Suggested citation: Troost, C., Bell, A. R., van Delden, H., Huber, R., Filatova, T., Le, Q.B., Lippe, M., Niamir, L., Polhill, J.G., Sun, Z. and Berger, T. (2022). How to Keep it Adequate: A Validation Protocol for Agent-Based Simulation. Preprint article available at SSRN: https://ssrn.com/abstract=4161475 or http://dx.doi.org/10.2139/ssrn.4161475

How to keep it adequate: A validation protocol for agent-based simulation

Authors:

Christian Troost, Hans-Ruthenberg-Institute, Universität Hohenheim, Stuttgart, Germany, christian.troost@uni-hohenheim.de, ORCID: 0000-0003-4626-7117 (Corresponding author)

Andrew R. Bell, Earth & Environment Department, Boston University, Boston, MA, USA, ORCID: 0000-0002-1164-312X

Hedwig van Delden, Research Institute for Knowledge Systems (RIKS), Maastricht, the Netherlands, ORCID: 0000-0001-6976-4832

Robert Huber, Agricultural Economics and Policy Group ETH Zürich, ORCID: 0000-0003-4545-456X

Tatiana Filatova, Multi Actor Systems Department, Faculty of Technology Policy and Management, TU Delft, The Netherlands, ORCID 0000-0002-3546-6930

Quang Bao Le, International Center for Agricultural Research in the Dry Areas (ICARDA), Cairo, Egypt, ORCID: 0000-0001-8514-1088

Melvin Lippe, Thünen Institute of Forestry, Hamburg, ORCID: 0000-0003-4323-8767

Leila Niamir, Energy, Climate, and Environment Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, ORCID: 0000-0002-0285-5542

J. Gareth Polhill, The James Hutton Institute, Craigiebuckler, Aberdeen AB15 8QH, United Kingdom. ORCID: 0000-0002-8596-0590

Zhanli Sun, Leibniz Institute of Agricultural Development in Transition Economies (IAMO), Germany, ORCID: 0000-0001-6204-4533

Thomas Berger, Hans-Ruthenberg-Institute, Universität Hohenheim, Stuttgart, Germany, ORCID: 0000-0003-3316-9614

Highlights

- A comprehensive understanding of validation for agent-based models and beyond
- 11 dimensions to characterise the modelling context and purpose
- A characterisation of the premises of common validation approaches
- A detailed protocol to guide context-adequate model construction and review
- A consistent tracking of uncertainty propagation through the modelling process

Abstract:

Agent-based models are used in a huge diversity of contexts, which complicates the establishment of a shared understanding of model validity and adequate methods for model construction, inference and validation. Starting from the tenet that model validity can only be judged with respect to a well-defined purpose and context, we conceptualise validation as systematically substantiating the premises on which conclusions from simulation analysis for a specific context are built.

We revisit the premises of empirical and structural validation and argue that validation should not be understood as an isolated step in the modelling process. Rather, sound conclusions from simulation analysis require context-adequate choices at all steps of simulation analysis.

To facilitate communication, we develop a protocol of guiding questions to analyse the modelling context, choose appropriate methods at each step, document the premises involved in a specific simulation analysis, and demonstrate the adequacy of the model for its context.

Key words:

Model validity – model inference – calibration – generalisation – regime shift - identifiability

1 Introduction

The increasing application of agent-based simulation models (ABM) for policy analysis in environmental and land system sciences, among other fields, demands improving and formalising methods of their validation (Heppenstall et al. 2021; Elsawah et al. 2020; An et al. 2020; Niamir et al. 2020b; Brown et al. 2016; Filatova, 2015; Filatova et al., 2013; Heckbert et al., 2010; Marshall & Galea, 2015; Rand & Rust, 2011; Siebers et al., 2010; Midgley et al., 2007).

A variety of approaches for constructing valid ABM have been suggested (e.g. Augusiak et al. 2014; Brenner & Werker, 2007; Deichsel & Pyka, 2009; Moss & Edmonds, 2005; Grimm et al., 2005) and many examples for formal empirical validation and calibration of ABM exist: Indirect inference methods for ABM calibration 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 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 & Chappin, 2016) and real estate market interactions (Filatova 2015; Magliocca et al., 2016; de Koning and Filatova, 2020); robust parameter uncertainty reduction in agricultural economics (Arnold et al. 2015; Troost & Berger 2015a, Berger et al. 2017).

A consensus or a formal guideline which method to choose for a specific ABM application context that transcends disciplines, has, however, not yet been established, even within the more confined field of ABM in environmental and land system sciences (An et al. 2020; Polhill & Salt 2017; Filatova 2015).

Empirical output validation, i.e. comparing model predictions to observations of a real-world system, is widely regarded as the ideal of validation because it entails reproducible protocols and quantitative, replicable and transparently communicable results. However, it has also been clearly demonstrated that overreliance on goodness-of-fit to observations is misleading and inadequate if the underlying (statistical) assumptions for empirical validation are not fulfilled in a specific research context (e.g. Oreskes et al. 1994; Polhill & Salt 2017).

As inherently structure-rich models, ABM are often used in contexts where simpler, statistical approaches are not applicable and as a consequence also the prerequisites for (system-level) empirical validation are typically not fulfilled (Berger & Troost 2014). The importance of structural validation and sensitivity analysis for such contexts has been widely recognised (Moss & Edmonds 2005; Troost & Berger 2015a; Marshall & Galea, 2015; Polhill & Salt 2017). Structural validation, i.e. ensuring adequate correspondence of model structure and processes with their real-world counterparts, is often less formalised. When using empirical validation for model components at the micro level, similar statistical prerequisites have to be considered as in empirical macrovalidation. While formal approaches for conducting sensitivity analysis have been clearly formulated (e.g. Saltelli et al. 2008), it is not necessarily obvious which uncertainties and criteria for robustness should be considered and how they relate to the encompassing modelling process (Ligmann-Zielinska et al. 2020).

The recognition that models are by definition abstractions from reality and ultimately their absolute truth cannot be proven empirically (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 & Senge, 1980; Gass, 1983; McCarl & Apland, 1986; Oreskes et al., 1994; Barlas, 1996; Kydland & Prescott, 1996; Rykiel, 1996; Beck et al., 1997; Jakeman et al., 2006; 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.

ABM are used for a large variety of purposes and contexts (Edmonds et al. 2019; Lippe et al. 2019; Schulze et al. 2017). Hence, on the one hand, formalising ABM validation cannot mean prescribing one statistical validation procedure to all ABM. On the other hand, context-dependency of validity does not mean 'anything goes'. There are fundamental principles that are essential for a valid analysis in certain contexts. There is a vast body of literature that suggests, justifies, discusses or criticises specific approaches for model selection, calibration, testing and analysis. Often, however, the modelling contexts for which these methods are applicable are not explicitly delineated, because they are implicit in the disciplinary context or even ignored.

