).

stronger changes in conclusions. In direct generalisation cases and when sampling error has been controlled for (e.g. by cross-validation, see below), the measured loss can be directly generalised to target situations. Hence, in this case, one can choose a loss function that is limited to output variables of interest and whose weighting directly reflects the precision, accuracy, relativity and symmetry required by the research question (see Step 2) penalising misclassifications based on their practical implications (e.g. prefer models with stronger deviations overall, but high reliability in critical areas) (Manderscheid 1965; Berger 1980; McCloskey 1985; Farahmand et al., 2017; Manski 2019).[15]

In indirect generalisation cases and structure-focused analysis, loss functions must reflect the impact of model errors on our confidence that the candidate reflects underlying system processes. In this case, loss functions should reflect the expected deviations of the model including sampling error, model bias and error correlation (Schoups and Vrugt 2010) regarding all observed output variables linked to the modelled mechanisms[16]: Theoretically anticipated deviations of candidate models are considered less severe than deviations unlikely to occur if the model predicts according to its theoretically expected precision (Hansen and Heckman 1996; Blavatskyy and Pogrebna, 2010). For example, if a model is designed to predict an upper bound, underestimation of observations should be penalised, overestimation not.[17]

If the model is expected to be well-specified and implies a well-defined tractable stochastic error distribution, a parametric likelihood function can be formulated. Using parametric likelihoods in cases where their underlying assumptions are not fulfilled or in doubt leads to biased model selection and overconfident conclusions (Beven et al., 2008; Stedinger et al., 2008). Robust loss functions allow for occasional outliers potentially generated by processes not captured in the model. (Willmott and Matsuura 2005; Hyndman and Koehler 2006). If the model is expected to capture the essential systematic relationship, but the exact error distribution is unknown or intractable, summary statistics that capture relevant systematic relationships can be estimated on both, observations and model output. A loss function can then be applied to the difference in the summary statistics rather than the individual observations (*Classical* and *Bayesian indirect inference*: Chen et al., 2012; Beaumont 2010; Drovandi et al., 2015). Pattern-Oriented Modelling generalises this principle to incorporate more qualitatively described strong and weak statistical patterns (Grimm and Railsback 2012). In other cases, qualitative criteria are used to define binary-valued acceptance functions (Spear and Hornberger 1980; Troost and Berger 2015a).

Often, absolute goodness-of-fit measures (e.g. model efficiencies) are used instead of pure loss functions or likelihoods (Step 9ii, Table 4). While the latter provide a relative ranking between candidate models, but their absolute values are specific to the sample used, absolute goodness-of-fit measures don't change the relative ranking, but take the sample variance into account in order to allow comparison between models estimated from different samples (Bennett et al., 2013; Hauduc et al., 2015). Implicitly, efficiency criteria compare the evaluated model with a benchmark or null model that employs only basic information of the data. R^2 and Model Efficiency, for example, contain the sample average as a null model. This null benchmark should be carefully chosen. The sample average is only one possible choice. Trend extrapolation, random allocation, or seasonal or group-specific averages are often more adequate benchmarks (Schaeffli and Gupta, 2007; Pontius and Millones 2011). As an alternative, Grimm and Railsback (2012) suggest to always explicitly include a benchmark null model among the candidates.

3.2.5.2. Adequate assessment of practical identifiability and posterior uncertainty. It is paramount to document uncertainty in measured predictive accuracy and model rankings and to assess how reliable the data could discriminate between candidates (practical identifiability) (Step 9 iii, iv, Table 4). Methods for behaviour-based inference considerably differ in the extent to which uncertainty in the selection process is characterised and to which prior uncertainty is considered and it is important to select a (combination of) method(s) whose premises fit the application case (Table 5). For example, classical minimum-loss or maximum likelihood-based parameter estimation presuppose that both the likelihood and the model structure are certain and correctly specified and all considered candidate parameterisations are *a priori* equally likely (Stigler 2007). They identify one best fitting model and limit quantification of posterior uncertainty to confidence intervals for parameters. While large confidence intervals point to low practical identifiability, they cannot conceptually be interpreted as posterior

[15] In the direct generalisation case: If we are interested in predicting deforestation, for example, then we can focus on the ability of the model to predict changes from forest to some other land use, without caring whether it also correctly predicts the new land use or changes among non-forest land use classes. (We thank Judith Verstegen for this example.)

[16] In the indirect generalisation case: Even if we are only interested in predicting deforestation, but the mechanisms that we have to trust to anticipate developments in unseen situations are supposed to also determine changes in other land uses accurately, then deviations in predictions of these other variables also undermine our trust in predicting deforestation. Since we cannot assume that predictive accuracy on deforestation observed in the sample is the same in the future, this holds even if prediction of deforestation in the sample is accurate.

[17] Bayes estimators allow combining a loss function for relevant errors in model application with a likelihood for the posterior probability of the model (Bassett and Deride 2019).

Table 5
An exemplary selection of methods and measures used in model inference from observed system behaviour (inverse modelling) and their characteristics and premises.

Method or Measure	Purpose	Loss function/Metric for deviations	Prior evidence	Posterior uncertainty	Premises	References
Maximum likelihood estimation	- identify best parameter combination	Parametric likelihood	- prior evidence only reflected in choice of candidate models tested	- Identifies only a single best estimate - Confidence intervals indicate uncertainty of estimates, but not posterior distribution for parameters	- correct model structure - correct formal likelihood that corresponds to the expected deviation of models	Hobbs and Hilborn (2006); Kukacka and Barunik (2017); Lux and Zwinkels (2018)
Bayesian maximum posterior density estimation	- identify best model = model with maximum posterior density	Parametric likelihood	- Prior evidence formalised as prior probability	- Identifies only a single best estimate - credible intervals	- correct formal likelihood that corresponds to the expected deviation of models - quantifiable prior evidence	Bassett and Deride (2019)
Bayesian (point) estimator	- identify best model = taking into account posterior density & decision-theoretic loss function	parametric likelihood	- prior evidence formalised as prior probability	- identifies only a single best estimate, but taking possible relevant (e.g. economic) loss into account - credible intervals	- correct formal likelihood that corresponds to the expected deviation of models - quantifiable prior evidence	Bassett and Deride (2019)
Bayesian posterior density simulation	- estimate posterior probability distribution for parameters and candidates	parametric likelihood	- prior evidence formalised as prior probability (possible for parameters and model structures)	- identifies the full quantifiable posterior density	- correct formal likelihood that corresponds to the expected deviation of models - quantifiable prior evidence	Hobbs and Hilborn (2006); Hartig et al., (2011); Grazzini et al., (2017); Lux and Zwinkels (2018)
Information criteria (AIC; BIC; DIC WAIC)	- identify a collection of best models	parametric likelihood	- corrects for bias towards more complex models	- ranking of candidate models based on bias-corrected maximum likelihood estimates - no objective posterior distribution - decision thresholds for inclusion/exclusion remain subjective	- correct formal likelihood that corresponds to the expected deviation of models - maximum likelihood parameter estimates for each candidate model	Burnham and Anderson (2004); Ward (2008); Brewer et al., (2016); Vehtari et al., (2017); Yates et al., (2021)
Bayesian indirect inference (incl. Approximate Bayesian Computation)	- identify a collection of best models/parameter values - estimate posterior probability distribution for parameters and candidates	- binary tolerance between auxiliary statistic/model estimated from model output and auxiliary statistic/model estimated from observation (sufficient to know systematic effects to be predicted by the model, full error distribution not needed)	- prior evidence formalised as prior probability	- approximates the full quantifiable posterior density	- expected systematic effects are well captured by (potentially misspecified) auxiliary model/summary statistic - quantifiable prior evidence - comprehensive inclusion of all candidates	Beaumont (2010); Hartig et al., 2011; Drovandi et al., (2015); Grazzini et al., (2017)
Indirect inference (frequentist)	- identify best model	- distance function between auxiliary statistical model estimated from model and auxiliary statistical model estimated from observation (sufficient to know systematic effects to be predicted by the model, full error distribution not needed)	- prior evidence only reflected in choice of candidate models tested (uniform)	- identifies only a single best estimate - confidence intervals indicate uncertainty of estimates, but not posterior distribution for parameter	- expected systematic effects are well captured by (potentially misspecified) auxiliary model/summary statistic - correct model structure - comprehensive inclusion of all candidates	Chen et al., (2012); Grazzini and Richiardi, 2015; Lux and Zwinkels (2018)
Pattern- oriented modelling	- identify a collection of acceptable/plausible models/parameter values	- summary statistics that capture statistical patterns to be matched (different degrees of formalisation from qualitative criteria to Bayesian indirect inference)	- prior evidence only reflected in choice of candidate models tested (uniform)	- approximates the posterior distribution to different degrees of formalisation	- expected systematic effects are well captured by (potentially misspecified) auxiliary model/summary statistic - comprehensive inclusion of all candidates	Grimm and Railsback (2012), Gallagher et al., 2021

(continued on next page)

Table 5 (*continued*)

