MIND STEP

MODELLING INDIVIDUAL DECISIONS TO SUPPORT THE EUROPEAN POLICIES RELATED TO AGRICULTURE

Deliverable Report D6.1
Report on options for quality management, validation requirements & suitability of validation tools

EDITED BY	Marten Graubner (IAMO)
APPROVED BY WP MANAGER:	Franziska Appel (IAMO)
DATE OF APPROVAL:	15 September 2022
APPROVED BY PROJECT COORDINATOR:	Hans van Meijl (WR)
DATE OF APPROVAL:	Day Month Year

CALL H2020-RUR-2018-2	Rural Renaissance
WORK PROGRAMME Topic RUR-04-2018	Analytical tools and models to support policies related to agriculture and food - RIA Research and Innovation action
PROJECT WEB SITE:	https://mind-step.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

This page is left blank deliberately

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

LIST OF FIGURES

LIST OF TABLES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

EXECUTIVE SUMMARY

The MIND STEP model toolbox will link and combine a variety of models designed for specific uses. Because of the heterogeneity of tools and applications within MIND STEP, validation requires a model-specific approach. These individual efforts represent an integral part of the overall evaluation of the MIND STEP toolbox additional techniques or their repeated application might be required given the level of model integration. The present deliverable reports on options for quality management and model evaluation based on a comprehensive literature review. Beyond traditional criteria as data quality and comprehensive documentation, quality management of research software features also aspects of the design of models (e.g., modularity) and the maintenance of models (in terms of providing sustainable use). Similarly, confronting model outputs with observational evidence, historically often the only and still an important tool for validation, needs to be accompanied by additional validation techniques, in particular, efficient interaction between modelers and stakeholders. This deliverable also presents the common framework and indicator system for model validation for the MIND STEP toolbox. The framework shall help to assess model quality and validity by detailing, e.g., the purpose and underlying concepts of the model, model evaluation (including verification, validation, and calibration) as well as transparent model documentation and resource accessibility (e.g., of data and/or source code).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

1. INTRODUCTION

The MIND STEP model toolbox will consist and combine a variety of models, each designed for a specific use. The underlying modeling approaches stretch from direct empirical investigations to agent-based systems. In each and every case, model validation is an important task to assess the usefulness of a model and ultimately its credibility (Baldos & Hertel, 2013; Rykiel, 1996), which is a cornerstone for evidence based policy assessment (Panhans & Singleton, 2017). Because of the large variety of tools and their tasks, validation generally requires a model-specific approach (Barlas & Carpenter, 1990) but, striving for an (partially) integrated framework of the „model toolbox", we need to account for this concept and provide a more comprehensive approach to validation.

The aim of Task 6.1 is to support validation across the work packages in MIND STEP and to provide a guideline and identification of indicators to assess model validity on different scales: The common framework and indicator system for model validation. This framework aims to ensure, e.g., that a) tools are suitable to investigate relevant policy measures, b) cover key indicators of interest, and c) provide valid results. Based on different concepts and perspective on model validation, existing guidelines in the literature, as well as experience and results of the project partners, this task developed a checklist with quality criteria and indicators for model validation, which will serve as a mean of quality management within MIND STEP.

The focus of this report is on simulation models as those build the core of the MIND STEP toolbox. Furthermore, simulation models in general and agent-based simulations in particular are increasingly used in impact assessment of agricultural policies (Fresco et al., 2021; Huber et al., 2018; Kremmydas et al., 2018; Reidsma et al., 2018) but there are no accepted standards for the validation of complex simulation models Jakeman et al. (2006); Kaye-Blake et al. (2014); Marks (2013). The use of any model to support any type of decision making requires models and their results to be credible (Rykiel, 1996). Model evaluation, including model validation and verification as well as the effective communication between the modeling team and the stakeholders, is the foundation to establish credibility (Bharathy & Silverman, 2013; Schlesinger, 1979) especially if counter intuitive results are generated (Smajgl et al., 2011). If computer modeling can exploit its potential, it could support better decision making (Bankes, 1993) but often model evaluation is insufficient (Janssen & Van Ittersum, 2007). For instance, validation, often does not follow formal, objective, and quantitative procedures (Barlas & Carpenter, 1990; Schwanitz, 2013) or, as van Vliet et al. (2016) found by reviewing studies on land change models, authors often do not report any validation at all.

Because of this and the importance for model evaluation to establish credibility, there is a clear need for more attention towards validation in practice (Beisbart, 2019). Bousquet and Le Page (2004) argue there are two strategies to enhance the credibility of (simulation) models: (a) Provide a comprehensive and transparent presentations of the structure and foundation (including underlying theories, concepts, and assumptions) of the model. (b) Compare the results of the simulation with other (types of) models or observational data. Roughly speaking, the first is the concern of quality management while the second is the concern of model validation. Over the last couple of years, there are increasing efforts in formalizing quality management and validation of simulation models, also in agricultural economics (Anzt et al., 2021; Bert et al., 2014; Britz et al., 2021; Fagiolo et al., 2019; Reidsma et al., 2018; Schwanitz, 2013). The present deliverable provides a report on options for quality management and validation for the MIND STEP project based on a comprehensive review of this literature.

The deliverable is structured as follows: The next section provides a brief overview of selected models used and further developed in the MIND STEP project to highlight the variety of approaches and potential applications. Chapter 3 discusses aspects of quality management with particular focus on the validation of simulation models and chapter 4 introduces concepts and approaches of model validation. Chapter 5 presents the common framework and indicator system for model validation within the MIND STEP project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

2. MODELS

This section provides a brief overview of selected models that will be used or developed within the project to highlight the variability of models that motivate the discussion of the (alternative) validation approaches and techniques in later sections. The models are ordered alphabetically.

2.1. AgriPoliS (Agricultural Policy Simulator)

AgriPoliS is a spatially explicit and dynamic agent-based model that is able to simulate the evolution of agricultural structures over time. It is mainly used to study the influence of policies on agricultural structural change (Happe et al., 2006). AgriPoliS can be calibrated with empirically collected data for real regions and contribute to a better understanding of past and future structural change. In AgriPoliS, individual farm agents are assumed to maximize profits or household income by use of a mixed-integer programing model, and are able to react to price or policy changes by renting or leasing land, changing their production system, or choosing to quit agriculture. These individual farm agents compete for land with their neighbors by interacting on the land market, which is implemented as a repeated auction.

2.2. AGRISPACE

AGRISPACE is a research activity to build a state-of-the-art agricultural sector model for Norway in order to analyze impacts of market and policy changes on the agricultural sector and farm structural change in Norway. It is a joint initiative of the Norwegian Institute of Bioeconomy Research in Oslo and the Institute for Food and Resource Economics of the University Bonn, drawing on a long-standing research co-operation between the two institutions. The development of the model was part of the AGRISPACE project (1/2014-10/2017), financed by the Norwegian Research Council, and involving several Norwegian and international partners. AGRISPACE consistently combines production, factor use and exit decisions for all individual farms in Norway with a regionalized partial equilibrium model. As such AGRISPACE also acts as a test-bed for the integration of IDM data and models in current models like MAGNET (see 2.6).

2.1. CAPRI

The Common Agricultural Policy Regional Impact (CAPRI) model is a global partial equilibrium model for the agricultural sector. It has been designed for ex-ante impact assessment of agricultural, environmental and trade policies. It iteratively links a supply module, focusing on the EU, Norway, Turkey and Western Balkans, with a global multi-commodity market module. The CAPRI model can be used for policy anticipation and formulation. It allows economic and environmental analysis of different policy scenarios regarding reforms of the Common Agricultural Policy (CAP). It is able to perform a regional level analysis of specific Common Market Organisations (e.g. sugar, dairies), trade of agricultural goods with the rest of the world (e.g. WTO proposals), environmental policies (e.g. greening, climate action and water) and different subsidy schemes in Europe (e.g. partial decoupling of agricultural subsidies). The model is frequently used in various Commission services (such as DG AGRI, DG ENV, DG CLIMA, Eurostat and the JRC) reporting on agricultural, environmental and climate policies at the regional dimension in the EU.

2.2. Econometric Models

To complement the simulation models, several econometric models will be used or developed within MIND STEP, including innovative micro-econometric production choice models for empirically analyzing farmers' crop management choices (WP3) and models that provide estimates of market power parameters or price transmission elasticity (WP4). Using Dutch farm accountancy data, a Data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

envelopment analysis (DEA) investigates to what extent can circular dairy farms improve its greenhouse gas (GHG) cycle and nutrient cycle as well as its productivity simultaneously by reallocating land between crop production and livestock grazing on the farm. The DEA model, an extension of the model by Ang and Kerstens (2016), will provide empirical support for the FarmDyn model (see below). Using the dual approach allows to compute the shadow price of GHG emissions, that is, the farmers' willingness to pay to give up one unit of GHG emission. The shadow price can be used to calibrate the objective function of the FarmDyn model.