In this article, we argue that, under a paradigm of adequacy, validity cannot be assured by the one confined, isolated step of the modelling process – typically located after calibration and before predictive simulations – which is commonly called validation. Instead, it requires context-adequate and mutually consistent choices at all stages of the simulation analysis including the choice of model components, choice of methods for parameterisation, model inference (inverse modelling, calibration, estimation), testing 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.

The ABM community has successfully adopted the ODD protocol (Grimm et al., 2020, 2010) for formal model documentation. Schmolcke et al. (2010) and Grimm et al. (2014) have suggested the TRACE format for formally documenting the modelling process. Though TRACE highlights that all the elements of a modelling process are relevant for assessing the validity of simulation analysis, it does not provide formal guidelines, which methods to use in which contexts. Our article, which has resulted from community discussions initiated in workshop W9 of the 2020 IEMSS conference, aims to fill this gap.

In the first part of this article, we conceptualise validation as “challenging and substantiating the premises on which the conclusions from simulation analysis are built”. We revisit premises typically used in simulation analysis and discuss in how far they are tested, respectively in how far they are actually presupposed by empirical and structural validation, uncertainty analysis, model selection, empirical parameter estimation and result interpretation.

On this basis, in the second part, we develop a protocol to help modellers keep it adequate (KIA): a protocol of guiding questions to characterise the modelling context for choosing adequate model components and methods of parameterisation, testing and uncertainty analysis step by step. The KIA protocol 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 validation and calibration methods.

2 Validation: 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 & Senge, 1980; Gass, 1983; McCarl & Apland, 1986; Oreskes et al., 1994; Barlas, 1996; Kydland & Prescott, 1996; Rykiel, 1996; Beck et al., 1997; Jakeman et al., 2006; 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 & Agassi, 1959; Caldwell 1991).

In its most generic form, scientific arguments that employ simulation modelling conform to the following logical proposition (Troost & 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 conditions U.” $R(t, x, y) \wedge U(t)$

Conclusion: “We conclude Z”. $\therefore Z$ *by* $A \wedge B$ *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 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. Recognising the uncertainty in the simulation process, the third subsection discusses 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 empirical validation and inverse modelling

The key underlying premise of empirical output validation 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*). For contrast, whenever the simulation purpose is prediction or counterfactual simulation, the target situations (life after climate changed, in our example) have not been (fully) observed. The same holds implicitly for 'explanation' where the objective typically is to find a generalisable explanatory model (Edmonds et al. 2019).

Generalisation of behaviour from observations to unobserved target situations needs to involve statistical considerations in order to avoid propagating spurious, unsystematic relationships (Hansen & Heckman 1996): *Direct generalisation* of statistical relationships, including X-Y relationships and predictive performance, is only possible if the 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*).

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 & Debels 2007; Gangl 2010; Gormley & 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).

It is very important to realise that these preconditions apply to any form of model inference by *inverse modelling* (i.e. calibration, empirical model selection or parameter estimation) using observed behaviour. In all cases, ignoring sampling error and bias leads to the generalisation of unsystematic, confounded or unstable relationships (overfitting) that cause inaccurate and misleading out-of-sample predictions (Browne 2000; Forster 2000; Hansen & Heckman 1996).

In non-representative sample settings, simulation 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 here: Structural knowledge often admits alternative model formulations or parameterisations (candidates). Even if a sample is not representative of the target situations, it can be used to discriminate between the candidates if it is representative (and sufficiently redundant) in

a domain 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 (*indirect generalisation*)[1]. Importantly, the predictive accuracy measured in the sample cannot be straightforwardly generalised to the target situation in these cases.

Using a sample to reliably discriminate between candidates or detect statistical relationships presupposes *structural* and *practical identifiability* (Bellman & Åström, 1970; Cobelli & 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 domain. Even a fully representative and redundant sample is not able to distinguish between models that predict the same output for the same input.[2]

Practical identifiability means that the variation in the observational data in connection with auxiliary 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 (Brown 2000; Burnham & Anderson 2004; Polhill & Salt 2017). Two model structures or parameter sets that cannot be discriminated by given data are termed ‘equifinal’ (Beven & Freer 2001).

2.2 Premises of structural validation and structure-based model choice

As argued above, 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 (1931) 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.”

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 & Senge, 1980).

---

[1] 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.

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

Sufficient accuracy in the representation of individual processes is the subject of micro-validation (Moss & 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 inverse modelling and estimation. The premises for empirical estimation and validation of process models at the micro level are the same as at the macro (full system) level: sample representativity, identifiability and control of sampling error. 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

In practice, all system knowledge and data used in simulation analysis are 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 a plausible alternative model may show different results. Rather, it is an essential component of *U(s)* to show 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 & Davis 2014; Berger & Troost 2014; Troost & Berger 2015a; Marchau et al. 2019).

This implies, firstly, that implications of uncertainty in structural knowledge and uncertainty in model inference from data must be carefully assessed. In predictive analysis, the uncertainty in the anticipated input for a target situation needs to be considered, additionally. Secondly, the type and degree of uncertainty and bias that are compatible with conclusion Z must be carefully specified in the major premise.

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 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 & 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 & 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 & 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 context-adequate agent-based simulation

Summarising the previous section, sound conclusions from simulation analysis require (i) a logically valid structure for a scientific argument targeted at a carefully defined research question; (ii) a convincing use of models and methods of analysis to support the premises of the argument; (iii) a transparent evaluation whether preconditions for the use of chosen models and methods hold in the specific modelling context.

This modelling context consists of the purpose (research question) and the available knowledge and data about the modelled system. We identified eleven dimensions to be derived from the modelling context which influence an adequate choice of models and methods. In order to improve clarity about distinct possible reasons for similar method choices, it is useful to make a distinction between dimensions that can be derived directly from the research question alone (Fig. 1 a-f), and those that require a more in-depth analysis of the relationship between research question and system knowledge and data during the modelling process (Fig. 1 g-k)

In the following sections, we sketch a protocol (Fig.1), a set of questions for each stage of simulation analysis, that helps characterise the modelling context (3.1) and guide the choice of context-adequate methods (3.2) based on these dimensions. Where available, we list formal methods of analysis with useful references and highlight the premises for their applicability. The protocol is organized in 12 steps and emphasises the documentation and consistent propagation of uncertainty through the modelling process, to ensure that the robustness of final conclusions can be comprehensively assessed (3.3).