Method or Measure	Purpose	Loss function/Metric for deviations	Prior evidence	Posterior uncertainty	Premises	References
Rejection sampling with acceptance criteria	- identify a collection of acceptable/ plausible models/ parameter values	- binary acceptance criteria: acceptable and not acceptable performance (qualitative, quantitative, informal**)**	- prior evidence only reflected in choice of candidate models tested (uniform)	- collection of accepted models without explicit posterior probabilities	- expected systematic effects reflected in acceptance criteria - comprehensive inclusion of all candidates	Spear and Hornberger, 1980; Troost and Berger (2015a)
Normalised goodness-of-fit (Model efficiency)	- benchmark the predictive accuracy of model	- parametric likelihood or robust loss function	No		- loss function adequate to the form of deviations - meaningful benchmark model	Schaeffli and Gupta (2007); Pontius and Millones (2011); Bennett et al. (2013); Hauduc et al. (2015)
Cross-validation (K-fold/Leave-one-out)	- to be combined with other estimation method - correct for bias towards more complex models in any estimation technique -estimate effect of sampling error on selection/ estimation results	- depends on method with which it is combined	- depends on method with which it is combined	- Non-parametric estimate of effect of sampling error on estimates and predictive accuracy	- data is representative and sufficiently redundant for resampling - data points are conditionally independent	Arlot and Celisse (2010); Vehtari et al., 2017; Browne (2000)
Bootstrapping	- to be combined with other estimation method - estimate effect of sampling error on selection/ estimation results	- depends on method with which it is combined	- depends on method with which it is combined	- Non-parametric estimate of effect of sampling error on estimates and predictive accuracy	- data is representative and sufficiently redundant for resampling	Efron and Tibshirani (1997)
Structural risk minimisation in model selection (e.g. by Rademacher complexity bounds, Vapnik-Chervonenkis dimension)	- to be combined with other estimation method - limit the allowed complexity of the model given a sample	- depends on method with which it is combined	- depends on method with which it is combined	- calculate bounds on the out-of-sample generalisation risk of differently complex model structures - include only models with acceptable risk	- applicable in direct generalisation cases	Bartlett and Mendelson (2002); Arlot and Celisse (2010)

probabilities for parameters. Bayesian frameworks (Hobbs and Hilborn 2006) can overcome the latter limitations if prior probabilities are specifiable.

K-fold cross-validation[18] is the essential non-parametric method to quantify sampling error in estimated expected loss or predictive accuracy for unseen situations from a sample (Browne 2000; Arlot and Celisse 2010; Bennett et al., 2013; Vehtari et al., 2017). It should be combined with any of the basic inference methods and also avoids the complexity bias when model structures are uncertain: Selecting model structures purely based on predictive accuracy measured in one sample is biased towards models with a higher number of freely adaptable parameters, which increases the danger of overfitting. Adequate model inference requires correcting this bias, e.g. by k-fold cross-validation. Only when parametric likelihoods are applicable (see above), information criteria (AIC, BIC, DIC, WAIC) or formal Bayesian frameworks with appropriately specified prior likelihoods (Burnham and Anderson 2004; Ward 2008; Vehtari et al., 2017) provide an alternative.

Statistical diagnostics for influential observations (e.g. Cook's distance) and multicollinearity in the data (e.g. variance inflation factors) common in econometric analysis should complement the analysis of posterior uncertainty.

3.3. *Part III: Adequate derivation and interpretation of simulation results and uncertainty*

Fig. 1 illustrated how an adequate modelling process structures, quantifies and potentially reduces uncertainty: The definition of a research question divides *uncertainty regarding the research question* from *uncertainty about wider implications* in the debate. Theory-based model selection structures the uncertainty about the research question into *prior model uncertainty* (represented by different candidate model structures and parameter ranges), *input uncertainty* (uncertainty in boundary and initial conditions), *expected deviation* (error terms, bias, aleatory uncertainty) and *unmodelled uncertainty* (alternative models not included in the analysis,[19] processes that have been ignored, potential exogenous events not considered, non-formalised error terms, unforeseeable events, critical assumptions for which no alternatives are tested, etc.). If applicable and successful, behaviour-based inference potentially reduces *prior model uncertainty* to *posterior model uncertainty*. If discrimination of candidate models by data is not possible, the posterior uncertainty remains the same as the prior uncertainty.

In structure-focused analysis (description, explanation), the resulting posterior model uncertainty is already the final uncertainty to be interpreted for conclusions. In output-focused analysis (prediction, scenario analysis, exploration), posterior uncertainty and input

[18] The traditional separation of data into one training and one validation dataset is the most basic form of cross-validation, but is subject to sampling error itself. K-fold cross-validation is the more robust extension.

[19] Brenner and Werker (2007) emphasise an inclusion of "all logically possible" parameter values and model structures consistent with structural and empirical knowledge. We recognise that this is often not feasible in practice, however, this needs to be acknowledged as unmodelled uncertainty and appropriately discussed when deriving conclusions.

Table 6
KIA Protocol, Part III: Guiding questions for the derivation of predictive uncertainty and the interpretation of results.

Item	Guiding questions	Outcome
Step 10: Interpret posterior uncertainty and expected predictive accuracy *(if applicable)*		
(i) Interpreting expected predictive accuracy (if measured)	What is the effect of sampling error on predictive accuracy (measured e.g. via cross-validation, bootstrapping, post-regression diagnostics.) and how does it influence interpretation (considering step 2)? Is there a bias in predictions that points to systematic model error (do disaggregate analysis of residuals!)? How do model predictions compare with the benchmarks? Have the limits to generalisability been respected (e.g. statements only relative to models included in the analysis and within the bounds of representativity of the sample used)?	- An indication to what extent the models capture the observed variation in the sample of system behaviour and whether it shows systematic biases. - An estimate on the possible effect of sampling variance on measured predictive accuracy. - Possibly: A qualitative judgment on the predictive accuracy (high, low, sufficient, etc.) based on an explicit and well-justified benchmark scale (e.g. restricted to comparison to a null model, long-term experience with similar models in similar situations) and the required precision derived from research question (from step 2). *(all to be used in step 11 and 12)*
(ii) Interpreting posterior uncertainty and the results of model inference (if applicable)	Considering identifiability, posterior uncertainty (from step 9) and unmodelled uncertainty (from step 7): Does the posterior uncertainty – if measured in step 9 – provide complete information about the effect of sampling error and practical identifiability of candidates (considering choice of method in step 9)? Was it possible to reduce prior uncertainty through inverse modelling? Can candidates (model structures, parameter values) be eliminated because we can clearly rule them out as implausible or highly unlikely *a posteriori*? Were parameters identifiable? Which alternative model formulations must be considered plausible enough to include into further analysis?	*In structure-focused analysis:* An interpretation of the evidence about system structure, cause-effect chains or influential system input that could be gained through the analysis which properly reflects the associated posterior uncertainty and plausible alternative model formulations. (→ step 12) *In output-focused analysis:* A set of models/parameter distributions for use in subsequent predictive simulation that reflects posterior uncertainty and does not neglect plausible alternative models and parameter estimates (→ step 11)
Step 11: Choose a simulation design for and run predictive simulations and analyse predictive uncertainty *(if the analysis is output-focused)*		
Design of predictive simulation experiments	Does the chosen design globally and representatively consider the full posterior model uncertainty as well as (scenario) input uncertainty and assess its effect on predictive outcomes? Is a form of prediction resp. method of sensitivity or explorative analysis chosen that is consistent with the level of uncertainty in the model and scenario input (see Table 5)? Does the assessment of predictive uncertainty focus on the simulated quantities relevant to the research question? Does it focus on the degree of accuracy, precision conditionality, relativity and symmetry relevant to the research question (step 2)? (For example, in policy analysis does it focus on the robustness of the policy effect rather than the uncertainty in unconditional prediction?)	A design for and the outcomes of simulation experiments that … … focuses on quantities and accuracy relevant for the research question (from step 2) … controls for the effect of aleatory uncertainty (e.g. by common random numbers schemes, e.g. Troost and Berger 2016, convergence over a large number of repetitions, assessments of case-wise or stochastic dominance)? … and … … covers the uncertainty space globally and representatively (Saltelli and Annoni, 2010) at a sampling rate adequate for the computational resources. (Consider efficient designs such as Sobol' sequences or LHS, see Tarantola et al., 2012) … or alternatively a comprehensive search for non-robust outcomes or strong deviations over the global uncertainty space (e.g. destructive verification, Midgley et al., 2007; stress testing and red-teaming, Lempert 2019). If uncertainty is nonnegligible, conduct global sensitivity analysis to detect which uncertain input factors have highest influence on output uncertainty (Helton et al., 2006; Campolongo et al., 2007; Saltelli et al. 2008, 2019; Borgonovo and Plischke 2016; Ligmann-Zielinska et al., 2020; Puy et al., 2021)
Step 12: Final interpretation, derivation of conclusions and documentation		
Conclusions	Is the communication of simulation outputs consistent with the level of uncertainty in model and scenario input (see Table 5)? Comparing the final predictive resp. posterior uncertainty (step 11, resp. 10) and the unmodelled uncertainty (step 7) with the precision and accuracy required by the research question (step 2): Which conclusions are possible? Are all the premises underlying the final conclusions clearly laid out (including assumptions on system complexity, alternative models, identifiability, representativity, error models etc.) and substantiated using the criteria set out in the previous steps? Is the posterior/ predictive uncertainty fully documented and discussed? Which of these premises are critical to maintain the conclusions? Does any theoretical or measured bias weaken or strengthen conclusions? Is there a clear delineation between what has been modelled with respect to the targeted question and the analysed target situations and what is further speculation in the context of the wider debate but not solely based on the discussed simulation analysis?	A summary of the results of running through the protocol explaining … … the purpose of the analysis and model (e.g. for the introduction of an article and the purpose section of the ODD protocol) … a summary justification of model and method choice following the steps, criteria and premises set out in the previous steps of this protocol (e.g. for the Methods & Results sections of an article, or for the Appendix) … the conclusions building on the comparison of model results and final uncertainty to research question requirements (e.g. for the Discussions and Conclusions sections of an article) … a documentation of prior (step 7), posterior (step 10) and predictive uncertainty (step 11), sensitivity to inputs (step 11) and specifically unmodelled uncertainty, i.e. critical and potentially value-laden assumptions for which plausible alternative assumptions could not be comprehensively tested in the analysis (step 7), e.g. following the schemes of NUSAP (van der Sluijs, 2017; Kloprogge et al., 2011), sensitivity auditing (Saltelli et al., 2013) or Fischhoff and Davis (2014)'s protocol. (for the Results & Discussions section of an article or as an extra document for policy advice).

uncertainty still need to be translated into *predictive uncertainty* for target situations (e.g. future or policy scenarios) by simulation experiments that include uncertainty analysis.