2.3. FarmDyn

FARMDYN as quite detailed bio-economic farm model provides a flexible, modular template to simulate farms with different branches (dairy, suckler cows, beef fattening, pig fattening, piglet production, arable farming, biogas plants). The model is parameterized for regional conditions in Germany and The Netherlands using highly detailed farm planning data in combination with farm structural statistics. The model is realized in GAMS, solved with the industry MIP solver CPLEX, linked to a Graphical User Interface realized in GGIG and hosted on a Software Versioning System. Design of experiments, building on R routines directly called from GAMS, can be used in combination with farm structural statistics to systematically simulate different farm realizations (assets, farm branches) and boundary conditions such as input and output prices or emissions ceilings using a computing server to solve several instances in parallel.

2.4. GLOBIOM

IIASA's Global Biosphere Management Model (GLOBIOM) is used to analyze the competition for land use between agriculture, forestry, and bioenergy, which are the main land-based production sectors. As such, the model can provide scientists and policymakers with the means to assess, on a global basis, the rational production of food, forest fiber, and bioenergy, all of which contribute to human welfare. GLOBIOM has been developed and used by IIASA since the late 2000s. The partial-equilibrium model represents various land use-based activities, including agriculture, forestry and bioenergy sectors. The model is built following a bottom-up setting based on detailed grid-cell information, providing the biophysical and technical cost information. This detailed structure allows a rich set of environmental parameters to be taken into account. Its spatial equilibrium modelling approach represents bilateral trade based on cost competitiveness. The model was initially developed for impact assessment of climate change mitigation policies in land-based sectors, including biofuels, and nowadays is also increasingly being implemented for agricultural and timber markets foresight, and economic impact analysis of climate change and adaptation, and a wide range of sustainable development goals.

2.5. MAGNET

MAGNET is a recursive dynamic, multi-region, multi-sector Computable General Equilibrium model used to analyze policy scenarios on agricultural economics, bioeconomy, food security, climate change and international trade. It was developed by the Wageningen Economic Research (WECR) in cooperation with JRC and the Thunen Institute. MAGNET is calibrated to the GTAP database and describes production, use and international trade flows of goods and services and primary factor use differentiated by sectors. The database distinguishes 141 countries or regions (including all EU member states), 65 sectors (plus several optional MAGNET-specific extensions) and 8 factors (e.g., labor, capital, land). A distinguishing feature of the model is its modular design which allows tailoring its structure to the research question. The GTAP model forms the MAGNET core while users choose among several extensions: different nesting structures or assumptions about factor markets, different agricultural-, trade- and biofuels-policy mechanisms and different assumptions relating to investment allocation. Other modules deal with the representation of the Common Agricultural Policies (including rural

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

development), land and labor supply, production quotas, tariff rate quotas, biofuels directive, bioenergy policies, water in agriculture, GHG emissions (marginal abatement curves) and tracking of Sustainable Development Goals (SDGs) to name a few. MAGNET can be used in policy formulation through ex-ante policy analysis. The model assesses policy scenarios related to agriculture and agri-food trade while taking into account other fields directly connected with agri-food production such-as bioeconomy (bioenergy, biofuel, biobased chemicals), sustainable use of resources (land and water), food security and nutrition (developing and developed countries) and climate change, but also feedback with the wider (non-agricultural) economy. Policy scenarios are compared against a baseline including the most recent macroeconomic (GDP and population) and agricultural (yields, land productivity, EU agricultural mid-term outlook) exogenous drivers. Focusing on ex-ante policy analysis, the model can be used to support policy formulation or to provide valuable information to policy makers in front of exogenous shocks.

2.6. IFM-CAP

IFM-CAP is a micro model designed for the ex-ante economic and environmental assessment of the medium-term adaptation of individual farmers to policy and market changes. IFM-CAP was developed by JRC in close cooperation with DG AGRI starting from 2013 for the purpose to improve the quality of agricultural policy assessment upon existing aggregate models and to assess distributional effects of policies over the EU farm population. Rather than providing forecasts or projections, the model aims to generate policy scenarios, or what if analyses. It simulates how a given scenario, for example, a change in prices, farm resources or environmental and agricultural policy, might affect a set of performance indicators important to decision makers and stakeholders. IFM-CAP is a comparative static positive mathematical programming model applied to each individual farm from the Farm Accountancy Data Network (FADN) to guarantee the highest possible representativeness of the EU agricultural sector. Farmers are assumed maximizing their expected utility at given yields, product prices and CAP subsidies, subject to resource endowments and policy constraints. The main strengths and capabilities of the model include the possibility to conduct a flexible assessment of a wide range of farm-specific policies and to capture the full heterogeneity of EU commercial farms in terms of policy representation and impacts (e.g. small versus big farms).

3. QUALITY MANAGEMENT

Quality management provides the tools to assess, monitor, and control the degree of product and/or process quality indicated by a set of desired characteristics (Balci, 2003, 2004). With respect to the MIND STEP project, this concerns mainly the quality of (simulation) models as well as data, where validity of both is a prime quality criteria and the focus of this report. The following subsections discuss the important aspects of data validity and good practice before a more detailed presentation of model validation is provided in chapter 4.

3.1. Data quality

Data validity is often not considered to be part of model validation (Sargent, 2013), but certainly a precondition for it and not only an issue in statistical validity (Anselin, 1988). Incomplete data or inconsistency in datasets can invalidate the results of any model and impair model credibility (Macal & North, 2005). Data validity certifies that the data meet a specified standard, i.e., quality assurance and quality control (Rykiel, 1996). Therefore, the development and use of adequate procedures is required for (i) collecting and maintaining data, (ii) testing the collected data, and (iii) screening the data for outliers (Sargent, 2013).
As models often rely on external data sources, some of these procedures are external to the MIND STEP project. For instance, a number of models (e.g., IFM-CAP and Farmdyn, see section 2) use the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Farm Accountancy Data Network (FADN), i.e., validated microeconomic data based on harmonized bookkeeping principles that allow to compare economic indicators across European regions. In general, criteria and procedures to ensure data quality for official statistics are typically provided by the European Commission and individual member states. For instance, the statistical office of the European Union, Eurostat, assesses data quality based on defined indicators including relevance, accuracy, timeliness and punctuality, accessibility and clarity as well as comparability and coherence along three aspects: (1) the characteristics of the statistical product (2) the perception of the statistical product by the user and (3) characteristics of the statistical production process (Bergdahl et al., 2007). The Federal Statistical Office in Germany, Destatis, just recently published a Handbook on data quality differentiating five levels of quality management that define (from level to level increasingly specific) requirements on data quality (Destatis, 2021).
Data acquisition and processing within a project can be time consuming and combining data from different data sources can cause additional issues, e.g., inconsistency in references (of data fields) or erroneous conversion of data to common units (Macal & North, 2005). Therefore, MIND STEP also develops and uses procedures for data acquisition, processing, and transfer. This is done mainly in work package 2 (Data requirements for indicators on European policies related to agriculture and data management) and work package 7 (ICT platform for MIND STEP and the MIND STEP model toolbox). For instance, Deliverable D2.2 provides a comprehensive handbook to build a conceptual framework for database interfaces to integrate data from multiple heterogeneous sources at flexible geographic and regional scales and to support analytical reporting as well as to allow structured and/or ad hoc queries (Gocht et al., 2021). Deliverable D7.6 describes the initial prototype of the MIND STEP data and download services to deliver and visualize the various geo-spatial results produced by the MIND STEP modeling teams to facilitate the transfer with stakeholders, the research community, and the public (McCallum & Subash, 2021).

3.2. Good practice and sustainable research software

As Jakeman et al. (2006) emphasize, good practice in the development of complex (simulation) models is crucial, because of the inherent difficulties in validating them. Deliverable 3.1 ”Specification of model requirements - Protocols for code and data” details some aspects of quality management within MIND STEP, including coding conventions and testing strategies (Mueller et al., 2021). The first concerns syntactic and code commenting guidelines such that the model code is understandable and maintainable independent of the original developer. In this respect, parameters, variables as well as dependencies can be described by commenting within the code. Mueller et al. (2021) further state, the implementation and documentation of testing strategies is crucial in multi-partner projects where linkages between models exist or several partners develop or work with the same model. As it will be discussed in section 4, testing shall ensure the correctness of the (computational) model and the credibility of the results. Mueller et al. (2021) propose that tests are documented, follow a protocol, and are periodically re-evaluated to ensure that new components of the model are included. An exemplary testing strategy (for the model FarmDyn) is also presented in Mueller et al. (2021).