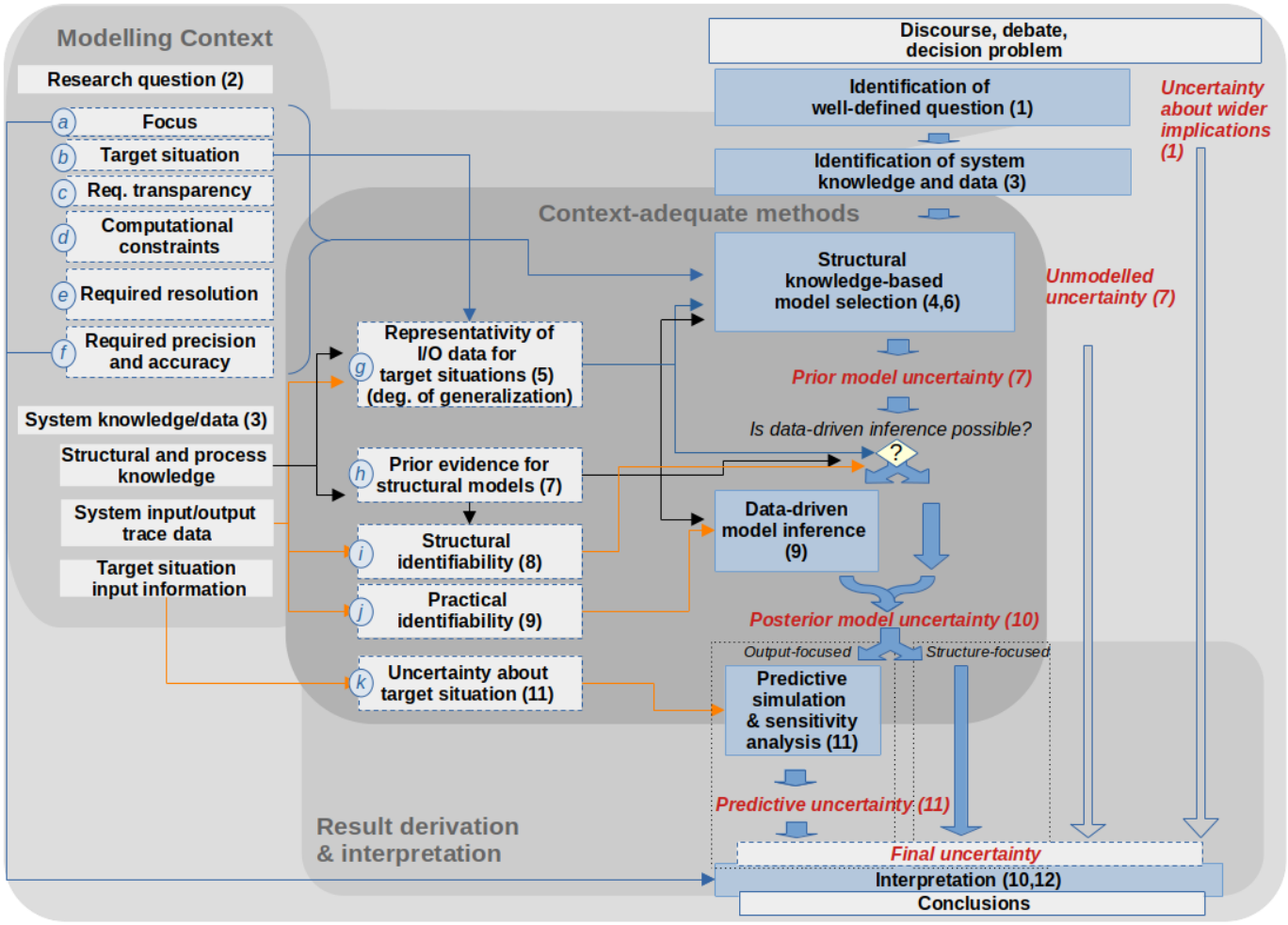

Fig. 1: Tracing the influence of the modelling context on adequate decisions in and conclusions from simulation analysis. Conceptual basis and structural overview of the protocol. (Note: Numbers refer to steps in the protocol. Colours of arrows help to visually trace crossing connections, but have no deeper significance.)

3.1 Context: Defining the modelling context

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

3.1.1 Precisely define the research question (Step 1)

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 the 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 & 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, respectively what precise argument it could contribute to the debate.

3.1.2 Characterise requirements implied by research question (Step 2)

Table 1 (a) provides guiding questions for identifying six dimensions of the modelling context from the research question itself without yet considering data or system knowledge: The most basic consideration is the *focus of interest:* Does it lie in anticipating system output for specific situations or in describing or understanding system structure? 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).

3.1.3 Identify knowledge and data about structure and behaviour of the modelled system (Step 3)

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) (Tab. 1b).

Table 1: 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			
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 *a specific sample of observed data* (description, compression) (specify which) (iii) 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 5)* *- To select an adequately comprehensive model in structural-knowledge-based model selection (steps 4,6)*

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.* *For 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.*
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.*
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 questions 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*	- *To be compared with the resolution of observational data to determine the degree of generalisation (step 5),* - *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 4,6)*

		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.)	
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 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?	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 ±2K 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)*

Step 3: Analyse system information			
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 5) 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?		*see steps 6, 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?		*see steps 5, 8, 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?		*see step 11*

3.2 Context-adequate model and parameter selection and uncertainty documentation

Appropriate simulation models can be selected in two steps: In a first 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. A set of multiple candidates fulfilling the requirements represents the *prior model uncertainty*[3]. In a potential second step, inference by inverse modelling on data of observed system behaviour 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).

Under suitable conditions, the two steps complement each other: The first step is key to ensure that only adequate candidates are considered in inverse modelling. 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). 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 & 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.

For the second step, it is key to ensure the adequacy of the inverse modelling process itself. 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, model inference by comparison to observed system behaviour is clearly not adequate.

3.2.1 ABM as composite models: Structuring component context (Step 4)

Regarding the application of agent-based simulation, the first thing to ask in structural model choice is certainly whether an ABM suits the given modelling context. ABM are typically composite models (model systems), which are composed of lower-hierarchy models 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 decisions (Schlüter et al. 2017). ABM also typically contain models of agent interactions, e.g. communication, markets, auction, collective action or network models (Schreinemachers & 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).[4]

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-

[3] 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).

[4] 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.

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 & Shugart 2013). In direct generalisation contexts, prediction is often achieved more efficiently with statistical or machine learning models (Polhill & Salt 2017).[5]

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 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. For example, even if the overall modelling context 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 even allow for the use of machine learning 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).

The next 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.***[6] The whole process may require iteratively moving back and forth between composite and components through steps 4-10 until an adequate composite structure for the overall modelling context has been established (Tab. 2, step 4).