In an adequate modelling process, in which uncertainty is properly analysed and propagated, the final posterior/predictive uncertainty and the unmodelled uncertainty describe the actual state of knowledge regarding the research question that can be defensibly extracted from the available data and structural system knowledge. This final model uncertainty can then be compared with the precision required by the research question for interpretation and derivation of conclusions.

Table 7
Adequacy of different types of predictive analysis depending on systematic and unsystematic model uncertainty and uncertainty in system input for target situations (scenario uncertainty), adapted and extended from Marchau et al. (2019).

Types of uncertainty				Use of predictive simulation analysis		
Systematic (posterior) model uncertainty	Aleatory/ unsystematic model uncertainty	Scenario (input/ boundary) uncertainty	Level of uncertainty according to Marchau et al. (2019)	Adequate type of predictive analysis	Simulation outcomes	Decision strategies
Very low	Very low	Very low	1	Unconditional prediction	The deterministic (or overwhelmingly probable) outcome	Simple deterministic decision
Low or Probabilistic	Probabilistic	Probabilistic	2	Probabilistic forecast	List of possible outcomes with probabilities for each outcome	Expected utility theory, traditional risk management
Medium (a small number of alternative system models)	Medium, probabilistic or specifiable	Medium (a few specifiable scenarios)	3	Conditional prediction (projection)	A limited number of possible outcomes for a few different possible states of nature without probabilities for each state of nature	Traditional scenario analysis and uncertainty/sensitivity analysis, robust policy choice
High	High	High	4	Exploration	Multiple possible outcomes for many different possible states of nature with unknown probabilities and without being able to explore all relevant states of nature	Strategies for robust decision making under deep uncertainty (assumptions-based planning, read-teaming, etc.) Marchau et al. (2019); Lempert (2019)

3.3.1. Step 10: Interpretation of predictive accuracy and posterior uncertainty

If sampling error has been properly controlled for (e.g. by cross-validation), expected predictive accuracy indicates how well the model predicts or explains the variation in the population of situations for which the sample is representative (subject to the importance weighting embodied in likelihood or loss function). This is valuable information in its own right. However, whenever using this information to draw further conclusions (Step 10i, Table 6), e.g. about the model being "sufficiently good" or the "correct" or "best explanation", care has to be taken (Oreskes et al., 1994). Even though absolute goodness-of-fit measures such as model efficiencies project predictive error onto an absolute scale between null model and perfect fit, defining any threshold to indicate 'sufficient fit' on this scale remains subjective or based on convention – similar to significance levels in statistical analysis – unless this threshold can be convincingly derived from the research question (Pontius and Millones 2011). The same holds for thresholds defined on posterior densities or relative differences in information criteria (Stephens et al., 2005).

The well-known problems of induction, under-determination and theory-ladenness imply that proving by comparison to observation that a model is the 'true' model is ultimately impossible (Oreskes et al., 1994; Quine, 1951). Expected predictive accuracy provides a relative ranking and allows identification of the "best" among the candidate models for the given sample. The more comprehensive the list of candidate models and parameterisations that has been tested and the more representative the sample, the higher can be the confidence in having identified a generalisable best model or parameterisation. As all other statistical relationships, measured expected predictive accuracy cannot be generalised to target situations across structural breaks.

Uncertainty in inference can be quantified as a posterior probability for the candidates if a formal Bayesian framework with proper prior probabilities and appropriate likelihood has been used in inverse modelling. However, also in those cases where posterior probabilities or credible intervals cannot be derived, it is important to consider posterior uncertainty (Step 10ii, Table 6) and recognise that the "best" model does not necessarily have or even approach a posterior probability of one (Troost and Berger 2015a). The potential explanatory and predictive power of alternatives should not be neglected in interpretation. If the analysis is structure-focused and interested in which model provides the better explanation, it remains inconclusive whenever two alternative models cannot be robustly discriminated by data or needs to employ additional theoretical considerations, e.g. parsimony as an epistemological principle[20] or correspondence to established theory, to justify a decision for one or the other model. In output-focused analysis, subsequent predictive simulation should use the full posterior distribution, consider confidence or credible intervals or at least a representative ensemble of all candidates that show nonnegligible explanatory power (ensemble modelling, model averaging).

3.3.2. Step 11: Analysis and interpretation of predictive uncertainty

Only in rare cases, it will be permissible to directly generalise expected predictive uncertainty from behaviour-based inference to the target situation (preconditions: representative sample, negligible input uncertainty, one clearly best model). Generally, predictive uncertainty for a target situation is a function of the uncertainty about the systematic effect of system input on behaviour that is captured in the set of models and parameterisations (posterior model uncertainty), the model error (bias and unsystematic aleatory uncertainty) and the uncertainty in system inputs (e.g. scenarios, boundary conditions) for target situations. Building on the considerations by Marchau et al. (2019) and Walker et al. (2003), Table 7 lists which forms of predictive simulation outputs are adequate depending on the level of uncertainty in each of these dimensions. Unconditional predictions require low uncertainty in all "locations" of uncertainty. For all higher levels of uncertainty, comprehensive uncertainty analysis is necessary (Step 11, Table 6). Depending on model complexity and available computational resources, one can choose from a considerable number of approaches for efficient uncertainty and sensitivity analysis[21] (Helton et al., 2006; Saltelli et al., 2008; Gramacy and Lee 2009; Troost et al., 2022). Clear conditions for appropriate choices have been formulated: Uncertainty analysis must be global, i.e. cover the full range of potential input values including interactions and correlation between input factors (Saltelli and Annoni, 2010). Probabilistic predictions require probability information in all locations. It is key that exploration of predictive uncertainty focuses on the output quantity, precision, and resolution relevant to answering the

[20] Parsimony as a epistemological principle (simpler models are always to be preferred) differs from a pragmatic argument for parsimony in estimating models for prediction (simpler models are less prone to overfitting).

[21] Following the definition of Helton et al. (2006), uncertainty analysis is concerned with quantifying the uncertainty (variance) in simulation outputs, while sensitivity analysis is concerned with linking this uncertainty to uncertainty in model inputs, i.e. determine which uncertain input factors are responsible to which degree for the uncertainty in outputs.

targeted research question. When we compare two target situations, we can distinguish the *apparent* (*or observable*) *difference,* i.e. the difference between two predictions that includes unsystematic, stochastic effects, and the *systematic difference*, i.e. the difference between two predictions controlled for unsystematic effects. In many decision support situations, the future may not be precisely predictable, but for a good decision it is enough if the systematic differences caused by decision options can be pointed out using pairwise comparison at each tested combination of input factor values (Berger and Troost 2014). For stochastic models, this requires Common Random Numbers schemes (Stout and Goldie, 2008; Troost and Berger 2016). The alternative is running sufficient repetitions and applying statistical comparison tests (e.g. Verstegen et al., 2019).[22] Especially when uncertainty is high in all locations, rather than trying to merely describe all possible outcomes, strategies to detect decision options that are robust under many different scenarios and assumptions should be emphasised (assumptions-based planning, stress testing, red teaming; Lempert 2019; Marchau et al., 2019).

3.3.3. Step 12: Interpretation and conclusions

The interpretation of results should compare the final uncertainty (Step 10 or 11) to the required precision and accuracy of the research question (Step 2). If the required certainty is reached, conclusions that are consistent with the simulated output can be considered valid and sound. If uncertainty is too high, we have to conclude that the knowledge employed in the process is insufficient for the desired type of conclusions (e.g. Carauta et al., 2021). It should not be necessary to emphasise that this is an equally valuable and relevant result (Leamer 2010).

The structure of the argument and the premises that are critical to support the conclusions must be clearly laid out (Step 12, Table 6). This involves the premises that are supported by simulation results, but also the auxiliary and hidden premises (prior model evidence, representativity of data, identifiability, posterior uncertainty).