Janssen and Van Ittersum (2007) define good practice based on a literature review focusing on bio-economic farm models for impact assessment. The authors state that a clear definition for the use of the model needs to be given. Furthermore, the model input both in terms of data and assumptions, e.g., the considered (farm) activities, needs to be described and model evaluation should be explicitly and comprehensively presented. Finally, model constraints should be mentioned and discussed. Because Janssen and Van Ittersum (2007) argue that transferability, i.e., model application to different regions and farm structures is an important feature, they call for a generic and modular structure of farm models to be developed. Reidsma et al. (2018) revisit the topic and review the use and the development of farm models for policy impact analysis. The authors conclude that, even though some progress has been made with respect to the criteria formulated by Janssen and Van Ittersum (2007),

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

this progress is limited especially regarding model evaluation and the development of generic, modular and easily transferable models, which limits the use of model results in policy-making and the re-use of improvements in modeling. Regarding the first issue, Reidsma et al. (2018) reinforce the call for thorough and consistent model evaluation and model comparison, with increased attention for model sensitivity and uncertainty. They also argue that the organization of a network of modelers as well as synthesizing research evidence into systematic reviews as an institutional element can support efficient communication at the science-policy-interfaces for agricultural systems. The authors also highlight that improved and timely data (collection) is crucial and advocate for stronger science-policy interaction, moving from a research-driven to a user-driven approach. Similarly, Eker et al. (2018) argue that empirical data most often play the lead role in model evaluation but qualitative and participatory approaches can enhance the usefulness and public reliability of models and their results.

As documentation and model quality are eminently important, Anzt et al. (2021) generally define mandatory criteria for transparency and quality of research software:

1. Source code should be publicly available.
2. Version control with meaningful commit messages and linked to an issue tracker needs to be used.
3. The license under which the software is distributed must be defined.
4. Documentation of the software needs to be publicly available comprising both user documentation (requirements, installation, getting started, user manual, release notes) and developer documentation.
5. Dependencies on libraries and technologies must be defined.

Additional criteria, according to Anzt et al. (2021), include: the availability of examples (comprising input data and reference results), interoperability (APIs / common and open data formats for input and output), and mechanisms for extensibility (modularity).

With respect to modularity, Britz et al. (2021) define requirements for the functionality and implementation of modular (bio-economic farm) models including necessary model features, model design and shared development. The authors define modularity as the option to replace, activate, or omit blocks of code holding equations and related variables, depending on the application of the model. The advantage is that this structural feature of models can ease their application by restricting data preparation, parameterization, model solving, and reporting to the actual use case, e.g., only the relevant farm branches, farming systems or policies are considered (Britz et al., 2021). With respect to the design of such models, the authors build on concepts from software engineering including „low coupling and high cohesion", where coupling refers to low dependencies between modules and cohesion refers to a strong dependence between elements within a module (Stevens et al., 1974). Britz et al. (2021) list a number of implications and advantages with respect to quality management: (1) Transparency: the model can be reviewed module by module, facilitating overall comprehension and quality control. (2) Maintainability: Code and database updates of a module do not affect other modules. (3) Extensibility: Modules can be extended or added to the core model without affecting others. (4) Distributed development: Modelers focus on specific modules which eases coordination of coding efforts.

One issue that might hamper the development of modular models includes the requirement of a clear strategy for model maintenance and distribution, including the use of version control, testing strategies, and documentation as part of quality management (Britz et al., 2021). On the one hand, models are often (and necessarily) developed for a specific use and thus rather a case study (Bert et al., 2014; Troost, 2015). Despite the fact that computational analysis, simulation and software are increasingly important in research, there are considerable inefficiencies because models are often not used or developed beyond a prototype stage (Appel & Loewe, 2021; Britz et al., 2021). On the other

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

hand, researchers tend to use models they have experience with, and rarely switch to competing model frameworks even if those offer some advantages or are more adequate given the study objectives (Addor & Melsen, 2019). In this respect, Appel and Loewe (2021) call for changes in the way research software is developed and maintained, including funding, structural and infrastructural support (e.g., project management), and legal considerations (e.g., licensing).

While most of the presented aspects of quality management will be reconsidered in section 5, the following section discusses model validation in more detail.

4. MODEL VALIDATION

Unfortunately, there is no standard theory or framework for model validation or verification (Kleijnen, 1995). In fact, a number of (partly) contradictory definitions exists on central concepts[1]. For instance, Oreskes et al. (1994) argue in their influential paper that the use of the terms validation and verification reflects "at best confirmation" and, as we will see below, both terms can bear different meanings or are used interchangeably. Some of the confusion in the literature reflects philosophical differences or perspectives among disciplines. As Eker et al. (2018) points out: "In decision sciences, validation usually implies establishing confidence in the model by judging its usefulness with respect to some purpose. In environmental modelling, validity is often used to indicate that model predictions are consistent with observational data, or that the model is an accurate representation of physical reality, or both." In any case, the need for clear definitions within a given context to establish a common understanding of the used terminology is obvious (Segerson, 2015). This is the aim of present section.

4.1. Verification and validation

Typically, validation represents a comprehensive evaluation of the model or the modeling process, while verification is commonly defined in a narrower, technical sense. By **verification** we determine that a simulation model performs as it was intended by the developer(s) (Kleijnen, 1995; Sargent, 2013). This includes documentation, program code debugging, and model testing to assess whether errors or inconsistencies exits within the model (Balci, 1998; Rand & Rust, 2011). In other words, verification ensures the correct implementation of a conceptual model[2] into a computer program similar to making sure that the arithmetic is correct within a mathematical model.
The ultimate objective of model validation is to establish or to increase model credibility (McCarl, 1984) by assessing the level of confidence that can be placed in the model and its results (Bert et al., 2014; Sargent, 2013). In this context, credibility is a sufficient degree of belief in the validity of a model to justify its use for research and decision making (Rykiel, 1996). For instance, showing that a simulation outcome has a certain level of accuracy (with respect to a given evaluation criteria) can induce this credibility (Beisbart, 2019). Accuracy can be measured objectively, which is commonly conducted within model testing (e.g., determining the deviation of the simulation result to a target level as discussed in section 4.2). In contrast, credibility is a subjective qualitative judgment, and cannot be quantified (Rykiel, 1996). In fact, what is "credible" as well as the specific procedure to judge that a model is validated depends on the context of the problem and the intended use of the model Barlas and Carpenter (1990). Several attempts have been made to structure and classify validation approaches and techniques.

[1] See also Breisbart (2019) for a recent discussion of the different definitions.

[2] For instance, the conceptual model can be a verbal description, mathematical formulation, governing relationships, or "natural laws" that approximate the real system (Schlesinger, 1979).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Among the first to provide a framework to review the credibility of simulation models were Schlesinger (1979), who state that model **validation** is the "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model". Before and since then clarification or modifications of this definition as well as alternative formulations are used in the literature e.g., (Anderson, 1974; Beisbart, 2019; Mitchell, 1997; Oreskes et al., 1994; Rand & Rust, 2011). The common denominator, however, is that models cannot be proved valid, but can only be judged to be so (Barlas & Carpenter, 1990). In this respect, Balci (1998) argues, the "adjective 'sufficient' must be used in front of terms such as model credibility, model validity, or model accuracy to indicate that the judgment is made with respect to the study objectives". Similarly, Kleijnen (1995) highlights that a model should be "good enough", because the result of validation is never a perfect model because this would be an exact copy of the real system itself. Rather the relation between the purpose of the analysis and the type of model to accomplish a given objective is central for the assessment of validity. Rykiel (1996) states that validation ensures that a "model is acceptable for its intended use, i.e., whether the model mimics the real world well enough for its stated purpose". Validation thus refers to a sequence of activities that determine the usefulness of a model (Eker et al., 2018; McCarl, 1984). Before such activities can be carried out, the purpose of the model, the performance criteria, as well as the model context need to be specified (Rykiel, 1996). In this respect, Swinton (2018) emphasizes that finally the "audience" (e.g., users, stakeholders or other experts) make the call whether credibility is assigned to a model or its outcomes. Thus, validity is conditioned on both purpose or topic *and* the addressees that will use the model or its results. Figure 1 illustrates this triad.

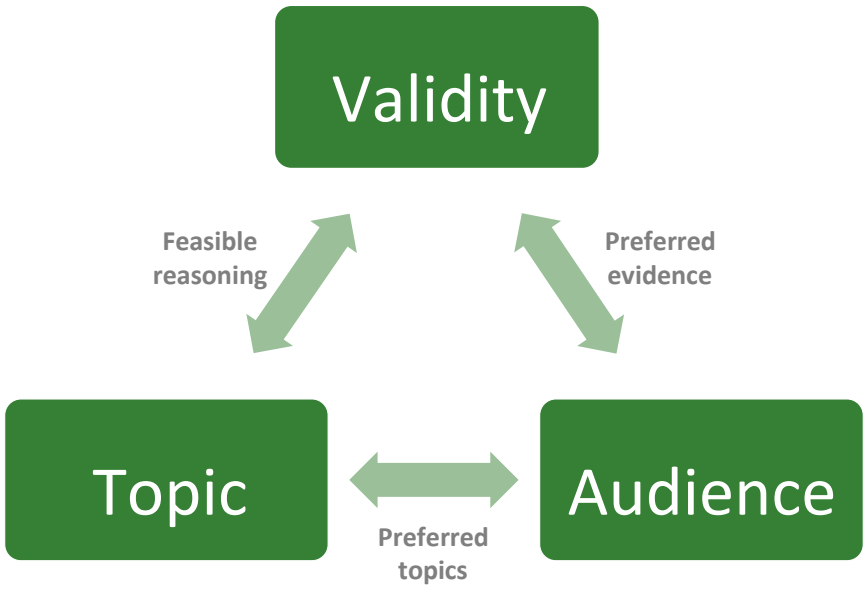

Figure 1: Validity triad (adopted from Swinton, 2018).