3.2.2 Representativity of data and degree of generalisation (Step 5)

The next step (Tab.2, step 5) in choosing an adequate model or model component is to contrast the observed or observable data with the target situation to *determine the degree of generalisation and extrapolation implied*: Can the observed sample of system behaviour be considered *representative* for the target situations? Are there regime shifts, non-stationarities or structural breaks or can statistical relationships be considered stable between observed and target situations? Are all relevant system states represented in the data with sufficient probability? 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.

3.2.3 Choosing structurally adequate candidate models and prior parameter ranges for each component (Step 6)

Table 2 contains guiding questions for assessing the adequacy of model candidates and parameter ranges for a (component) modelling context from a structural point of view. The third column

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

6 Especially for inverse modelling it may be useful to subdivide the composite into observational units that do not necessarily have to correspond to lower hierarchy models, but may also use different boundaries if that, for example, allows exploiting better identifiability by subsystem input-output datasets.

indicates selected literature sources that expand on the relevant theory or suggest formal tests for the assessment of the questions.

Logical consistency, correct technical implementation, and fit to the required resolution and resource constraints are obvious preconditions for candidate models that have to be carefully assessed even if the component context allows for direct generalisation.

For adequate structure-based model selection, it is useful to first sketch a comprehensive conceptual system model, even if not all system process 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 applicabilit*y and *sufficient completeness* of processes for the target situations (Parker et al. 2008). It has to 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) (Lucas 1976; Engle & 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 & Berger 2015b; Troost et al. 2022).

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. 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) 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).[7] 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.[8]

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

[8] This principle mirrors the conservative rationale in statistical hypothesis testing: Type II errors, false-negatives, are preferred over type I errors, false-positives.

Table 2 Guiding question and formal methods for adequate structural knowledge-based model choice

Dimension	Guiding questions	Outcome
Step 4. Structure the tasks into components and observational units		
Component definition	Structure the modelling task into components (and functional links between them) Identify and characterise the specific modelling context of components Identify potential observational units for inverse modelling	Outcome: A structuration of the simulation task into components and the characterisation of the modelling context for each component
Step 5: 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: 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 & Debels 2007; Gangl 2010; Gormley & Matsa 2014; Jager et al. 2020; Smith 2020) Considering the above, is direct generalisation of statistical relationships from observations to target situations potentially possible?	Outcome: A well-argued decision for either of - *No generalisation implied* (target situations fully contained in observed data) - *Direct generalisation potentially possible* (data sufficiently representative for target situations, possible biases correctable) - *Direct generalisation not possible* (data not representative, structural breaks, non-stationarity)

Dimension	Guiding questions	Formal methods
Step 6. Identify structurally adequate candidate models and (prior) parameter ranges		
Domain of applicability / Structure and parameter constancy	Do target situations correspond to situations for which the model was designed or estimated? If not, do parameters and model structure represent relationships considered constant and stable across all relevant target situations? Can all relevant differences between 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 & Davis 2014, Rosenzweig & Udry 2016) Extreme condition tests (Forrester & Senge 1980) Behaviour-sensitivity tests (Barlas 1996)
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 & Senge 1980) Pattern-oriented modelling (Grimm & Railsback 2012) Face validation, Stakeholder participation (Voinov & 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)
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 & 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 & Railsback 2012)

Expected deviations	Are the expected deviations (residuals) of the model a priori consistent with the precision and accuracy (certainty, relativity, symmetry) required by the modelling context?	
Match of effective resolution	What is the *effective* (temporal, spatial, thematic) resolution of the model? Does the *effective resolution* of the model match the required resolution of the modelling context? 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), then the effective resolution would be the size of this neighbourhood. As an extreme case, if 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 a class-specific probability distribution) 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).	Aumann 2007; van Delden et al. 2011; Díaz-Pacheco et al. 2018; García-Álvarez et al. 2019.
Transparency and resource constraints:	Does the model match transparency, interpretability and resource use restrictions implied by the research question?	
Logical consistency	Is the model formulation in itself logically consistent?	Face validation for logical errors; Formal ontologies and ontology assessment tools (Polhill & Salt 2017)
Technical verification	Has the conceptual model been correctly implemented in computer code?	Formal testing (see overview in Midgeley et al. 2007); Unit testing (Onggo & Karatas 2016); Statistical debugging & trace validation (Gore et al. 2017); Model checking (Clarke et al. 2018)

3.2.4 Documenting prior and input data uncertainty and assessing structural identifiability (Steps 7, 8)

Structure-based model selection typically results in a number of plausible model structures and parameter values and this prior uncertainty should be documented (even if not all plausible alternatives can be implemented and tested). The first step to determine whether data-driven model inference (calibration, model selection) can help reduce this prior uncertainty is to assess the structural identifiability of candidates in the observed range of data, i.e. analyse whether the behaviour of candidate models differ in the domain for which the data is representative. A variety of analytical and numerical approaches to assess structural identifiability exists (Guillaume et al. 2019; Chis et al. 2012) including numerical parameter screening methods from sensitivity analysis (Campolongo et al. 2007; Troost & Berger 2015a).

In addition to uncertain model structures and parameters, also uncertain auxiliary assumptions must be documented and represented in parameters (e.g. error distributions for expected deviations, imputation to deal with incompleteness in the data, alternative choices in data curation, preparation or aggregation) that may decrease identifiability. 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, parameters that affect dynamic behaviour or accumulative development over several periods require panel data (Troost & Berger 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 & Berger 2015a).

3.2.5 Choosing adequate methods for model inference and measurement of predictive accuracy (Step 9)

If structural identifiability is given or direct generalisation is possible, one can choose an adequate method for data-driven model inference. If not, it is often still useful to measure sample predictive accuracy of candidates and compare it against a null model to ensure the models do completely go astray.

Inverse modelling employs algorithms to characterise the distribution of a loss function over candidates (exploration/estimation of posterior parameter distribution) or find the candidate with the optimal loss function value (optimisation, calibration). Available methods considerably differ in the extent to which uncertainty in the selection process is characterised and to which prior uncertainty is considered (Table 3).