Both, unstructured uncertainty about wider implications (Step 1) and unmodelled uncertainty (Step 7) remain qualitative and unquantified in the modelling process. Nevertheless, they must be an important part of the interpretation: Conclusions must be qualified with respect to the information omitted from the modelling process. Hypotheses on how omitted processes or alternative system conceptualisations could affect conclusions must be discussed (Forrester and Senge, 1980). Banerjee et al. (2016) argue for an explicit and structured section for ‘Speculation’ about external validity (generalisability) of results obtained from case studies. Especially, when using models to inform decision-makers in the face of deep uncertainty, transparent documentation of critical and potentially value-laden fundamental assumptions (see protocols in Kloprogge et al., 2011; Saltelli et al., 2013; Fischhoff and Davis, 2014; van der Sluijs, 2017) and additional effort to assess the robustness of decision option outcomes to these assumptions is essential (Lempert 2019; Marchau et al., 2019).

4. Discussion and conclusions

The purpose of validation is to ensure the adequacy of simulation analysis for answering a specific well-defined research question. This requires a careful analysis of the logical argumentative structure and assessment of the critical premises that conclusions from simulation analysis build upon. Such premises rest on simulation outcomes, but are also implicit in the choice of models and methods of inference from data. Especially the latter is not always obvious to modellers, reviewers, and addressees of simulation results. For example, empirical validation and model inference presuppose representativity of data, identifiability, and control of sampling error. Moreover, specific methods such as maximum likelihood estimation rely on even more restrictive, not always obvious premises (see Tables 5 and 7). Validation needs to ensure that models and methods chosen fit the modelling context, which comprises the research question and available system knowledge and data on system behaviour. And it needs to assess whether the final uncertainty in simulation results fits the requirements on precision and accuracy implied by the research question.

In most cases this is more complex and subtler than a single-step matching of context to a method. Rather it is a *hierarchical* process, i.e. outcomes of earlier steps affect choices in later steps (e.g. behaviour-based inference should not be pursued without first ensuring representative data and structural identifiability). It is *recursive*, i.e. in composite models such as ABM the context of each component must be assessed, and *iterative,* i.e. outcomes of subsequent steps may encourage receding a number of steps and reconsidering choices: For example, if the evaluation of structural identifiability, practical identifiability or predictive uncertainty leads to unsatisfactory results, it may be useful to go back to structure-based model selection or even to a redefinition of the research question: It may be possible to answer a more restricted question that is already useful, where the context does not allow to reliably answer the original question.

The KIA protocol that we have proposed in this article is intended to guide modellers in making adequate choices during the process of simulation analysis and justify them with adequate argumentation. It provides a guideline to reviewers who can use it by starting from the final conclusions and their premises, and working backward to evaluate whether the steps taken during the modelling process adequately support the premises in the given context. Moreover, it is intended to structure documentation: (i) as a checklist to ensure modelling context and justification for all relevant modelling decisions have been discussed in the main body of an article and (ii) as a template for well-structured tabular documentation in an appendix.

The protocol mirrors and is compatible with established recommendations for a structured modelling process (e.g. Jakeman et al., 2006), but it emphasises the linkages and propagation of uncertainty between modelling stages and highlights general criteria for the choice of adequate methods at each stage. It operationalises the principle “as empirical as possible, as general as necessary” coined for ABM by Brenner and Werker (2007). It incorporates the different levels of uncertainty of Walker et al. (2003) and Marchau et al. (2019), but also explains how this uncertainty comes about in the modelling process. Similar to Polhill and Salt (2017), it highlights the importance of structural model choice compared with purely data-driven model inference. While we have not extensively discussed stakeholder participation, the protocol is meant to be open to valuable stakeholder input and feedback at any step of the process: e.g. in shaping the encompassing debate, defining the targeted research questions, providing information in model selection and inference and shared interpretation (Voinov et al., 2016; Barreteau et al., 2010).

The exhaustive discussion of many of the guiding questions listed in the tables of the protocol would warrant their own articles. Our intention here has been to comprehensively list them and highlight their interlinkages. We have linked many of the guiding questions to literature with more detailed explanation or formal assessment methods. This list of methods does not claim to be complete and it will certainly have to be extended over time as new approaches for model testing, selection or estimation are developed to deal with the formulated questions. We actually hope that this protocol sparks interest in developing new methods and then assists in clearly communicating the conditions for which they are suitable.

In defining eleven dimensions for the characterization of modelling contexts, we have moved beyond discrete typologies of model purpose (e.g. Edmonds et al., 2019; Epstein 2008). Typologies, such as Edmonds et al. (2019), and especially terms such as prediction, forecast, projection or exploration, whose understanding and usage differ between and

[22] Common random number schemes are more efficient in terms of required model runs, but sometimes quite difficult to implement (see example in Troost and Berger 2016).

sometimes even within disciplines (Bray and von Storch, 2009), can be mapped onto these dimensions to allow for more precise communication (see Appendix A.2). The dimensions are intended to improve communication on methodology by helping to identify which ABM applications share a similar modelling context and might learn from each other and which not. For example, Troost and Berger (2015a) and Carrella et al. (2020) both deal with unknown or intractable likelihoods for model inference. However, the former face both low structural and practical identifiability, while the latter assume few parameters and a large number of identifying summary statistics, i.e. high practical identifiability. As both are explicit about the assumed modelling context, this can be read from their articles, but may still be easily overlooked. Our protocol is intended to highlight these differences and in this way avoid common pitfalls in discussions between modellers and reviewers about adequate and valid model use and inference: e.g. avoid discussions about an appropriate loss function, when structural identifiability is the more important issue; avoid overemphasis on separation of training and validation data, when validation data is not representative for target situations; avoid discussions about unreliability of unconditional predictions when these are neither possible nor necessary; avoid suggesting model simplification to increase practical identifiability when model complexity is required for structural reasons and direct generalisation is not adequate, etc.

Given the breadth of application contexts for ABM and their potential components, we strived to be general in redacting the protocol. We believe that the principles discussed here are applicable to any modelling endeavour and most disciplinary standards that have been established form special cases that are in principle covered by the protocol. In this sense, we expect that it can be useful for many different types of simulation, not only for ABM.

At this point, the KIA protocol itself is a theory-based hypothesis that requires practical testing. We propose it to the community of agent-based modellers for adoption in model construction, documentation, and review. Its use in practice will tell if it proves useful as guidance for model development and a communication device in documentation and review. Based on practical experience, it should then be reviewed and improved.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

We thank all participants of Workshop W9 “Best Practice in Agent-Based Model Parameterisation and Validation” at the 10th International Congress on Environmental Modelling & Software 2020 for their valuable and constructive input and comments. CT and TB acknowledge funding by the Federal Ministry of Education and Research of Germany (BMBF) for the project SimLearn (01IS19073C). LN acknowledges funding from the Energy Demand changes Induced by Technological and Social innovations (EDITS) project provided by Ministry of Economy, Trade, and Industry (METI), Japan. GP acknowledges funding by the Scottish Government Strategic Research Programme 2022–27, Project ID JHI-C5-1. QBL acknowledges support by the CGIAR Research Program on Grain Legumes and Dry Cereals (CRP-GLDC) and Initiative on Sustainable Intensification of Mixed Farming Systems. TF acknowledges the support of the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Program (grant agreement number 758014). Author contributions: CT conceptualized and wrote the manuscript. RH, AB, HvD, TF, QBL, ML, LN, JGP, ZS, and TB contributed ideas, comments, and corrections at all stages of manuscript writing. We thank Judith Verstegen and a further anonymous referee for highly constructive and valuable feedback during the review process.

Appendix A

A.1 Notes on differences in structural identifiability of parameters

Structural identifiability in the data can considerably differ between different groups of parameters or model components. For example, parameters that relate short-term agent behaviour to static characteristics can be estimated from sufficiently heterogeneous cross-sectional data. For contrast, parameters that affect dynamic behaviour or accumulative development over several periods require panel data (Troost and Berger, 2015a, 2020). Parameters that affect the probability of low probability events can only be identified if enough low probability events have been observed (Filatova et al., 2016). Structural non-identifiability cannot be resolved by more of the same data, but requires either widening the range of situations observed or more dimensions of the data. Under certain conditions, unidentifiable parameters may be temporarily fixed to allow identification of other components. However, fixing has to be reversed for latter predictive simulation in order not to obscure model uncertainty (noninfluence in the observed domain does not necessarily mean noninfluence in the target situation, see example in Troost and Berger 2015a).

A.2 Mapping purposes to modelling contexts

We believe that terms like prediction, forecast or projection, which are often ambiguous or defined differently between disciplines, as well as typologies of Edmonds et al. (2019) can be communicated more precisely using the suggested dimensions of the modelling context.