To bluntly summarize, verification means building the model right while validation denotes building the right model (Balci, 1998)[3] knowing that this is rarely an objective, binary judgment - not only by

[3] Note that Oreskes et al. (1994) famously argues to the contrary such that validation is to ensure a model does not contain known or detectable flaws and is internally consistent, while verification is the assertion or establishment of truth, which is almost never possible, because real world systems are never closed. However, verification and validation, as used in this report, is in line with the majority of works on simulation models (e.g., Sargent, 2013; Balci, 2004; Rykiel, 1996).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

the modeling team but also the audience - and conditioned on the context of the model's application. With validation taken a broader perspective, verification is part of the validation process of simulation models such that a validated model always needs to be verified but verification does not necessarily yield a valid (or validated) model. Throughout the report we speak of a valid model if it has been validated and the validation did meet some quality criteria (Rykiel, 1996). A valid model in our context does not represent the universally true model rather the model structure or its outputs are sufficiently close to the workings or observed states of the real system (Mitchell, 1997). For the sake of brevity, we also use the term *validation* to refer to the model evaluation process instead, as it is often used in the literature, to refer to *verification and validation* (Oberkampf & Roy, 2010) or *verification, validation, and testing* (Balci, 1998) because both testing and verification is always part of the validation process of simulation models.

4.2. Model testing and calibration

Both, model testing and calibration are typically considered as part of the validation process. However, model testing is less a separate category. Rather, specific activities within verification and/or validation can be considered as model testing. Based on Macal and North (2005), Rand and Rust (2011) differentiate programmatic testing and test cases (Table 1). The first pertains to verification to ensure that the implemented model works as intended by the modeling team and it involves unit testing, code walkthroughs, debugging walkthroughs, and formal testing. Test cases use artificially generated data to check for correct model functionality, e.g., by specific scenario tests, corner cases, sampled cases, and relative value testing.

Table 1: Model testing

Test	Description
Unit testing	Every unit or section of code has a test written for it
Code walkthroughs	Researcher/user and programmer together examine the code step by step
Debugging walkthroughs	Running the code examined at each or defined steps to check for the correct results at execution time
Formal testing	Correctness of the code is inferred by logic, which typically requires limited complexity of the model
Specific scenario testing	Given specific input, there is conceptual or expert knowledge about the results, which is compared to actual model behavior under this scenario
Corner cases	Extreme values of the inputs are used to test the model behavior

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Sampled cases	Based on a subset of input values the model results should be within a known range of possible outcomes
Relative value testing	If the relationship between an input and an output variable is known (by tendency), changing this input should cause the predicted change in output

Source: Adopted from Rand and Rust (2011).

While some form of model testing is always performed (and required) within the modeling process, calibration has a direct link to empirically data (Moss, 2008). Hence, there might be instances of applications where calibration is not necessarily part of the validation process. For instance, deterministic, theory-based (numerical) simulations may not need to be calibrated or calibration takes only place in defining a potential range, averages or reasonable values of some exogenous variables. For most purposes, however, empirical data are employed in two ways Werker and Brenner (2004): (a) to set up the simulation model, i.e., to parameterize it and/or (b) to test the simulation model based on statistical approaches. In this respect, calibration is the estimation and adjustment of model parameters and constants to improve the agreement between model output and a data set (Rykiel, 1996). For instance, van Vliet et al. (2016) review calibration approaches with respect to land-change models and find that statistical analyses and automated procedures are the two most common calibration approaches, while expert knowledge, manual calibration, and transfer of parameters from other applications are less frequently used.
There are a number of techniques that can be applied within the validation process, but these techniques are not mutually exclusive for the one or the other purpose. In fact, a clear delineation between verification and validation, testing or calibration techniques themselves is not (and often cannot be) made, but the context in which these techniques are applied matters, i.e., whether to evaluate the sound transformation of the conceptual model into the simulation model or whether the simulation is a sufficiently good representation of the real system Bharathy and Silverman (2013). Before we present these techniques in more detail, that context within the simulation model life cycle, i.e, the occasions for model validation will be discussed.

4.3. Modeling context

Models are developed for specific purpose such that some can generate recommendations for action while others support decision-making (Eker et al., 2018). In general, models can be used for three purposes in particular (McCarl, 1984): structural exploration, (to discover the determinants influencing economic behavior), prediction (to forecast the consequences of decisions), and prescription (to identify the best or a desirable action for a given decision problem). Barlas and Carpenter (1990) argue that non-causal (statistical/correlational) models should only be used for prediction while causal (theory-like) models can be used for prediction and explanation. In this respect, the potential to use simulation models for “what-if” studies (e.g., in terms of sensitivity analysis) is valuable and even more so if it challenges prior assumptions used to represent a system (Oreskes et al., 1994).
According to Bankes (1993) there are two very different modeling approaches: consolidative and exploratory modeling. In the first case, one consolidates known facts into a model to use it as a surrogate for the real world system. Marks (2013) calls such models *descriptive models*. The advantage of this approach is that it can represent system behavior sufficiently closely so that it can be used to predict consequences of decision or policies. The disadvantage is that comprehensive and adequate knowledge of the system is required. If these information or data are not available, one has to resort

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

to exploratory modeling, which uses a series of computational experiments to explore the implications of varying assumptions and hypothesis, i.e., these models address “what-if” questions and they are used as heuristics to guide decision making (Eker et al., 2018; Marks, 2013). Exploratory modeling can be used for three types of applications (Bankes, 1993): (a) data-driven exploration, (b) question-driven exploration, and (c) model-driven exploration.

(a) data-driven exploration starts with a data set and attempts to derive insight from it by searching over an ensemble of models to find those that are consistent with the available data

(b) question-driven exploration searches over an ensemble of models believed to be plausible to answer a question of interest

(c) model-driven exploration involves neither a fixed data set nor a particular question or policy choice, but rather is a theoretical investigation into the properties of a class of models

For any modeling approach, there is a variety of different models that can be used. In this respect, Kleijnen (1995) refers to Karplus (1983) who differentiates a whole spectrum of mathematical models, ranging from black box (non-causal) models (e.g., regression analysis in the social sciences) through gray box models (e.g., linear programming in ecology) to white box (causal) models (e.g., in physics and astronomy). Whatever type of model is used or whatever approach is taken, model validation depends on the intended use of that model. A valid (or validated) model for one purpose may not be valid for 10 another, e.g., a simulation model for a descriptive "what is” question may not be useful in a "what should be” context (Burton, 2003).

4.4. Occasions for validation: The modeling process

It is long been recognized that there needs to be a stage of validation within the modeling process (Mitchell, 1997; Pachepsky et al., 1996), but, in fact, model validation and eventually evaluation is a continuous process and not the end to a project or model’s development. On the one hand, validation needs to be conducted through the whole modeling process, which is illustrated in Figure 2. On the other hand, validation is an iterative but not a time-linear processes for two reasons: (a) validation criteria may evolve along with the model (Rykiel, 1996) and (b) new model applications may require repeated or adjusted model validation (McCarl, 1984; Reidsma et al., 2018).
Basically any model development follows a similar process (Anderson, 1974; Balci, 1998; McCarl, 1984) with some deviations regarding the focus and context of the particular study. A general and simplified modeling process as shown in Figure 2 starts with the identification and formulation of the problem, including e.g., the definition of the research question(s) and modeling gaps as well as the future use of the model (1). Based on prior work, literature or logic the relevant system that should be investigates is characterized and a simplified representation of the real system is developed and synthesized into a conceptual model (2). This step of system analysis includes the identification of relevant features and interrelationships in the system and its environment with explicit consideration of stochastic elements (Anderson, 1974). The selection of the modeling approach (3) should be guided by these characteristics of the real world system and the intended use and objective of the model but might also be affected by (potential) data availability (4) or lack thereof. The model construction (5) represents the transformation of the conceptual model into the operational model, e.g., an empirical, theoretical, or simulation model. Calibration (6) refers to formal or informal statistical fitting of model parameters compared to the real world. While informal approaches include trial-and-error, i.e., best guesses by the modeler or expert knowledge, i.e., reasonable assumptions regarding values and functional relationships, formal approaches are based on statistical techniques and (often comprehensive) empirical data (McCarl, 1984; Troost, 2015). The outcome of the calibration stage is a base model (7)

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

that can represent the status quo or a benchmark to be used for scenario analysis. The identification of relevant scenarios belongs to stage of formulating the experiments (8) where these computational experiments correspond to "what-if" scenarios (Eker et al., 2018). In the case of complex simulation models with a large number of variables and parameters, concepts as design of experiment might be employed at this stage (Rand & Rust, 2011). The execution of different experiments in the final stage (9) of the presented modeling process generates a dataset that can be supplied to further analysis.
It is quite clear that the outcome of stage (9) should be used to answer the initially stated (research) question or provide a solution to the problem statement (1). Hence, there is a natural feedback within the linearly presented modeling process of Figure 2, but not only in this case. Often validation is considered a necessary activity after the model is developed but validation is an iterative process and should start from the beginning and run through all stages of the modeling process (Balci, 1998; Bert et al., 2014; Bharathy & Silverman, 2013). In this respect, every stage of the modeling process provides the opportunity or need to revisit prior stages namely if the model fails a validation test (Rykiel, 1996). Examples include that the model outcome or behavior (observed in stage 9) does not resemble the real world system or that theoretical predictions (e.g., from stage 2) are not supported by empirical observations.