Table 3 An exemplary selection of methods and measures used in model inference (inverse modelling) from observed system behaviour and their characteristics and premises

Method or Measure	**Purpose**	**Loss function**	**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 parameter	- correct model structure - correct formal likelihood that corresponds to the expected deviation of models - representative data	Hobbs & Hilborn 2006; Kukacka & Barunik, 2017; Lux, & 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 & 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 & 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 & Hilborn 2006; Hartig et al. 2011; Grazzini et al. 2017; Lux, & 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 & 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 & Ricchiardi 2015; Lux, & 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 & Railsback 2012, Gallagher et al. 2021
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	Hornberger & Spear 1980; Troost & Berger 2015a

Normalised Goodness-of-fit (Model efficiency)	- benchmark predictive accuracy of model	- Parametric likelihood or robust loss function	no		- loss function adequate to the form of deviations - meaningful benchmark model	Schaeffli & Gupta, 2007; Pontius & 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 basic method	- depends on basic method	- 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 & Celisse 2010; Vehtar et al. 2017; Browne 2000
Bootstrapping	- to be combined with other estimation method - estimate effect of sampling error on selection/estimation results	- depends on basic method	- depends on basic method	- Non-parametric estimate of effect of sampling error on estimates and predictive accuracy	- data is representative and sufficiently redundant for resampling	Efron & 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 basic method	- depends on basic method	- 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 & Mendelson 2002; Arlot & Celisse 2010

3.2.5.1 Adequate choice of loss function or likelihood

Loss functions 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 stronger changes in conclusions. Hence, in principle loss functions can be specified to directly reflect the precision, accuracy, relativity and symmetry required by the research question and penalise 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). In direct generalisation cases and when sampling error has been controlled for (e.g. by cross-validation, see below), the measured loss can also be directly generalised to target situations.

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 & Vrugt 2010): 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 & Heckman 1996; Blavatsky & Progrebna 2010). For example, if a model is designed to predict an upper bound, underestimation of observations should be penalised, overestimation not.[9]

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 & Matsuura 2005; Hyndman & 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 & Railsback 2012). In other cases, qualitative criteria are used to define binary-valued acceptance functions (Spear & Hornberger 1980; Troost & Berger 2015a).

Pure loss functions and likelihoods provide a relative ranking between candidate models, but their absolute values are specific to the sample used. Absolute goodness-of-fit measures (e.g. model efficiencies) 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 models 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. However, the sample average is only one possible choice for the null benchmark. Trend extrapolation, random allocation, or seasonal or group-specific averages can often be more adequate benchmarks

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

(Schaeffli & Gupta, 2007; Pontius & Millones 2011). As an alternative, Grimm & Railsback (2012) suggest to always explicitly include a benchmark null model among 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. Classical least-squares or maximum likelihood-based parameter estimation identify one best fitting model and quantify posterior uncertainty in the form of confidence intervals for parameters. This quantification is very limited: It presupposes 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). Moreover, while large confidence intervals point to low practical identifiability, they cannot conceptually be interpreted as posterior probabilities for parameters. Bayesian frameworks (Hobbs & Hilborn 2006) can overcome the latter limitations if formal prior probabilities and certain parametric likelihoods are specifiable.

Predictive accuracy measured in a sample is a biased measure of *expected predictive accuracy out-of-sample*: It favours models with a higher number of freely adaptable parameters, which increases the danger of overfitting. Adequate model inference requires correcting this bias: This can be achieved by the use of information criteria (AIC, BIC, DIC, WAIC) or appropriately specified prior probabilities in formal Bayesian frameworks (Burnham & Anderson 2004; Ward 2008; Vehtari et al. 2017). Both, require parametric likelihoods.

Cross-validation[10] and bootstrapping are the essential non-parametric alternatives to obtain unbiased estimates of expected predictive accuracy from a sample (Browne 2000; Arlot & Celisse 2010; Bennet et al. 2013; Vehtari et al. 2017). 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.

[10] 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.

Table 4 Guiding questions for model inference from observed system behaviour (inverse modelling, calibration, model selection)

Dimension	Guiding questions	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? Which parts of the model 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 inverse modelling) List of auxiliary parameters used to represent data and data preparation uncertainty Ranges or if available prior probabilities for these models and parameters List of alternative models and parameter ranges theoretically suitable, but not explored in the analysis List of critical assumptions for which no alternative assumptions will be tested during the further analysis
Step 8. Assess structural identifiability in the population represented by observed sample		
Structural identifiability	How large is the expected difference generated by two candidates in the observed domain? Are outcomes unique to a candidate or do different candidates produce the same outcome? If they are not identifiable: Have all suitable dimensions (variables) of the observed data been employed? Can we subdivide the model into components/parameter groups that are identifiable? Can we reparametrise the model by aggregating unidentifiable ones to identifiable ones without violating structural knowledge on parameter stability?	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 a) identified from a theoretical perspective (e.g. Guillaume et al. 2019; Chis et al. 2012) 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 & Berger 2015a)

Step 9: Adequate model inference and predictive accuracy measurement by inverse modelling (if applicable)		
Choice of loss function / acceptance criteria /predictive accuracy measure	*In direct generalisation cases:* Which prediction errors would have the strongest effect on conclusions? Does the loss function appropriately reflect this? *In indirect generalisation or for structure-focus:* Does the loss function appropriately weight errors by the expected deviations of the model candidate? 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? Consider formal likelihoods for well-specified models with tractable, well-defined error distributions. Consider indirect likelihoods based on summary statistics, robust loss functions, and 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 & Vrugt 2010; Hansen & Heckman 1996; Kukacka & Barunik, 2017; Lux, & Zwinkels, 2018 *Indirect/Approximate likelihoods*: Chen et al. 2012; Beaumont 2010; Drovandi et al. 2015; Grazzini & Richiardi 2015; Carrella et al. 2021; *Robust loss functions:* Willmott & Matsuura 2005; Troost & Berger 2015a (ABM example) *Qualitative criteria*: Pattern-oriented modelling (Grimm & Railsback 2012, Gallagher et al. 2021); Binary acceptance (Spear & Hornberger 1980) *Landscape metrics* (as qualitative criteria or summary statistics in approximate likelihoods): e.g. Hagen-Zanker (2009); Chen 2011; Pontius & Millones 2011; Van Vliet et al. 2013; McGarigal 2014
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	A first quick assessment whether practical identifiability can at all be expected and it is worth to try model inference from the data.

	candidate models can be disentangled (E.g. assess using variance inflation factors)? Is the whole domain well represented or are we likely to have a strong influence of outliers?	
Choice of method or algorithm for model/parameter inference, predictive accuracy and posterior uncertainty quantification	Does the method or algorithm chosen … a) … consider all (operational) alternative model formulations and parameter sets? b) … adequately consider prior evidence/probability of model structures and parameter values (if available)? 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?	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 model 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
Benchmarking	Is a proper benchmark null model (e.g. sample average, random allocation, trend extrapolation) included in the analysis? Either by explicit inclusion in the set of candidate models (Grimm & Railsback 2012) or implicitly in an absolute goodness-of-fit measure (model efficiency)? Does this benchmark model reflect the best simple alternative model or can it be replaced by a better benchmark (Schaeffli & Gupta, 2007; Pontius & Millones 2011)?	Outcome: Potentially: The expected predictive accuracy of the candidate models in predicting situations for which the available I/O data is representative put in relation to the expected predictive accuracy of a simple alternative model.