For example, the seven modelling purposes of Edmonds et al. (2019) could be coarsely mapped onto our characterisations of modelling context as follows: In ‘theoretical exposition’ and ‘illustration’ the system under study is the model itself, with the former being output-focused (moving from an insufficient sample situation to an in sample-situation by exhaustive simulation) and the latter putting emphasis on transparency and interpretability. ‘Analogy’ does relate to a real system and is structure-focused with a low demand on precision and comprehensiveness, but high demands on transparency and interpretability. In this three cases, conclusions about the relationship of the model to the real-world are left-aside for a moment or discussed as unmodelled uncertainty. ‘Social learning’ and education can happen in all contexts, can be about the model, opinions of participants or the real system, output or structure, but require transparency and interpretability. ‘Description’ corresponds to structure-focused, in-sample analysis. (Output-focused in-sample analysis – not mentioned by Edmonds et al. – could be termed ‘compression’: storing and reproducing observations in a more resource-efficient way than explicitly listing them.) ‘Explanation’ is structure-focused, out-of-sample generalization. ‘Prediction’ is any output-focused analysis in out-of-sample or non-representative sample settings. This wide scope of prediction still opens up a lot of room for misunderstanding and clearer definitions of modelling context using the dimensions of required precision and accuracy, transparency, etc. can help in this context to link to appropriate forms of simulation analysis (e.g. Marchau et al., 2019).

References

Alexandrov, G.A., Ames, D., Bellocchi, G., Bruen, M., Crout, N., Erechtchoukova, M., et al., 2011. Technical assessment and evaluation of environmental models and software. Environ. Model. Software 26 (3), 328–336.

An, L., Grimm, V., Turner II, B.L., 2020. Editorial: meeting grand challenges in agent-based models. JASSS 23, 13.

Andersen, T., Carstensen, J., Hernandez-Garcia, E., Duarte, C.M., 2009. Ecological thresholds and regime shifts: approaches to identification. Trends Ecol. Evol. 24 (1), 49–57.

Argent, R.M., Sojda, R.S., Guipponi, C., McIntosh, B., Voinov, A.A., Maier, H.R., 2016. Best practices for conceptual modelling in environmental planning and management. Environ. Model. Software 80, 113–121. https://doi.org/10.1016/j.envsoft.2016.02.023.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics Surveys, Statist. Surv. 4, 40–79.

Arnold, R.T., Troost, C., Berger, T., 2015. Quantifying the economic importance of irrigation water reuse in a Chilean watershed using an integrated agent-based model. Water Resour. Res. 51, 648–668. https://doi.org/10.1002/2014WR015382.

Augusiak, J., Van den Brink, P.J., Grimm, V., 2014. Merging validation and evaluation of ecological models to 'evaludation': a review of terminology and a practical approach. Ecol. Model. 280, 117–128. https://doi.org/10.1016/j.ecolmodel.2013.11.009.

Aumann, C.A., 2007. A methodology for developing simulation models of complex systems. Ecol. Model. 202, 385–396. https://doi.org/10.1016/j.ecolmodel.2006.11.005.

Banerjee, A., Chassang, S., Snowberg, E., 2016. Decision Theoretic Approaches to Experiment Design and External Validity. NBER Working Paper No, 22167.

Barlas, Y., 1996. Formal aspects of model validity and validation in system dynamics. Syst. Dynam. Rev. 12, 183–210.

Barreteau, O., Bots, P.W.G., Daniell, K.A., 2010. A framework for clarifying "Participation" in participatory research to prevent its rejection for the wrong reasons. Ecol. Soc. 15 (2), 24.

Bartlett, P.L., Mendelson, S., 2002. Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. 3, 463–482.

Bassett, R., Deride, J., 2019. Maximum a posteriori estimators as a limit of Bayes estimators. Math. Program. 174, 129–144. https://doi.org/10.1007/s10107-018-1241-0.

Beaumont, M.A., 2010. Approximate Bayesian computation in evolution and ecology. Annu. Rev. Ecol. Evol. Systemat. 41, 379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621.

Beck, M.B., Ravetz, J.R., Mulkey, L.A., Barnwell, T.O., 1997. On the problem of model validation for predictive exposure assessments. Stoch. Hydrol. Hydraul. 11, 229–254.

Bellman, R., Åström, K.J., 1970. On structural identifiability. Math. Biosci. 7, 329–339. https://doi.org/10.1016/0025-5564(70)90132-X.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Software 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011.

Berger, J., 1980. Statistical Decision Theory: Foundations, Concepts, and Methods. Springer, New York.

Berger, T., Troost, C., 2014. Agent-based modelling of climate adaptation and mitigation options in agriculture. J. Agric. Econ. 65, 323–348. https://doi.org/10.1111/1477-9552.12045.

Berger, T., Schilling, C., Troost, C., Latynskiy, E., 2010. Knowledge-brokering with agent-based models: some experiences from irrigation-related research in Chile. In: Swayne, David A., Yang, Wanhong, Voinov, A.A., Rizzoli, A., Filatova, T. (Eds.), 2010 International Congress on Environmental Modelling and Software. Ottawa, Canada.

Berger, T., Troost, C., Wossen, T., Latynskiy, E., Tesfaye, K., Gbegbelegbe, S., 2017. Can smallholder farmers adapt to climate variability, and how effective are policy interventions? Agent-based simulation results for Ethiopia. Agric. Econ. 48, 693–706.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11–49.

Beven, K.J., Smith, P.J., Freer, J.E., 2008. So just why would a modeller choose to be incoherent? J. Hydrol. 354, 15–32.

Blavatskyy, P.R., Pogrebna, G., 2010. Models of stochastic choice and decision theories: why both are important for analyzing decisions. J. Appl. Econ. 25, 963–986. https://doi.org/10.1002/jae.1116.

Borgonovo, E., Plischke, E., 2016. Sensitivity analysis: a review of recent advances. Eur. J. Oper. Res. 248 (3), 869–887.

Bray, D., von Storch, H., 2009. "Prediction" or "projection"? The nomenclature of climate science. Sci. Commun. 30 (4), 534–543.

Brenner, T., Werker, C., 2007. A taxonomy of inference in simulation models. Comput. Econ. 30, 227–244. https://doi.org/10.1007/s10614-007-9102-6.

Brewer, M.J., Butler, A., Cooksley, S.L., 2016. The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity. Methods Ecol. Evol. 7, 679–692. https://doi.org/10.1111/2041-210X.12541.

Brown, C., Alexander, P., Holzhauer, S., Rounsevell, M.D., 2017. Behavioral models of climate change adaptation and mitigation in land-based sectors. Wiley Interdisciplinary Reviews: Clim. Change 8 (2), e448.

Browne, M.W., 2000. Cross-validation methods. J. Math. Psychol. 44, 108–132. https://doi.org/10.1006/jmps.1999.1279.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in model selection. Socio. Methods Res. 33, 261–304.

Caldwell, B.J., 1991. Clarifying popper. J. Econ. Lit. 29, 1–33.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environ. Model. Software 22, 1509–1518. https://doi.org/10.1016/j.envsoft.2006.10.004.

Carauta, M., Troost, C., Guzman-Bustamante, I., Hampf, A., Libera, A., Meurer, K., Bönecke, E., Franko, U., de Aragão Ribeiro Rodrigues, R., Berger, T., 2021. Climate-related land use policies in Brazil: how much has been achieved with economic incentives in agriculture? Land Use Pol. 109, 105618 https://doi.org/10.1016/j.landusepol.2021.105618.

Carrella, E., Bailey, R., Madsen, J., 2020. Calibrating Agent-Based Models with Linear Regressions. J. Artif. Soc. Soc. Simulat. 23 (1), 7 https://doi.org/10.18564/jasss.4150.

Chen, Y., 2011. Derivation of the functional relations between fractal dimension of and shape indices of urban form. Comput. Environ. Urban Syst. 35 (6), 442–451.

Chen, S.-H., Chang, C.-L., Du, Y.-R., 2012. Agent-based economic models and econometrics. Knowl. Eng. Rev. 27, 187–219. https://doi.org/10.1017/S0269888912000136.

Chis, O.-T., Banga, J.R., Balsa-Canto, E., 2011. Structural identifiability of systems biology models: a critical comparison of methods. PLoS One 6, e27755. https://doi.org/10.1371/journal.pone.0027755.

Clarke, E.M., Henzinger, T.A., Veith, H., Bloem, R., 2018. Handbook of Model Checking, 1st. Springer, Cham.

Cobelli, C., DiStefano, J.J., 1980. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. Am. J. Physiol. Regul. Integr. Comp. Physiol. 239, R7–R24.

Constantino, S.M., Schlüter, M., Weber, E.U., Wijermans, N., 2021. Cognition and behavior in context: a framework and theories to explain natural resource use decisions in social-ecological systems. Sustain. Sci. 16, 1651–1671. https://doi.org/10.1007/s11625-021-00989-w.

de Koning, K., Filatova, T., 2020. Repetitive floods intensify outmigration and climate gentrification in coastal cities. Environ. Res. Lett. 15, 034008 https://doi.org/10.1088/1748-9326/ab6668.

Deichsel, S., Pyka, A., 2009. A pragmatic reading of Friedman's methodological essay and what it tells us for the discussion of ABMs. J. Artif. Soc. Soc. Simulat. 12, 6.

Díaz-Pacheco, J., van Delden, H., Hewitt, R., 2018. The importance of scale in land use models: experiments in data conversion, data resampling, resolution and neighborhood extent. In: Camacho Olmedo, M.T., Paegelow, M., Mas, J.F., Escobar, F. (Eds.), Geomatic Approaches for Modeling Land Change Scenarios. Springer, Cham, CH, pp. 163–186.

Drovandi, C.C., Pettitt, A.N., Lee, A., 2015. Bayesian indirect inference using a parametric auxiliary model. Stat. Sci. 30 (1), 72–95.

Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montañola-Sales, C., Ormerod, P., Root, H., Squazzoni, F., 2019. Different modelling purposes. J. Artif. Soc. Soc. Simulat. 22 (3), 6. https://doi.org/10.18564/jasss.3993.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc. 92 (438), 548–560. https://doi.org/10.2307/2965703.

Elsawah, S., Filatova, T., Jakeman, A.J., Kettner, A.J., Zellner, M.L., Athanasiadis, I.N., Hamilton, S.H., Axtell, R.L., Brown, D.G., Gilligan, J.M., Janssen, M.A., Robinson, D. T., Rozenberg, J., Ullah, I.I.T., Lade, S.J., 2020. Eight grand challenges in socio-environmental systems modelling. Socio-Environmental Systems Modelling 2, 16226. https://doi.org/10.18174/sesmo.2020a16226, 16226.

Engle, R.F., Hendry, D.F., 1993. Testing super exogeneity and invariance in regression models. J. Econom. 56, 119–139.

Epstein, J.M., 2008. Why model? J. Artif. Soc. Soc. Simulat. 11 (4), 12. http://jasss.soc.surrey.ac.uk/11/4/12.html.

Farahmand, A., Barreto, A., Nikovski, D., 2017. Value-aware loss function for model-based reinforcement learning. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics in: Proceedings of Machine Learning Research, pp. 1486–1494, 54. https://proceedings.mlr.press/v54/farahmand17a.html.

Filatova, T., 2015. Empirical agent-based land market: integrating adaptive economic behaviour in urban land-use models. Comput. Environ. Urban Syst. 54, 397–413. https://doi.org/10.1016/j.compenvurbsys.2014.06.007.

Filatova, T., Verburg, P.H., Parker, D.C., Stannard, C.A., 2013. Spatial agent-based models for socio-ecological systems: challenges and prospects. Environ. Model. Software 45, 1–7. https://doi.org/10.1016/j.envsoft.2013.03.017.

Filatova, T., Polhill, J.G., van Ewijk, S., 2016. Regime shifts in coupled socio-environmental systems: review of modelling challenges and approaches. Environ. Model. Software 75, 333–347. https://doi.org/10.1016/j.envsoft.2015.04.003.

Fischhoff, B., Davis, A.L., 2014. Communicating scientific uncertainty. Proc. Natl. Acad. Sci. USA 111 (4), 13664–13671.

Forrester, J.W., Senge, P.M., 1980. Tests for building confidence in system dynamics models. In: Legasto Jr., A.A., Forrester, J.W., Lyneis, J.M. (Eds.), System Dynamics, TIMS Studies in the Management Sciences. North-Holland, New York, Amsterdam, pp. 209–228.

Forster, M., 2000. Key concepts in model selection: performance and generalizability. J. Math. Psychol. 44, 205–231. https://doi.org/10.1006/jmps.1999.1284.

Frisch, R., 1933. Editorial. Econometrica 1, 1–5.

Gangl, M., 2010. Causal inference in sociological research. Annu. Rev. Sociol. 36 (1), 21–47.

García-Álvarez, D., Lloyd, C.D., Van Delden, H., Olmedo, M.T.C., 2019. Thematic resolution influence in spatial analysis. An application to Land Use Cover Change (LUCC) modelling calibration. Comput. Environ. Urban Syst. 78.

Gass, S.I., 1983. Decision-aiding models: validation, assessment, and related issues for policy analysis. Oper. Res. 31, 603–631.

Ghaffarian, S., Roy, D., Filatova, T., Kerle, N., 2021. Agent-based modelling of post-disaster recovery with remote sensing data. Int. J. Disaster Risk Reduc. 60, 102285 https://doi.org/10.1016/j.ijdrr.2021.102285.

Gore, R.J., Lynch, C.J., Kavak, H., 2017. Applying statistical debugging for enhanced trace validation of agent-based models. Simulation 93 (4), 273–284.

Gormley, T.A., Matsa, D.A., 2014. Common errors: how to (and not to) control for unobserved heterogeneity. Rev. Financ. Stud. 27 (2), 617–661. https://doi.org/10.1093/rfs/hht047.

Gramacy, R.B., Lee, H.K.H., 2009. Adaptive design and analysis of supercomputer experiments. Technometrics 51, 130–145.

Grazzini, Jakob, Richiardi, Matteo, 2015. Estimation of ergodic agent-based models by simulated minimum distance. J. Econ. Dynam. Control 51, 148–165. https://doi.org/10.1016/j.jedc.2014.10.006.

Grazzini, J., Richiardi, M.G., Tsionas, M., 2017. Bayesian estimation of agent-based models. J. Econ. Dynam. Control 77, 26–47.

Grimm, Volker, Railsback, S.F., 2012. Pattern-oriented modelling: a 'multi-scope' for predictive systems ecology. Phil. Trans. Biol. Sci. 367 (1586), 298–310. https://doi.org/10.1098/rstb.2011.0180.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.-H., Weiner, J., Wiegand, T., DeAngelis, D.L., 2005. Pattern-oriented modelling of agent-based complex systems: lessons from ecology. Science 310, 987–991. https://doi.org/10.1126/science.1116681.

Grimm, V., Berger, U., DeAngelis, D.L., Polhill, J.G., Giske, J., Railsback, S.F., 2010. The ODD protocol: a review and first update. Ecol. Model. 221, 2760–2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019.

Grimm, V., Augusiak, J., Focks, A., Frank, B.M., Gabsi, F., Johnston, A.S.A., Liu, C., Martin, B.T., Meli, M., Radchuk, V., Thorbek, P., Railsback, S.F., 2014. Towards better modelling and decision support: documenting model development, testing, and analysis using TRACE. Ecol. Model. 280, 129–139. https://doi.org/10.1016/j.ecolmodel.2014.01.018.

Grimm, V., Railsback, S.F., Vincenot, C.E., Berger, U., Gallagher, C., DeAngelis, D.L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J., Johnston, A.S.A., Milles, A., Nabe-Nielsen, J., Polhill, J.G., Radchuk, V., Rohwäder, M.-S., Stillman, R.A., Thiele, J.C., Ayllón, D., 2020. The ODD protocol for describing agent-based and other simulation models: a second update to improve clarity, replication, and structural realism. J. Artif. Soc. Soc. Simulat. 23, 7.

Guillaume, Joseph H.A., Jakeman, John D., Marsili-Libelli, Stefano, Asher, Michael, Brunner, Philip, Barry, Croke, Hill, Mary C., et al., 2019. Introductory overview of identifiability analysis: a guide to evaluating whether you have the right type of data for your modelling purpose. Environ. Model. Software 119, 418–432. https://doi.org/10.1016/j.envsoft.2019.07.007. September.

Hagen-Zanker, A., 2009. An improved fuzzy kappa statistic that accounts for spatial autocorrelation. Int. J. Geogr. Inf. Sci. 23 (1), 61–73.

Hands, D.W., 2001. Reflection without Rules - Economic Methodology and Contemporary Science Theory. Cambridge University Press., Cambridge.

Hansen, L.P., Heckman, J.J., 1996. The empirical foundations of calibration. J. Econ. Perspect. 10, 87–104.

Hartig, F., Calabrese, J.M., Reineking, B., Wiegand, T., Huth, A., 2011. Statistical inference for stochastic simulation models - theory and application. Ecol. Lett. 14 (8), 816–827.

Hauduc, H., Neumann, M.B., Muschalla, D., Gamerith, V., Gillot, S., Vanrolleghem, P.A., 2015. Efficiency criteria for environmental model quality assessment: a review and its application to wastewater treatment. Environ. Model. Software 68, 196–204. https://doi.org/10.1016/j.envsoft.2015.02.004.

Heckbert, S., Baynes, T., Reeson, A., 2010. Agent-based modelling in ecological economics. Ann. N. Y. Acad. Sci. 1185, 39–53. https://doi.org/10.1111/j.1749-6632.2009.05286.x.

Helton, J.C., Johnson, J.D., Sallaberry, C.J., Storlie, C.B., 2006. Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliab. Eng. Syst. Saf. 91, 1175–1209.

Hendry, D.F., 1996. On the constancy of time-series econometric equations. Econ. Soc. Rev. 27 (5), 401–422.

Heppenstall, A., Crooks, A., Malleson, N., Manley, E., Ge, J., Batty, M., 2021. Future developments in geographical agent-based models: challenges and opportunities. Geogr. Anal. 53 (1), 76–91.

Hobbs, N.T., Hilborn, R., 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecol. Appl. 16, 5–19.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688.

Jager, K.J., Tripepi, G., Chesnaye, N.C., Dekker, F.W., Zoccali, C., Stel, V.S., 2020. Where to look for the most frequent biases? Nephrology 25, 435–441. https://doi.org/10.1111/nep.13706.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Software 21, 602–614.

Jensen, T., Chappin, É.J.L., 2016. Agent-based modelling automated: data-driven generation of innovation diffusion models. In: Sauvage, S., Sánchez-Pérez, J.M., Rizzoli, A.E. (Eds.), Proceedings of the 8th International Congress on Environmental Modelling and Software, July 10-14, Toulouse. FRANCE.

Klappholz, K., Agassi, J., 1959. Methodological prescriptions in economics. Economica, New Series 26, 60–74.