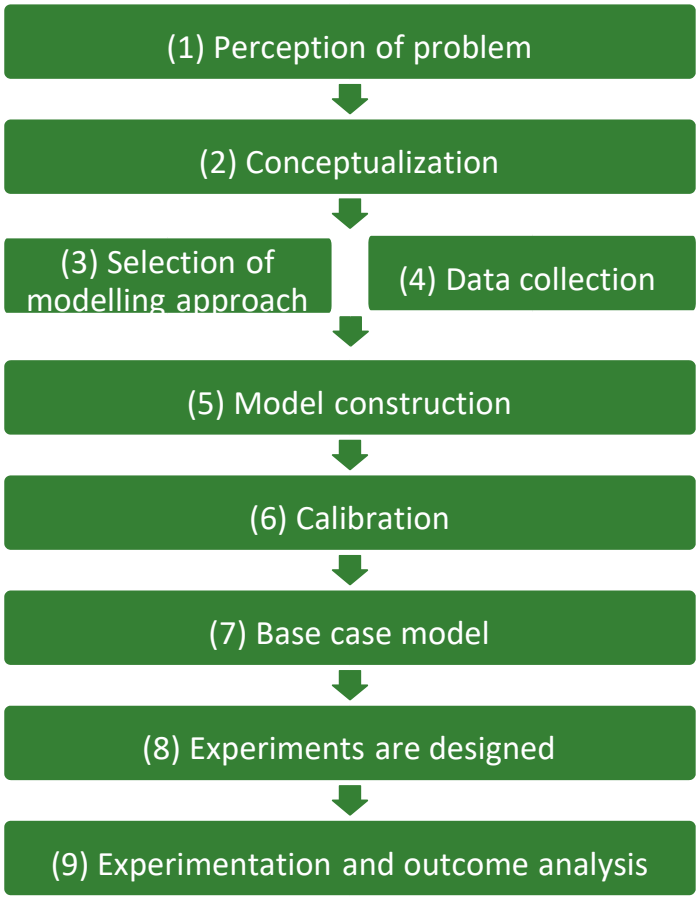

Figure 2: A general representation of the modeling process.

4.5. Validation approaches and concepts

Balci (1998) differentiates validation approaches by who performs the tests and at which level of model development these tests are conducted:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

1. Private Testing. Represents informal, individual evaluation by the modeler, mostly without documentation.

2. Submodel (Module) Testing. Each submodel (or module) independently passes through tests that are designed, executed, and documented by the modelling team.

3. Integration Testing. These tests, conducted by the modelling team, aim to verify that no inconsistencies in interfaces and communications exist when submodels are combined.

4. Model (Product) Testing. The modeling team seeks to access the model's overall validity. This stage usually involves empirical validation (see section 4.5.1).

5. Acceptance Testing. Stakeholders or third parties independently design, execute, and document model evaluation. The aim is to establish the simulation model's credibility so that the results can be accepted and used by the stakeholders.

The details of each level and what levels are required depends on the context and objectives of the model. For instance, integration testing is certainly required if the model consists of (a large number of) different modules while simpler applications do not necessarily feature this level. Also, acceptance testing might show different forms, e.g., if a simulation study is subject to peer-review, eventually published, and increasingly used (cited) in the academic literature, which lends an increasing degree of credibility to simulation results. Figure 3 illustrates the five levels of model testing, spanning from private testing (by the modeler itself) to acceptance testing (by stakeholders), with increasing complexity of planning and management of model evaluation from the inner to the outer level.

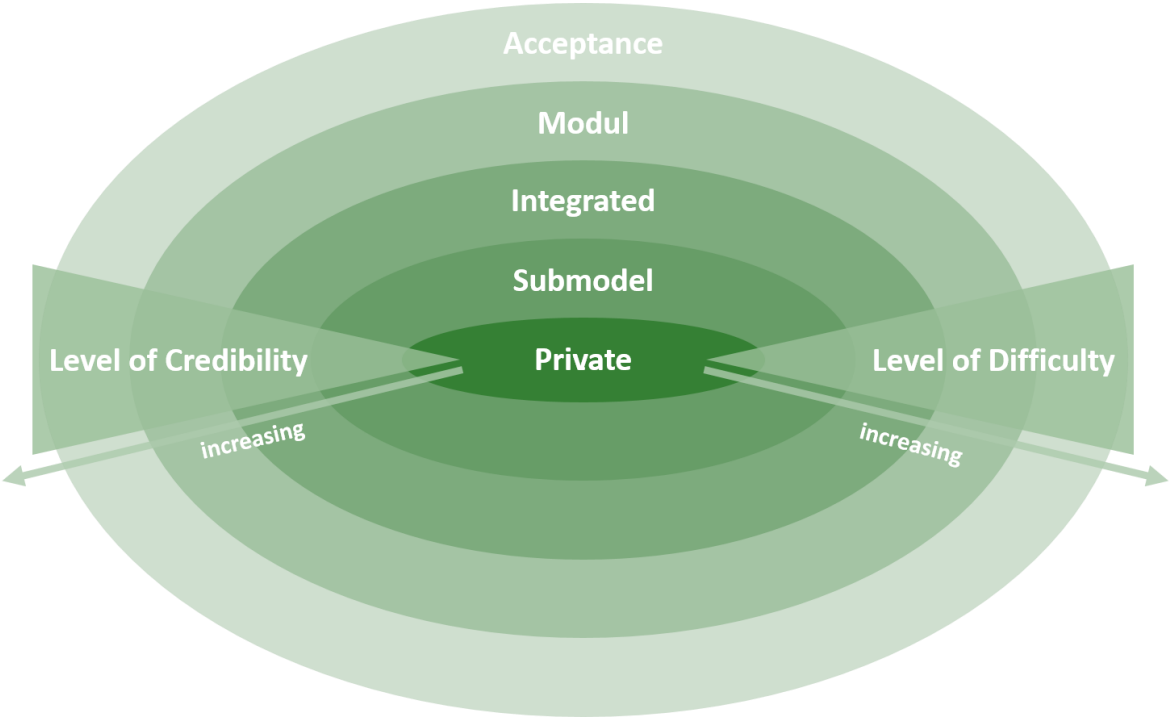

Figure 3: Levels of model testing and trade-off between credibility and Difficulty.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Another common characterization of the validation process reflects the phases of model development. McCarl (1984) differentiates technical and operational validation. Technical validation concerns to test of a models assumptions and data, its formal logic (e.g., in terms of equations), and its predictive-prescriptive ability. Operational validation is directed on the intended application of the model and includes (a) identifying the acceptable domain of use, (b) tests of mechanisms that adapt the model to a particular use case, (c) tests of model updating procedures, and (d) repeated technical validation for different applications of the model. Accordingly, the focus of technical validation are the early stages of the modeling process (e.g., stages 1 through 5 in Figure 3) while operational validation focuses mainly on the latter stages (e.g., stage 6 to 9). Among others, Sargent (2013) separates conceptual validation, verification, and also operational validation with conceptualization, model construction, and experimentation at the core of the respective validation stages. Figure 4 illustrates this validation process. Sargent (2013) defines conceptual model validation as the process to determine that the theories and assumptions underlying the conceptual model are correct and that the representation of the problem is reasonable for the intended purpose of the model. As defined in Section 4.1, verification ensures the correct implementation of a conceptual model into the operational (e.g., simulation) model Rand and Rust (2011). As before, operational validation determines that the models output behavior is sufficiently accurate with respect to the models intended use Sargent (2013).