3.3 Adequate derivation and interpretation of simulation results and uncertainty

Figure 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[11], processes that have been ignored, potential exogenous events not considered, unformalised error terms, unforeseeable events, critical assumptions for which no alternatives are tested, etc.). If applicable and successful, model 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 uncertainty still need to be translated into *output or 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.

3.3.1 Interpretation of predictive accuracy and posterior uncertainty (Step 10)

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, care has to be taken when using this information to draw further conclusions, e.g. about the model being "sufficiently good" or the "correct" or "best explanation" (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 and its encompassing debate (Pontius & 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

[11] Brenner & 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 recognised as unmodelled uncertainty and appropriately discussed when deriving conclusions.

allows to identify 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 and recognise that the "best" model does not necessarily have or even approach a posterior probability of one (Troost & 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 a philosophical principle[12] 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 Analysis and interpretation of predictive uncertainty (Step 11)

Only in rare cases, it will be permissible to directly generalise expected predictive uncertainty from inverse modelling to the target situation (representative sample, negligible input uncertainty, one clearly best model). Generally, comprehensive uncertainty and sensitivity analysis is necessary. Uncertainty analysis must be global, i.e. cover the full range of potential input values including interactions and correlation between input factors (Saltelli & Annoni, 2010). A considerable number of approaches for efficient uncertainty analysis is available that adapt to different model complexities and available computational resources (Helton et al., 2006; Saltelli et al., 2008; Gramacy & Lee 2009; Troost et al. 2022). Stochastic models require sufficient repetitions and statistical comparison tests or, more efficiently, Common Random Numbers schemes to isolate systematic differences from stochastic ones (Stoute & Goldie 2008; Troost & Berger 2016).

[12] Parsimony as a philosophical 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).

Table 5 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)

Locations 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 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)

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 5 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. Probabilistic predictions require probability information in all locations. Simulation analysis can however also provide useful insights if uncertainty is high in one or all locations. It is key that exploration of predictive uncertainty focuses on the output quantity, precision and resolution relevant to answering the 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 systematic difference is much more relevant than the apparent one: 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 (Berger & Troost 2014) and strategies that are robust under many different scenarios and assumptions can be detected (Marchau et al. 2019; Lempert 2019).

Table 6 Guiding questions for the analysis of predictive uncertainty and the interpretation of results

Dimension	Guiding questions	Outcome
Step 10: Interpretation of posterior uncertainty and expected predictive accuracy *(if applicable)*		
Interpreting expected predictive accuracy (if measured)	What is the possible effect of sampling error on predictive accuracy (e.g. via cross-validation, bootstrapping, post-regression diagnostics for outlier influence etc.) and how does it influence interpretation? Is there a bias in predictions that points to systematic model error (disaggregate analysis of residuals)? How do model predictions compare with an appropriate benchmark model? Have the limits to generalisability (e.g. statements only relative to models included in the analysis and within the bounds of representativity of the sample used) been respected?	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, required precision derived from research question, long-term experience with similar models in similar situations)
Interpreting posterior uncertainty and the results of model inference	Does the posterior uncertainty – as measured – provide complete information about the effect of sampling error and practical identifiability of candidates? Considering identifiability, posterior uncertainty and unmodelled uncertainty: 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? 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 obtained from the analysis which properly reflects the associated posterior uncertainty and plausible alternative model formulations. *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: Predictive simulations and analysis of predictive uncertainty *(if the analysis is output-focused)*		
Design of predictive simulation experiments	Do simulations 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? (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 … controls for the effect of aleatory uncertainty (e.g. by common random numbers schemes, e.g. Troost & 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 & Annoni, 2010) at a sampling rate adequate for the computational resources … or alternatively a comprehensive search for non-robust outcomes or strong deviations over the global uncertainty space (e.g. destructive verification, Midgeley et al. 2007; stress testing and red-teaming, Lempert 2019).
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 and the unmodelled uncertainty with the precision and accuracy required by the research question: 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	A compact summary explaining … … the conclusions building on the comparison of model results and final uncertainty to research question requirements … a summary justification of model and method choice following the criteria and premises set out in the previous steps of this protocol … a documentation of prior, posterior and predictive uncertainty 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, e.g. following the schemes of NUSAP (van der Sluijs 2017; Kloprogge et al. 2011), sensitivity auditing (Saltelli et al. 2013) or Fischhoff & Davis (2014)'s protocol.

	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?	

3.3.3 Interpretation and conclusions (Step 12)

The interpretation of results should compare the final uncertainty to the required precision and accuracy of the research question. 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 (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. 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 and unmodelled uncertainty 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 & 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; Fischbach & Davis 2014; van der Sluijs 2017) and additional effort to assess the robustness of decision option outcomes to these assumptions is essential (assumptions-based planning, stress testing, red teaming; Lempert 2019; Marchau et al. 2019).

4 Discussion and conclusions

Adequate conclusions from simulation analysis require a careful analysis of the logical argumentative structure and the critical premises they 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. Even if – as we demonstrated – premises in the overarching argumentative structure vary, the preconditions for the use of specific methods of analysis are invariable and their violation makes the analysis inadequate. For example, empirical output validation and inverse modelling 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 Table 3).

A number of previous studies (e.g. Edmonds et al. 2019; Epstein 2008) highlighted how different modelling purposes require different data and methods. In the presented protocol we have moved beyond discrete typologies of model purpose and define concrete dimensions of research question and available system knowledge and data that together characterise the modelling context. Typologies of Edwards et al. (2019), and especially terms such as prediction, forecast, projection or exploration, whose understanding and usage differ between and sometimes even within disciplines (Bray & von Storch 2009), can be mapped onto these dimensions to allow for more precise communication (see Appendix A.1).

When understood comprehensively, the process of ensuring adequate model conclusions is, however, more complex and subtler than a single-step matching of context type to a method. Rather it is a process that is *hierarchical*, i.e. outcomes of earlier steps affect choices in later steps (e.g. inverse modelling should not be pursued without first ensuring representative data and structural identifiability). It is *recursive*, i.e. in composite models the context of each component must be assessed, and *iterative,* i.e. outcomes of subsequent steps may encourage receding a number of steps and reconsider 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 further possible to answer a more restricted question that is already useful where the context does not allow to reliably answer the original question as presented above in our initial example.

The KIA protocol that we suggest 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 statement of 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 - 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 as a template for well-structured tabular documentation in an appendix.

We strived to be general in redacting the protocol. We do not advocate one common method for all ABM, rather the dimensions of the modelling context that we introduced are intended to help identify which ABM applications share a similar modelling context and might learn from each other and which not. For example, Troost & Berger (2015a) and Carrella et al. (2021) 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. to avoid discussions about an appropriate loss function, when structural identifiability is the more important issue; to avoid overemphasis on separation of training and validation data, when validation data is not representative for target situations; to avoid discussions about unreliability of unconditional predictions when these are neither possible nor necessary; to avoid suggesting model simplification to increase practical identifiability when model complexity is required for structural reasons and direct generalisation is not adequate, etc.