Klaver, R., Haarsma, R., Vidale, P.L., Hazeleger, W., 2020. Effective resolution in high resolution global atmospheric models for climate studies. Atmos. Sci. Lett. 21 (4), e952.

Kloprogge, P., Van der Sluijs, J.P., Petersen, A.C., 2011. A method for the analysis of assumptions in model-based environmental assessments. Environ. Model. Software 26 (3), 289–301.

Kukacka, J., Barunik, J., 2017. Estimation of financial agent-based models with simulated maximum likelihood. J. Econ. Dynam. Control 85, 21–45.

Kydland, F.E., Prescott, E.C., 1996. The computational experiment: an econometric tool. J. Econ. Perspect. 10, 69–85.

Laprise, R., 1992. The resolution of global spectral models. Bull. Am. Meteorol. Soc. 73, 1453–1455. https://doi.org/10.1175/1520-0477-73.9.1453.

Le, Q.B., Seidl, R., Scholz, R.W., 2012. Feedback loops and types of adaptation in the modelling of land-use decisions in an agent-based simulation. Environ. Model. Software 27–28, 83–96.

Leamer, E., 2010. Tantalus on the way to asymtopia. J. Econ. Perspect. 24 (2), 31–46.

Lempert, R., 2019. Robust decision making (RDM). In: Marchau, V., Walker, W., Bloemen, P., Popper, S. (Eds.), Decision Making under Deep Uncertainty - from Theory to Practice. Springer, Cham, Switzerland.

Ligmann-Zielinska, A., Siebers, P.-O., Magliocca, N., Parker, D.C., Grimm, V., Du, J., et al., 2020. 'One size does not fit all': a roadmap of purpose-driven mixed-method pathways for sensi-tivity analysis of agent-based models. J. Artif. Soc. Soc. Simulat. 23 (1), 6. https://doi.org/10.18564/jasss.4201.

Lippe, M., Bithell, M., Gotts, N., Natalini, D., Barbrook-Johnson, P., Giupponi, C., Hallier, M., Hofstede, G.J., Le Page, C., Matthews, R.B., Schlüter, M., Smith, P., Teglio, A., Thellmann, K., 2019. Using agent-based modelling to simulate social-ecological systems across scales. GeoInformatica 23, 269–298. https://doi.org/10.1007/s10707-018-00337-8.

Longino, H., 1992. Essential tensions - phase two: Feminist, philosophical and social studies of science. In: McMullin, E. (Ed.), The Social Dimensions of Science. University of Notre Dame Press, Notre Dame, IN, pp. 198–216.

Lucas, R.E., 1976. Econometric policy evaluation: a critique. In: Brunner, K., Meltzer, A. (Eds.), The Phillips Curve and Labor Markets, Vol. 1 of Carnegie- Rochester Conferences on Public Policy. North-Holland Publishing Company, Amsterdam, pp. 19–46.

Lux, T., Zwinkels, R.C., 2018. Empirical validation of agent-based models. In: Handbook of Computational Economics, vol. 4. Elsevier, pp. 437–488.

Magliocca, N.R., McConnell, V., Walls, M., 2016. The role of subjective risk perceptions in shaping coastal development dynamics. In: Sauvage, S., Sánchez-Pérez, J.M., Rizzoli, A.E. (Eds.), Proceedings of the 8th International Congress on Environmental Modelling and Software, July 10-14, Toulouse. FRANCE.

Manderscheid, L.V., 1965. Significance levels. 0.05, 0.01, or? J. Farm Econ. 47 (5), 1381–1385. https://doi.org/10.2307/1236396.

Manski, C.F., 2019. Treatment choice with trial data: statistical decision theory should supplant hypothesis testing. Am. Statistician 73 (Suppl. 1), 296–304. https://doi.org/10.1080/00031305.2018.1513377.

Marchau, V., Walker, W., Bloemen, P., Popper, S., 2019. Introduction. In: Marchau, V., Walker, W., Bloemen, P., Popper, S. (Eds.), Decision Making under Deep Uncertainty - from Theory to Practice. Springer, Cham, Switzerland.

Marshall, B.D.L., Galea, S., 2015. Formalizing the role of agent-based modelling in causal inference and epidemiology. Am. J. Epidemiol. 181, 92–99. https://doi.org/10.1093/aje/kwu274.

McCarl, B., Apland, J., 1986. Validation of linear programming models. South. J. Agric. Econ. 18, 155–164.

McCloskey, D.N., 1983. The rhetoric of economics. J. Econ. Lit. 21, 481–517.

McCloskey, D.N., 1985. The loss function has been mislaid: the rhetoric of significance tests. Am. Econ. Rev. 75 (2), 201–205.

McGarigal, K., 2014. Landscape pattern metrics. In: Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., Teugels, J.L. (Eds.), Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat07723.

Midgley, D., Marks, R., Kunchamwar, D., 2007. Building and assurance of agent-based models: an example and challenge to the field. J. Bus. Res. 60, 884–893. https://doi.org/10.1016/j.jbusres.2007.02.004.

Moss, S., Edmonds, B., 2005. Towards good social science. J. Artif. Soc. Soc. Simulat. 8, 13.

Mössinger, J., Troost, C., Berger, T., 2022. Bridging the gap between models and users: a lightweight mobile interface for optimized farming decisions in interactive modeling sessions. Agric. Syst. 195, 103315 https://doi.org/10.1016/j.agsy.2021.103315.

Niamir, L., Kiesewetter, G., Wagner, F., et al., 2020a. Assessing the macroeconomic impacts of individual behavioral changes on carbon emissions. Climatic Change 158, 141–160. https://doi.org/10.1007/s10584-019-02566-8.

Niamir, L., Ivanova, O., Filatova, T., 2020b. Economy-wide impacts of behavioral climate change mitigation: linking agent-based and computable general equilibrium models. Environ. Model. Software 134, 104839.

Nolan, J., Parker, D., van Kooten, G.C., Berger, T., 2009. An overview of computational modelling in agricultural and ressource economics. Can. J. Agric. Econ. 57, 417–429.

Onggo, B.S., Karatas, M., 2016. Test-driven simulation modelling: a case study using agent-based maritime search-operation simulation. Eur. J. Oper. Res. 254 (2), 517–531. https://doi.org/10.1016/j.ejor.2016.03.050.

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

Parker, D.C., Entwisle, B., Rindfuss, R.R., Vanwey, L.K., Manson, S.M., Moran, E., et al., 2008. Case studies, cross-site comparisons, and the challenge of generalization: comparing agent-based models of land-use change in frontier regions. J. Land Use Sci. 3 (1), 41–72.

Perron, P., 2006. Dealing with structural breaks. In: Patterson, K., Mills, T.C. (Eds.), Palgrave Handbook of Econometrics. Palgrave-Macmillan, pp. 278–352.

Pielke, R.A., 1991. A recommended specific definition of "resolution". Bull. Am. Meteorol. Soc. 72, 1914. https://doi.org/10.1175/1520-0477-72.12.1914, 1914.

Polhill, G., Salt, D., 2017. The importance of ontological structure: why validation by 'Fit-to-data' is insufficient. In: Edmonds, B., Meyer, R. (Eds.), Simulating Social Complexity: A Handbook, Understanding Complex Systems. Springer International Publishing, Cham, pp. 141–172. https://doi.org/10.1007/978-3-319-66948-9_8.

Pontius Jr., R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Rem. Sens. 32, 4407–4429. https://doi.org/10.1080/01431161.2011.552923.

Puy, A., Piano, S.L., Saltelli, A., 2021. Is VARS more intuitive and efficient than Sobol'indices? Environ. Model. Software 137, 104960.

Quine, W.V.O., 1951. Two dogmas of empiricism. Philos. Rev. 60, 20–43.

Rand, W., Rust, R.T., 2011. Agent-based modelling in marketing: guidelines for rigor. Int. J. Res. Market. 28, 181–193. https://doi.org/10.1016/j.ijresmar.2011.04.002.

Rosenzweig, M., Udry, C., 2016. External validity in a stochastic world. NBER Working Paper, 22449. https://doi.org/10.3386/w22449.

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

Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. Environ. Model. Software 25, 1508–1517.

Saltelli, A., Guimaraes Pereira, A., van der Sluijs, J., Funtowicz, S., 2013. What do I make of your latinorum? Sensitivity auditing of mathematical modelling. Int. J. Foresight Innovation Policy 9 (2–4), 213–234.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis: the Primer. John Wiley & Sons, Hoboken, NJ.

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., et al., 2019. Why so many published sensitivity analyses are false: a systematic review of sensitivity analysis practices. Environ. Model. Software 114, 29–39.

Schaeffli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21, 2075–2080.

Schlüter, M., Baeza, A., Dressler, G., Frank, K., Groeneveld, J., Jager, W., Janssen, M.A., McAllister, R.R.J., Müller, B., Orach, K., Schwarz, N., Wijermans, N., 2017. A framework for mapping and comparing behavioural theories in models of social-ecological systems. Ecol. Econ. 131, 21–35. https://doi.org/10.1016/j.ecolecon.2016.08.008.

Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V., 2010. Ecological models supporting environmental decision making: a strategy for the future. Trends Ecol. Evol. 25, 479–486. https://doi.org/10.1016/j.tree.2010.05.001.