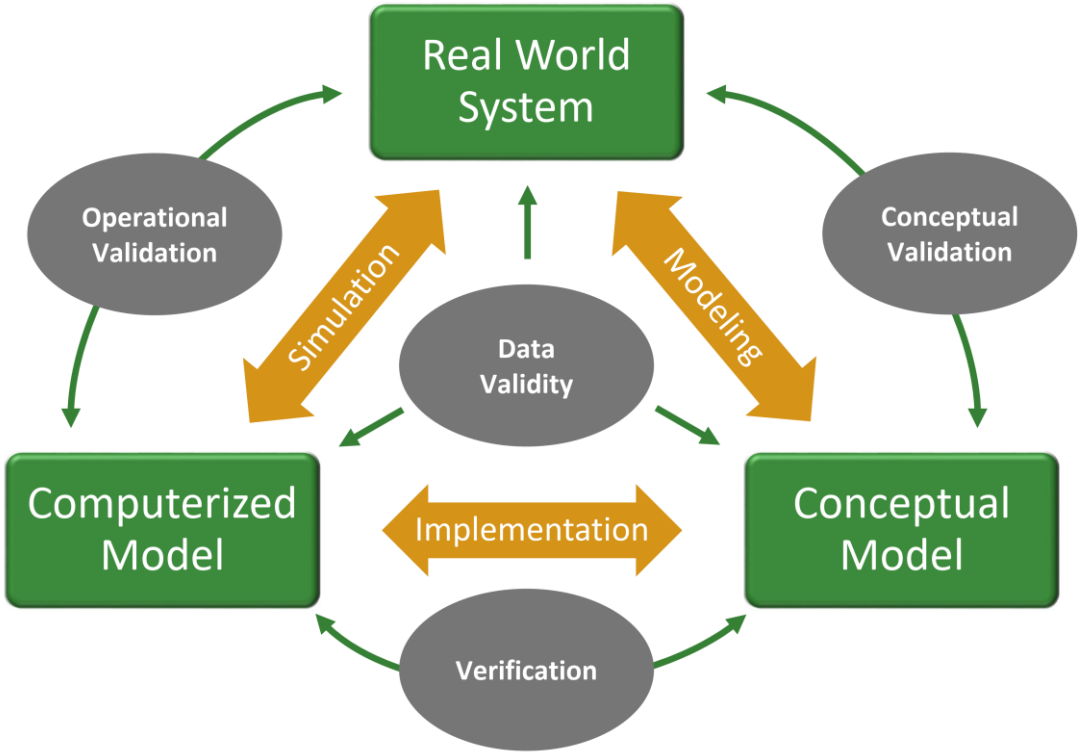

Figure 4: Simplified version of the model development process (Sargent, 2013).

As Sargent (2013) observes, data validity is often not considered as part of model validation even though there is a strong emphasis on empirical data in validation (Eker et al., 2018). In fact, data that are collected from the real world system can inform the development of the conceptual model as well as the computer model (e.g., in terms of calibration) and can be used for model validation and experimentation (with the validated model) (Robinson, 1999). Figure 4 illustrates that data are crucial at most if not all stages of model validation. Data validity ascertains that data necessary for model building, model evaluation and testing, and conducting the model experiments to solve the problem

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

are adequate and correct (Sargent, 2013). This becomes particularly important because operational validity is often based on empirical validation, i.e., activities that involve comparisons of model outputs against real world data (Bert et al., 2014). The (simulation) literature places a priority on empirical validation as the compliance of model results and real world observations is an often used criteria for model credibility (Bharathy & Silverman, 2013; Eker et al., 2018; Moss, 2008; Topping et al., 2012). Because of the importance we will address empirical validation in more detail before aspects of integrated model evaluations are discussed.

4.5.1. Empirical validation

An empirically validated model is grounded on qualitative and quantitative data collected from the system of interest (Garcia et al., 2007). A single data gap or inconsistency can invalidate the results of any model and destroy the model's credibility. Therefore, data validation is necessarily a part of the model validation process (Macal & North, 2005). Power (1993) distinguishes three types of empirical validity:

- **Replicative validity:** The model matches data already acquired from the real system and used in the formulation and estimation phases of model design and construction.
- **Predictive validity:** The model matches data before the data are acquired from the real system.
- **Structural validity:** The model reproduces real system behavior such that it reflects operating characteristics of the real system.

Especially for replicative and predictive validity it is often necessary to model the real world system closely to obtain good correspondence between model and reality (Topping et al., 2012). Depending on what type of empirical validity is the objective, different methods need to be employed. For instance, Power (1993) stresses that standard statistical techniques useful for the replicative validation of models include tests of means and variances, analysis of variance, goodness-of-fit, regression and correlation analysis and confidence interval construction.
Power (1993) also discusses options for predictive validity including data splitting (cross-validation) and the evaluation and comparison of model predictions. While Diebold and Mariano (2002) discuss tests for evaluating predictive accuracy of competing model forecasts, Mitchell (1997) emphasizes that regressions are not appropriate for empirical validation with respect to model predictions.
Lamperti (2018) provides a recent review of the literature on empirical validation and argues that a prerequisite for policy analysis with respect to macro-oriented models is their ability to replicate key empirical stylized facts. For instance, Giannone et al. (2006) and Canova and Sala (2009) present details regarding estimation and validation 15 of dynamic stochastic general equilibrium models where the vector autoregressive model is the basic econometric tool for empirical validation. On the other side of the spectrum, i.e., micro-(or firm-)level models, Windrum et al. (2007) review empirical validation of agent-based models and discuss three alternative approaches: (a) the history-friendly approach, (b) Werker-Brenner calibration, and (c) indirect calibration. While all are based or informed by empirical knowledge, the resemblance between simulation history and real world history decreases from (a) to (c). While the first aims for the highest correspondence between simulated and observed history of the system, the latter focuses on empirical evidence on stylized facts to restrict the parameter space (Moss, 2008).
All of the above concern *validation by result* (McCarl, 1984) as model output actually is or can be compared to the real world system. In contrast, *validation by assumption* is another important type of validation, where prior theoretical or expert knowledge is used, especially regarding conceptual model

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

building or model calibration, e.g., if appropriate empirical data are not available. Ideally, the real world system should be adequately represented by the model on both a micro and a macro-level (Garcia et al., 2007). In this respect, Utomo et al. (2018) references literature distinguishing black box and white box validation. The authors define black box validation as the evaluation of whether the model outputs either reflect the empirical observations for the same set of inputs or are consistent with the result from a mathematical model. White box validation evaluates whether the decision rules of agents represent the decision rules of actors in the real world and whether the structure of the model (such as the network between agents) represents reality (Utomo et al., 2018).

4.5.2. Integrated evaluation

It is generally accepted that the modeling process includes a stage of validation (Mitchell, 1997), e.g., contrasting model output with empirical data. An integrated approach to model evaluation emphasizes the fact the validation needs to start from the beginning of the modeling process, i.e., the conceptual design stage (Bert et al., 2014) and needs to run through the entire life cycle of the (simulation) model (Bharathy & Silverman, 2013; McCarl, 1984; Rykiel, 1996). The more validation techniques (see section 4.7) are applied, the higher the credibility of models (Eker et al., 2018) and validation experiments, in contrast to traditional experimentation, particularly aim to determine the usefulness of a model, e.g., in terms of its predictive capability (Oberkampf & Roy, 2010). In this respect, integrated model evaluation is a process of evidence accumulation to support the case for using simulation models in complex, risky decision making situations (Trucano et al., 2006).
While Schwanitz (2013) points out that there is little understanding let alone consensus on how complex models can be evaluated and accepted standards are missing, Jakeman et al. (2006) argue there are specific points that need to be considered to obtain credible results. The authors present ten steps that, independent of the modeling problem, should be followed and the present section is based on this discussion as it summarizes and synthesizes considerations from different fields and approaches. The ten steps laid out by Jakeman et al. (2006) are illustrated in Figure 5. The figure also highlights that validation itself can be decomposed in a number of steps and can be presented sequentially, but model evaluation is iterative in nature (Oberkampf & Roy, 2010).
Jakeman et al. (2006) provide a detailed description of these steps. The important implication for the MIND STEP project is that each of these steps require decisions between alternative options in the modeling process and respective justification for choosing one alternative over another. Transparent documentation of these decisions are apt to increase model credibility. For instance, Schwanitz (2013) develop an evaluation framework for models of global climate change based on systematic and transparent step-by-step demonstration of a models usefulness testing and the plausibility of its behavior. The author emphasizes that setting up an evaluation framework, evaluation of the conceptual model, code verification and documentation, model evaluation, uncertainty and sensitivity analysis, documentation of the evaluation process, and communication with stakeholders are important steps. Bert et al. (2014) provide an evaluation framework for land use modeling based on the validation of model processes and components (including the comparison with alternative models and the involvement of stakeholders) as well as empirical validation. Reidsma et al. (2018) further argue that model structure and design as well as underlying assumptions and model constraints have considerable effect on model results and there is a clear need for the comparisons of different models or approaches. As emphasized by Marks (2013), model evaluation among several contending models can point the researcher and stakeholders to the “best” model or modeling framework.
In their recent overview, Fagiolo et al. (2019) critically review existing validation techniques for agent-based models. The authors develop a conceptual framework along three dimensions: (i) comparison between artificial and real-world data; (ii) calibration and estimation of model parameters; and (iii) parameter space exploration. Despite their focus on agent-based models, some of the conclusions apply more broadly: (a) Pros and Cons of different validation are not always available and an "if-then

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

map" for selecting the right tool for specific situations is still not available. (b) The development of better empirical-validation techniques is a never-ending process, which must naturally co-evolve together with the developments of new models, new statistical techniques and with the increase in computational power.