The KIA 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 concretises the principle “as empirical as possible, as general as necessary” coined for ABM by Brenner & 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 & 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).

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.

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

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 also provide guidance for non-ABM simulation when facing similar challenges.

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).

References

Alexandrov, G. A., Ames, D., Bellocchi, G., Bruen, M., Crout, N., Erechtchoukova, M., ... & Samaniego, L. 2011. Technical assessment and evaluation of environmental models and software. Environmental Modelling & 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. Softw. 80, 113–121. doi:10.1016/j.envsoft.2016.02.023

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. doi:10.1002/2014WR015382

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

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., Population Models for Ecological Risk Assessment of Chemicals 280, 117–128. doi: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. doi:10.1016/j.ecolmodel.2006.11.005

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

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

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, *Ecology and Society*, 15(2), 24.

Bartlett, P. L., & Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, 3, 463-482.

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

Beaumont, M.A., 2010. Approximate Bayesian Computation in Evolution and Ecology. Annual Review of Ecology, Evolution, and Systematics 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. doi: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. Environmental Modelling & Software 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Berger, T., Schilling, C., Troost, C., Latynskiy, E., 2010. Knowledge-Brokering with Agent-Based Models: Some Experiences from Irrigation-Related Research in Chile. In: David A. Swayne,

Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.), 2010 International Congress on Environmental Modelling and Software, Ottawa, Canada.

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

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.

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

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

Beven, K.J., Smith, P.J., Freer, J.E., 2008. So just why would a modeller choose to be incoherent? Journal of Hydrology 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

Brenner, T., Werker, C., 2007. A Taxonomy of Inference in Simulation Models. Comput. Econ. 30, 227–244. doi: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. doi: 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: Climate Change, 8(2), e448.

Browne, M.W., 2000. Cross-Validation Methods. Journal of Mathematical Psychology 44, 108–132. doi: 10.1006/jmps.1999.1279

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in Model Selection. Sociological Methods & Research 33, 261–304.

Caldwell, B.J., 1991. Clarifying Popper. Journal of Economic Literature 29, 1–33.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environmental Modelling & Software 22, 1509–1518. https://doi.org/doi: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 Policy 109, 105618. doi:10.1016/j.landusepol.2021.105618

Chen, Y., 2011. Derivation of the functional relations between fractal dimension of and shape indices of urban form. Computers, Environment and Urban Systems, 35(6), 442-451.

Chen, S.-H., Chang, C.-L., Du, Y.-R., 2012. Agent-based economic models and econometrics. The Knowledge Engineering Review 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

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. doi:10.1007/s11625-021-00989-w

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. Simul. 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. 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. Statistical Science, 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. and Squazzoni, F., 2019. Different Modelling Purposes. Journal of Artificial Societies and Social Simulation 22 (3) 6. doi: 10.18564/jasss.3993

Efron, B., Tibshirani, R., 1997. Improvements on Cross-Validation: The 632+ Bootstrap Method, Journal of the American Statistical Association, 92:438, 548-560, doi: 10.100/01621459.1997.10474007

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, 16226–16226. https://doi.org/10.18174/sesmo.2020a16226

Engle, R. F. and D. F. Hendry, 1993. Testing Super Exogeneity and Invariance in Regression Models, Journal of Econometrics 56, 119–139.

Epstein, J. M., 2008. Why model?. Journal of Artificial Societies and Social Simulation, 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. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics in: Proceedings of Machine Learning Research 54:1486-1494 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. doi: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. Softw. 45, 1–7. doi: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. Softw. 75, 333–347. doi:10.1016/j.envsoft.2015.04.003

Fischhoff, B., Davis, A.L. 2014. Communicating scientific uncertainty. Proceedings of the National Academy of Sciences, 111 (4): 13664–13671.

Forster, M. 2000. Key Concepts in Model Selection: Performance and Generalizability. Journal of Mathematical Psychology 44, 205-231. doi: 10.1006/jmps.1999.1284

Forrester, J.W., Senge, P.M., 1980. Tests for Building Confidence in System Dynamics Models, in: Legasto, A.A., Jr., Forrester, J.W., Lyneis, J.M. (Eds.), System Dynamics, TIMS Studies in the Management Sciences. North-Holland, New York, Amsterdam, 209–228.

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

Gallagher, Cara A., Magda Chudzinska, Angela Larsen-Gray, Christopher J. Pollock, Sarah N. Sells, Patrick J. C. White, and Uta Berger. n.d. "From Theory to Practice in Pattern-Oriented Modelling: Identifying and Using Empirical Patterns in Predictive Models." *Biological Reviews* n/a (n/a). Accessed May 27, 2021. https://doi.org/10.1111/brv.12729.

Gass, S.I., 1983. Decision-Aiding Models: Validation, Assessment, and Related Issues for Policy Analysis. Oper. Res. 31, 603–631.

Gangl, M, 2010. Causal Inference in Sociological Research. Annual Review of Sociology. 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, Computers, Environment and Urban Systems 78

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

Grazzini, J., Richiardi, M. G., Tsionas, M. 2017. Bayesian estimation of agent-based models. Journal of Economic Dynamics and Control, 77, 26-47.

Gramacy, R. B., Lee, H. K. H., 2009. Adaptive Design and Analysis of Supercomputer Experiments. Technometrics, 51, 130-145

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. doi: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. doi:10.1016/j.ecolmodel.2010.08.019

Grimm, Volker, and Steven F. Railsback. 2012. "Pattern-Oriented Modelling: A 'Multi-Scope' for Predictive Systems Ecology." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1586): 298–310. https://doi.org/10.1098/rstb.2011.0180.

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., Population Models for Ecological Risk Assessment of Chemicals 280, 129–139. doi: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. Simul. 23, 7.

Gore RJ, Lynch CJ, 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, The Review of Financial Studies, 27(2): 617–661, doi: 10.1093/rfs/hht047

Guillaume, Joseph H. A., John D. Jakeman, Stefano Marsili-Libelli, Michael Asher, Philip Brunner, Barry Croke, Mary C. Hill, et al. 2019. "Introductory Overview of Identifiability Analysis: A Guide to Evaluating Whether You Have the Right Type of Data for Your Modelling Purpose." *Environmental Modelling & Software* 119 (September): 418–32. https://doi.org/10.1016/j.envsoft.2019.07.007.

Hagen-Zanker, A., 2009. An improved fuzzy kappa statistic that accounts for spatial autocorrelation. International Journal of Geographical Information Science, 23(1): 61-73

Hansen, L.P., Heckman, J.J., 1996. The empirical foundations of calibration. Journal of Economic Perspectives 10, 87–104.

Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., & Huth, A., 2011. Statistical inference for stochastic simulation models - theory and application. Ecology letters, 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. Softw. 68, 196–204. doi: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. doi: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. The Economic and Social Review, 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. Geographical Analysis, 53:1, 76-91.

Hobbs, N. T. and Hilborn, R., 2006. Alternatives to statistical hypothesis testing in ecology: A guide to self teaching, Ecological Applications 16, 5–19.

Hyndman, R. J. and Koehler, A. B., 2006. Another look at measures of forecast accuracy, International Journal of Forecasting 22, 679–688

Jager, KJ, Tripepi, G, Chesnaye, NC, Dekker, FW, Zoccali, C, Stel, VS. 2020. Where to look for the most frequent biases? Nephrology, 25: 435-441. Doi: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. Softw. 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. Atmospheric Science Letters, 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. Environmental Modelling & Software, 26(3), 289-301.

de Koning, K. , Filatova, T., 2020. Repetitive floods intensify outmigration and climate gentrification in coastal cities, Environmental Research Letters, 15 034008, doi: 10.1088/1748-9326/ab6668

Kukacka, J., Barunik, J., 2017. Estimation of financial agent-based models with simulated maximum likelihood. Journal of Economic Dynamics and 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. Bulletin of the American Meteorological Society, 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. Environmental Modelling & Software 27-28, 83-96.

Leamer, E., 2010. Tantalus on the way to Asymtopia. Journal of Economic Perspectives, 24 (2): 31-46

Lempert , R., 2019. Robust Decision Making (RDM). In: Marchau, V., Walker, W., Bloemen, P., Popper, S. (Ed.), 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 Sensitivity Analysis of Agent-Based Models, Journal of Artificial Societies and Social Simulation 23:1:6, doi: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 K Brunner and A. Meltzer (eds.), The Phillips Curve and Labor Markets, Vol. 1 of Carnegie- Rochester Conferences on Public Policy, pp. 19-46, Amsterdam: North-Holland Publishing Company.

Lux, T., Zwinkels, R. C., 2018. Empirical validation of agent-based models. In: Handbook of computational economics Vol. 4, 437-488. Elsevier.

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?, Journal of Farm Economics 47: 5, 1381-85. doi:/10.2307/1236396.

Manski, C.F., 2019: Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing, The American Statistician, 73:sup1, 296-304, doi: 10.1080/00031305.2018.1513377

Marchau, V., Walker, W., Bloemen, P., Popper, S., 2019: Introduction. In: Marchau, V., Walker, W., Bloemen, P., Popper, S. (Ed.), 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. doi: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. The American Economic Review 75 (2): 201-5

McGarigal, K., 2014. Landscape Pattern Metrics. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). 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. doi: 10.1016/j.jbusres.2007.02.004

Moss, S., Edmonds, B., 2005. Towards Good Social Science. J. Artif. Soc. Soc. Simul. 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. Agricultural Systems 195, 103315. doi: 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. doi: 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. Environmental Modelling & 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. European Journal of Operational Research, 254 (2): 517-531, doi: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., ..., Malanson, G., 2008. Case studies, cross-site comparisons, and the challenge of generalization: comparing agent-based models of land-use change in frontier regions. Journal of Land Use Science, 3(1), 41-72.

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

Pielke, R.A., 1991. A recommended specific definition of "resolution". Bulletin of the American Meteorological Society, 72, 1914-1914. doi: 10.1175/1520-0477-72.12.1914.

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. doi: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. International Journal of Remote Sensing 32, 4407–4429. https://doi.org/10.1080/01431161.2011.552923

Rosenzweig, M., Udry, C., 2016. External Validity in a Stochastic World. NBER Working Paper 22449. doi: 10.3386/w22449

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. Mark. 28, 181–193. doi:10.1016/j.ijresmar.2011.04.002

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. Softw. 25, 1508–1517.

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.

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. doi: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 in Ecology & Evolution 25, 479–486. doi: 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 Resources Research 46, 1–17.

Schreinemachers, P., Berger, T., 2011. An agent-based simulation model of human environment interactions in agricultural systems. Environmental Modelling & Software 26, 845-859. doi: j.envsoft.2011.02.004

Schulze, J., Müller, B., Groeneveld, J. and Grimm, V., 2017: Agent-Based Modelling of Social-Ecological Systems: Achievements, Challenges, and a Way Forward, Journal of Artificial Societies and Social Simulation 2017 Vol. 20 Issue 2 Pages 8, doi: 10.18564/jasss.3423

Siebers, O.P., Macal, M.C., Garnett, J., Buxton, D., Pidd, M., 2010. Discrete-event simulation is dead, long live agent-based simulation! J. Simul. 4, 204–210. doi: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. doi: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 Research 14, 43–49. https://doi.org/10.1016/0043-1354(80)90040-8

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

Stigler, S. M., 2007. The epic story of maximum likelihood. Statistical Science, 598-620.

Stigter, J. D., M. B. Beck, and J. Molenaar. 2017. “Assessing Local Structural Identifiability for Environmental Models.” Environmental Modelling & Software 93 (July): 398–408. https://doi.org/10.1016/j.envsoft.2017.03.006.

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

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. Simul. 17, 11. doi: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. American Journal of Agricultural Economics 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, August 9-14, 2015, Milan, Italy. doi: /10.22004/ag.econ.211929

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. Accepted for publication in Computational Economics (09 May 2022). doi: 10.1007/s10614-022-10276-0

Vandecasteele, L., Debels, A., 2007. Attrition in Panel Data: The Effectiveness of Weighting, European Sociological Review, 23(1): 81–97, doi: 0.1093/esr/jcl021

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., J. van Vliet, D.T. Rutledge, M.J. Kirkby, 2011. Comparison of scale and scaling issues in integrated land-use models for policy support. Agriculture, Ecosystems and Environment, 142(1-2), 18-28

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. doi: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. Ecological Modelling: 261-262: 32-42.

Vehtari, A., Gelman, A. & Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27, 1433. doi: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. Softw. 81, 72–85. doi:10.1016/j.envsoft.2016.03.011

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. Softw. 25, 1268–1281. doi: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. Environmental Modelling & Software, 77, 196-220, doi:10.1016/j.envsoft.2015.11.016.

Ward, E. J., 2008. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. Ecological Modelling, 211(1-2): 1-10.

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 Assessment, 4(1), 5–17.

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, Climate Research 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., and Brook, B. W., 2021. Parsimonious model selection using information theory: a modified selection rule. Ecology 102 (10): e03475. 10.1002/ecy.3475

Appendix A

A.1 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 Edwards 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 the three latter 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 requires 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 and appropriate forms of simulation analysis (e.g. Marchau et al. 2019) can help in this context.