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res. 46, 1–17.

Schreinemachers, P., Berger, T., 2011. An agent-based simulation model of human environment interactions in agricultural systems. Environ. Model. Software 26, 845–859 j.envsoft.2011.02.004.

Schulze, J., Müller, B., Groeneveld, J., Grimm, V., 2017. Agent-based modelling of social-ecological systems: achievements, challenges, and a way Forward. J. Artif. Soc. Soc. Simulat. 20 (2), 8. https://doi.org/10.18564/jasss.3423, 2017.

Siebers, O.P., Macal, M.C., Garnett, J., Buxton, D., Pidd, M., 2010. Discrete-event simulation is dead, long live agent-based simulation. J. Simulat. 4, 204–210. https://doi.org/10.1057/jos.2010.14.

Smith, L.H., 2020. Selection mechanisms and their consequences: understanding and addressing selection bias. Curr Epidemiol Rep 7, 179–189. https://doi.org/10.1007/s40471-020-00241-6.

Spear, R.C., Hornberger, G.M., 1980. Eutrophication in Peel inlet—II. Identification of critical uncertainties via generalised sensitivity analysis. Water Res. 14, 43–49. https://doi.org/10.1016/0043-1354(80)90040-8.

Stedinger, J. R, Vogel, R.M., Lee, S.U., Batchelder, R., 2008. Appraisal of the generalized likeli- hood uncertainty estimation (GLUE) method. Water Resour. Res. 44, 1–18. W00B06.

Stephens, P.A., Buskirk, S.W., Hayward, G.D., Martinez Del Rio, C., 2005. Information theory and hypothesis testing: a call for pluralism. J. Appl. Ecol. 42, 4–12. https://doi.org/10.1111/j.1365-2664.2005.01002.x.

Stigler, S.M., 2007. The epic story of maximum likelihood. Stat. Sci. 598–620.

Stigter, J.D., Beck, M.B., Molenaar, J., 2017. Assessing local structural identifiability for environmental models. Environ. Model. Software 93, 398–408. https://doi.org/10.1016/j.envsoft.2017.03.006. July.

Stout, N.K., Goldie, S.J., 2008. Keeping the noise down: common random numbers for disease simulation modelling. Health Care Manag. Sci. 11, 399–406. https://doi.org/10.1007/s10729-008-9067-6.

Tarantola, S., Becker, W., Zeitz, D., 2012. A comparison of two sampling methods for global sensitivity analysis. Comput. Phys. Commun. 183, 1061–1072. https://doi.org/10.1016/j.cpc.2011.12.015.

Thiele, J.C., Kurth, W., Grimm, V., 2014. Facilitating parameter estimation and sensitivity analysis of agent-based models: a cookbook using NetLogo and R. J. Artif. Soc. Soc. Simulat. 17, 11. https://doi.org/10.18564/jasss.2503.

Troost, C., Berger, T., 2015a. Dealing with uncertainty in agent-based simulation: farm-level modelling of adaptation to climate change in southwest Germany. Am. J. Agric. Econ. 97, 833–854. https://doi.org/10.1093/ajae/aau076.

Troost, C., Berger, T., 2015b. Process-based simulation of regional agricultural supply functions in Southwestern Germany using farm-level and agent-based models. In: International Association of Agricultural Economists, 2015 Conference. https://doi.org/10.22004/ag.econ.211929. August 9-14, 2015, Milan, Italy.

Troost, C., Berger, T., 2016. Advances in probabilistic and parallel agent-based simulation: modelling climate change adaptation in agriculture. In: Sauvage, S., Sánchez Pérez, J.-M., Rizzoli, A.E. (Eds.), Proceedings of the 8th International Congress on Environmental Modelling and Software, July 10-14. Toulouse, France.

Troost, C., Berger, T., 2020. Formalising validation? Towards criteria for valid conclusions from agent-based simulation. In: van Griensven, A., Nossent, J., Ames, D.P. (Eds.), 10th International Congress on Environmental Modelling and Software. Brussels, Belgium.

Troost, C., Duan, X., Gayler, S., Heinlein, F., Klein, C., Aurbacher, J., Demyan, M.S., Högy, P., Laub, M., Ingwersen, J., Kremer, P., Mendoza Tijerino, F., Otto, L.H., Poyda, A., Warrach-Sagi, K., Weber, T.K.D., Priesack, E., Streck, T., Berger, T., 2020. The bioeconomic modelling system MPMAS-XN: simulating short and long-term feedback between climate, crop growth, crop management and farm management. In: van Griensven, A., Nossent, J., Ames, D.P. (Eds.), 10th International Congress on Environmental Modelling and Software. Brussels, Belgium.

Troost, C., Parussis-Krech, J., Mejail, M., Berger, T., 2022. Boosting the scalability of farm-level models: efficient surrogate modeling of compositional simulation output. Comput. Econ. https://doi.org/10.1007/s10614-022-10276-0.

van Asselt, M.B.A., 2000. Perspectives on Uncertainty and Risk - the PRIMA Approach to Decision Support. Kluwer Academic Publishers, Boston, Dordrecht, London.

van Delden, H., van Vliet, J., Rutledge, D.T., Kirkby, M.J., 2011. Comparison of scale and scaling issues in integrated land-use models for policy support. Agric. Ecosyst. Environ. 142 (1–2), 18–28.

van der Sluijs, J., 2017. The NUSAP Approach to Uncertainty Appraisal and Communication. In: Spash, C.L. (Ed.), Routledge Handbook of Ecological Economics: Nature and Society. Routledge, London.

van der Vaart, E., Beaumont, M.A., Johnston, A.S.A., Sibly, R.M., 2015. Calibration and evaluation of individual-based models using Approximate Bayesian Computation. Ecol. Model. 312, 182–190. https://doi.org/10.1016/j.ecolmodel.2015.05.020.

van Vliet, J., Hagen-Zanker, A., Hurkens, J., Van Delden, H., 2013. A fuzzy set approach to assess the predictive accuracy of land use simulations. Ecol. Model. 261–262, 32–42.

Vandecasteele, L., Debels, A., 2007. Attrition in panel data: the effectiveness of weighting. Eur. Socio Rev. 23 (1), 81–97, 0.1093/esr/jcl021.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27, 1433. https://doi.org/10.1007/s11222-016-9696-4.

Verhoog, R., Ghorbani, A., Dijkema, G.P.J., 2016. Modelling socio-ecological systems with MAIA: a biogas infrastructure simulation. Environ. Model. Software 81, 72–85. https://doi.org/10.1016/j.envsoft.2016.03.011.

Verstegen, J.A., Karssenberg, D., van der Hilst, F., Faaij, A.P., 2016. Detecting systemic change in a land use system by Bayesian data assimilation. Environ. Model. Software 75, 424–438.

Verstegen, J.A., van der Laan, C., Dekker, S.C., Faaij, A.P., Santos, M.J., 2019. Recent and projected impacts of land use and land cover changes on carbon stocks and biodiversity in East Kalimantan, Indonesia. Ecol. Indicat. 103, 563–575.

Vester, F., 2002. Die Kunst vernetzt zu denken: Ideen und Werkzeuge für einen neuen Umgang mit Komplexität; ein Bericht an den Club of Rome. Dt. Taschenbuch-Verlag.

Voinov, A., Bousquet, F., 2010. Modelling with stakeholders. Environ. Model. Software 25, 1268–1281. https://doi.org/10.1016/j.envsoft.2010.03.007.

Voinov, A., Kolagani, N., McCall, M.K., Glynn, P.D., Kragt, M., Ostermann, F., Pierce, S., Ramu, P., 2016. Modelling with stakeholders – next generation. Environ. Model. Software 77, 196–220. https://doi.org/10.1016/j.envsoft.2015.11.016.

Voinov, A., Shugart, H.H., 2013. 'Integronsters', integral and integrated modeling. Environ. Model. Software 39, 149–158. https://doi.org/10.1016/j.envsoft.2012.05.014.

Walker, W.E., Harremoës, P., Rotmans, J., van der Sluijs, J.P., van Asselt, M.B.A., Janssen, P., et al., 2003. Defining uncertainty: a conceptual basis for uncertainty management in model- based decision support. Integrated Assess. 4 (1), 5–17.

Ward, E.J., 2008. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. Ecol. Model. 211 (1–2), 1–10.

Williams, T.A., Sweeney, D.J., Anderson, D.R., 2022. Sample survey methods. In: Encyclopedia Britannica. https://www.britannica.com/science/statistics/Sample-survey-methods.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82.

Windrum, P., Fagiolo, G., Moneta, A., 2007. Empirical validation of agent-based models: alternatives and prospect. J. Artif. Soc. Soc. Simulat. 10, 8.

Yates, L.A., Richards, S.A., Brook, B.W., 2021. Parsimonious model selection using information theory: a modified selection rule. Ecology 102 (10), e03475. https://doi.org/10.1002/ecy.3475.

Gallagher, C.A., Chudzinska, M., Larsen-Gray, A., Pollock, C.J., Sells, S.N., White, P.J.C, Berger, U., 2021. From Theory to Practice in Pattern-Oriented Modelling: Identifying and Using Empirical Patterns in Predictive Models. Biological Reviews 56 (5), 1868–1888. https://doi.org/10.1111/brv.12729.