4.6. On validation criteria

One precondition for successful model evaluation is that the performance criteria are specified (Anderson, 1974; Rykiel, 1996), i.e., quality indicators that lead to the judgment of the model being valid (or sufficiently valid). We need to recognize that model performance may be assessed against many criteria (Jakeman et al., 2006), but it is important that these criteria are defined and justified with reference to the purpose of the model (Mitchell, 1997).

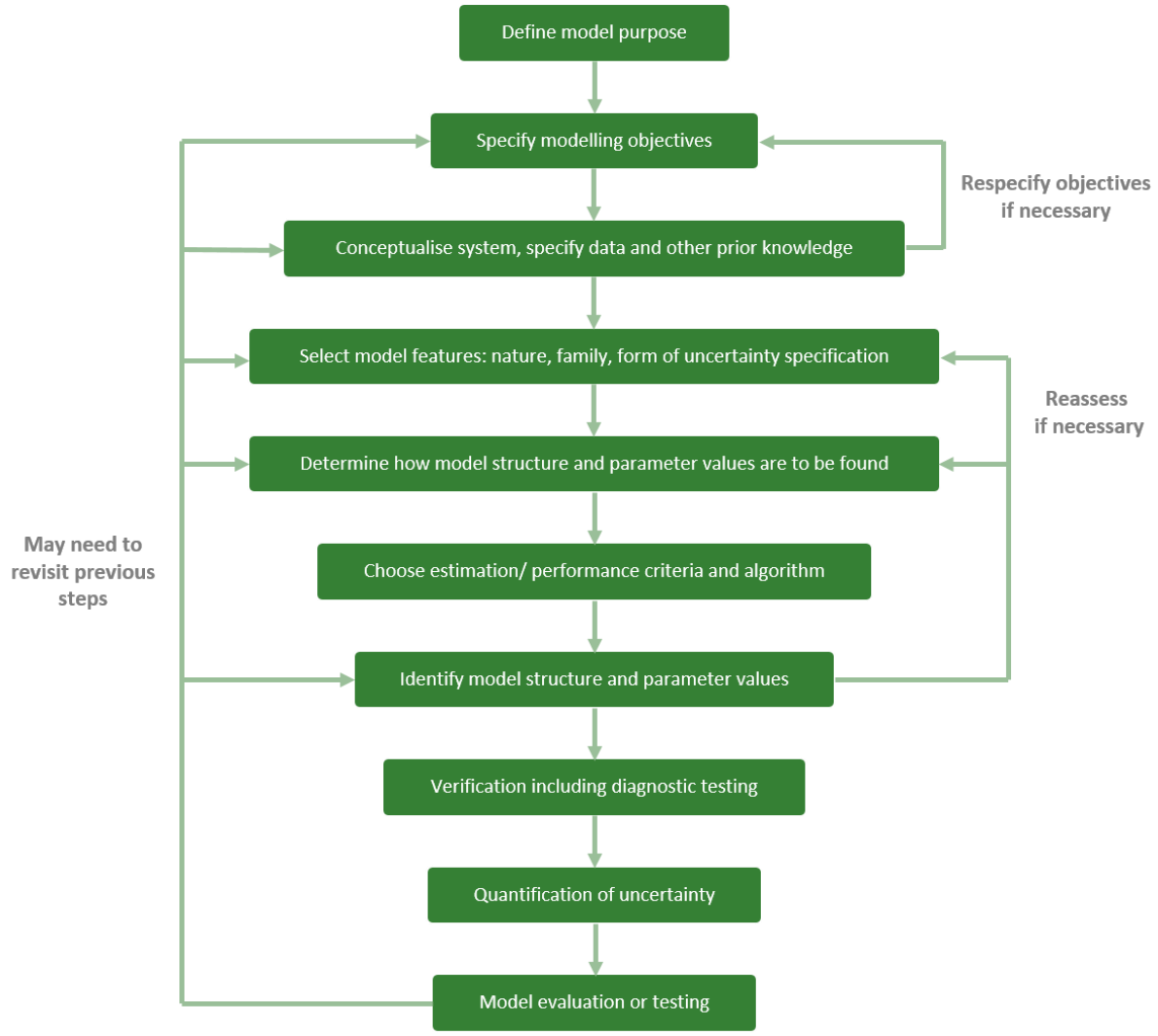

Figure 5: Iterative relationship between model building and evaluation steps according to Jakeman et al. (2006).

For econometric models, validity is usually examined by a set of statistical tests that typically include objective criteria (Anselin, 1988), e.g., goodness of fit. This objectiveness or typically agreed standards account for demanded "'transparency' of the techniques for obtaining 'credible' causal effects for 'evidence-based policy evaluation'" (Panhans & Singleton, 2017). In the case of simulation models,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

however, a common and unique framework for model validation is missing (Kaye-Blake et al., 2014). Instead, the evaluation process is strongly dependent on a couple of factors such as purpose of the model, data (availability), or conceptual framework. If inputs and outputs of the real system can be measured, techniques for comparing simulated and real data can also rely on statistical tests. In this respect, the ability to replicate the history of a system is a commonly used validity criterion concerning its predictive power (Eker et al., 2018). Among others, Power (1993) discusses criteria for predictive model validation including tests of means and variances, goodness-of-fit testing, regression and correlation analysis.

Other evaluation criteria can be derived by comparing (the outcome of) different models. On the one hand, the equivalency between two models provides a validation for each model (Burton, 2003) or, as Macal and North (2005) citing Axelrod (1997) put it: "replication of results from multiple models is one of the hallmarks of cumulative science". On the other hand, one is interested in whether a model performs better compared to alternatives (Marks, 2013; McCarl, 1984). In this respect, there are two different levels of equivalency between models (Axelrod, 1997; Burton, 2003): distributional and relational. Numerical and distribution equivalency requires the same (numerical) results; relational equivalency requires equivalent internal relation results. Depending on the purpose of the model, priority might be placed on the one or the other aspect of this comparison. For instance, Axtell et al. (1996) presents concepts and methods for the alignment of computational models that deal with the same phenomena. One challenge, however, is that even within one stream of modeling, there are often difficulties comparing the relative explanatory power of models from different methodological lineages (Barde, 2017). Hence, several authors propose indicators that can be applied to any model able to simulate or predict time series data (Barde, 2017; Lamperti, 2018; Marks, 2013).

If models and submodels include unobservable inputs and outputs, they can not be subjected to common statistical tests and instead sensitivity analysis is required, i.e., what-if analysis in terms of a systematic investigation of changes in model inputs or model structure on model outputs (Kleijnen, 1995). Because each computational experiment corresponds to an uncertain what-if scenario, the model may generate a large ensemble of exploratory scenarios and representation accuracy may be less important in validation than the qualitative behavior of the model. In this respect, there is an important difference to econometric model "validation" because the nature of quality criteria are openly subjective, e.g., whether model behavior is inline with the judgments of experts (Jakeman et al., 2006). This is especially true for economic models (or modeling human behavior in general) while most advanced methods of validation were developed in engineering where model behavior is determined by fundamental physical laws (Louie & Carley, 2008). In this respect, Burton and Obel (1995) argue that "realism in computational modeling is clearly germane; but, balance is a more demanding criterion. Without some degree of realism, computation modeling becomes a logical and/or numerical exercise. At the other extreme, total realism may create an imbalance with all the experimental and analytical issues that any real world field experiment has." In fact, perfect confirmation in terms of a model's consistency with all available knowledge shouldn't be the criterion for such model evaluation rather the fitness for purpose and transparency of the process by which the model is produced needs to be accounted for (Jakeman et al., 2006). In this respect, the common argument is that the entire life cycle of model development, use and improvement needs to be considered to assess model validity (Bharathy & Silverman, 2013; McCarl, 1984; Rykiel, 1996).

In their review, Eker et al. (2018) identify important validation criteria based on a survey among practitioners. The authors show that important criteria for a model's validity or credibility are

1. how useful it is for a given purpose (79% of respondents agree or strongly agree),
2. how well it represents reality (67%),
3. if uncertainties and critical assumptions are communicated well (65%),
4. if it can replicate historical data (62%).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

While these criteria can be defined and accounted for during model development, one of the most important criteria indicating the validity of a model is its acceptance by the model user (Swinton, 2018), which Schlesinger (1979) denote as "model certification" and which can only be accessed *ex post*. One example, relevant in the MIND STEP project can be the introduction of a model(ling framework) into the "Modelling Inventory and Knowledge Management System of the European Commission", the MIDAS database (see section 5). Another important criteria is the use of models and their results in academic publications. While Finger et al. (2022) are critical about a single indicator with respect to evaluating the performance of agricultural economics journals, citations are still one of the major quality criteria for any type of research and so indicate credibility of simulation studies as well. Typically, this is the only external validation of simulation models in agricultural economics. As Balci (1998) argues (see section 4.8) the developer with the most knowledge of the model may be the least independent and external validation by an independent party can improve the credibility of a model and the very fact that this type of validation was carried out can serve as validity criteria. Of course, such validation is the most time and cost consuming but, e.g., Sargent (2013) argue that if independent validation is conducted on a completed simulation model, it is usually best to only evaluate the verification and validation that has already been performed. In any case, the trade-off between the cost of the validation process and individual techniques employed within it and the benefits of the validity information need to be considered and will determine the validation strategy (McCarl, 1984; Sargent, 2013).

4.7. Validation techniques

In the following, some selected validation techniques are briefly described while additional options are presented by Balci (1998), who provides a comprehensive discussion of over 70 techniques. Appendix 1 contains an excerpt of Balci (1998) taxonomy. The following list is composed based on Power (1993), Rykiel (1996), Baldos and Hertel (2013) and Sargent (2013). The techniques are presented in alphabetical order.

Animation: The model's dynamic behavior is displayed graphically. For instance, a grid of plots may represent the fields in a region and colors illustrate its current use while a change of the color represents land use change during the simulation.

Comparison to other models: The output of the (simulation) model is compared to the results of other (validated) models.

Cross validation/data splitting: The available data is split into two data sets: an estimation and a prediction data set. The estimation data set is used to estimate model parameters and to assess the replicative validity of the resulting model. The prediction data set is used exclusively for predictive validation.

Data relationship correctness: Requires data to have the proper values regarding relationships that occur within a type of data and between different types of data. For example, are the input-output levels for a given production activity correctly presented.

Degenerate tests: The degeneracy of the model's behavior is tested by appropriate selection of values of the input and internal parameters. For instance, does the number of farms in a region decrease if the farm exit rate is larger than the farm entry rate.

Event validity: A comparison between the model and system is made of the occurrence, timing and magnitude of simulated and actual events. For instance, farm exit might be triggered by profitability or generation change.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Extreme condition test: The model structure and outputs should be plausible for any extreme and unlikely combination of levels of factors in the system.

Face validity: Expert assessment if the model and its behavior are reasonable, i.e., whether the model logic and input-output relationships appear reasonable on the face of it given the model's purpose.

Historical data validation: Evaluate the model performance (results and dynamic behavior) under given specifications against the historical system behavior.

Internal validity: Several replications (runs) of a stochastic model are made to determine the amount of (internal) stochastic variability in the model.

Multistage validation: Validation methods are applied to critical stages in the model building process: (1) developing the model on theory, observations, and general knowledge; (2) validating the model's assumptions where possible by empirically testing them; and (3) comparing (testing) the input-output relationships of the model to the real system.

Predictive validation: The model is used to forecast the system behavior and comparisons are made to determine if the system's behavior and the model's predictions are the same. The system data may come from data sets not used in model development or from future observations of the system. The strongest case is when the model output is generated before the data are collected.

Sensitivity analysis: The (systematic) change of input values and internal parameters of a model to determine the effect upon the model's behavior or output. Parameters that are sensitive, that is, cause significant changes in the model's behavior or output (qualitatively and/or quantitatively), should be made sufficiently accurate prior to using the model.

Structured walkthrough: The entity under review is formally presented usually by the developer to a peer group to determine the entity's correctness. An example is a formal review of computer code by the code developer explaining the code line by line to a set of peers to determine the code's correctness.

Trace: The behavior of specific variables is traced through the model and through simulations to determine if the behavior is correct and if necessary accuracy is obtained.

Turing tests: Knowledgeable individuals are asked if they can discriminate between system and model outputs.

Statistical validation: A variety of tests performed during model calibration and operation. Three cases are most common: (1) the model produces output that has the same statistical properties as the observations obtained from the real system; (2) the error associated with critical output variables falls within specified or acceptable limits; (3) several models are evaluated statistically to determine which test fits the available data.

Visualization techniques: The dynamical behaviors of performance indicators are visually displayed as the (simulation) model runs through time to ensure that the performance measures and the model are behaving correctly. Such visualization can form the basis for comparisons between system and model and a subjective statement concerning the visual goodness of fit.

4.8. Principles of model evaluation

Despite the wide variety of possible validation techniques, occasions and the context of the simulation study, any activity intended to demonstrate validity and to increase credibility of model results should follow a number of principles for a successful simulation study. The following list is adopted from Balci (1998).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Principle 1: Model evaluation, i.e., verification, validation, and testing has to be conduct at every stage of a study's life cycle.

To ensure the validity, accuracy, and reliability, model evaluation is an ongoing activity. In this way, the quality shortcomings can be recognized and be corrected as the project progress through the simulation phases.

Principle 2: The outcome of model evaluation is rarely that a model is absolutely correct or absolutely incorrect.

Because models are abstraction from reality, perfect representation of the real system cannot be expected and a dualistic view that a model is either correct or incorrect is not adequate. Instead, it is reasonable to consider the outcome of model evaluation as a degree of credibility on a scale between 0 to 100, where 0 represents incorrect, and 100 represents correct. As model credibility increases, the model development cost will increase. Likewise, the model utility will increase but at a slower rate.

Principle 3: A simulation model is built and its credibility is judged with respect to the study objectives.

For a simulation study to be successful, the objectives and specifications of the model must be precise. The level of representation required from the model will depend on the study's objectives and a higher or lower representation accuracy will be required depending on the importance of the decisions based on the simulation results.

Principle 4: Model evaluation requires independence to prevent developer's bias.

The model developer with the most knowledge of the model may be the least independent when it comes to testing because they may fear that negative results impede the credibility of the model. Model testing is most credible in itself if conducted by an independent, unbiased party. Two alternatives can achieve independence: (a) There is an independent team within the organization to design, conduct, and document model testing or (b) a third party is responsible for this task.

Principle 5: Model evaluation is difficult and requires creativity and insight.

Adequate model evaluation requires understanding the problem, command of the simulation model, experience with the modeling methodology, ability to identify suitable test cases, and familiarity with the evaluation framework. While developers are best qualified to demonstrate creativity and knowledge of a model's internals since they are intimately familiar with them, they are not independent. Independent testing makes the evaluation process more complicated, which requires good planning and management.

Principle 6: Credibility can be claimed only for the conditions for which the (simulation) model is tested.

Because the initialization of a simulation model impacts the precision of the input-output transformation, a transformation that works for one set of input conditions might produce nonsensical results under a different set of input conditions.

Principle 7: Complete model evaluation is not possible.

An exhaustive (complete) model evaluation may require testing it under every possible input condition. Instead, model testing may focus on increasing confidence in the model's validity as determined by the study objectives. To estimate what percentage of the valid input domain is covered by the test data is important because credibility of a model increases with an increase in this coverage.

Principle 8: Model evaluation must be planned and documented.

Testing is not a phase or step in the model development life cycle, it is a continuous activity throughout the entire life cycle. For successful testing, careful planning and transparent documentation is

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

required. For instance, a test plan should describe what is selected for testing, data and code, test specifications, standards and conventions, test tools, and the expected and/or obtained results. Beyond the development team, project management and stakeholders should be involved in these tasks.

Principle 9: Type 1, 2, and 3 errors must be prevented.

Type 1 error occurs when a sufficiently credible result is rejected. Type 2 errors represent the opposite, i.e., invalid simulation result are accepted. Type 3 errors occur when a simulation fails to solve the actual problem, i.e., a solution is accepted but the problem formulation does not completely (or adequately) contain the actual problem.

Principle 10: Errors should be detected as early as possible in the modeling process.

Because detecting and correcting errors in later stages of the model life cycle (e.g., at the implementation stage) can be time-consuming, complex, and expensive, the main goal is to identify and resolve problems as soon as possible.

Principle 11: Multiple response problem must be recognized and resolved properly.

Simulation models with several output variables (responses) cannot be adequately validated with a univariate statistical approach. Instead, a multivariate statistical procedure is required to consider correlations among output variables by comparing them to system output (observations).

Principle 12: Successfully testing each submodel (module) does not imply overall model credibility.

Even if each submodel is credible (individually), the whole, aggregated model may not be (sufficiently) credible. Submodel errors can accumulate for the aggregated model. Therefore, the whole model must be tested despite sufficient credibility at the submodel-level.

Principle 13: Double validation problem must be recognized and resolved properly.

If data can be collected on both system input and output, model validation can be conducted by comparing model and system outputs obtained by running the model with the ”same” input data that drives the system. Determination of the “same” is yet another validation within model validation: the double validation problem. The “same” is determined by validating the input data models before validating simulation results.

Principle 14: Simulation model validity does not guarantee the credibility and acceptability of simulation results.

Model validity is a necessary but not sufficient condition for the credibility and acceptability of simulation results. There is a difference between the model credibility and the credibility of simulation results. The first is judged with respect to the system (requirements) definition and the study objectives, whereas the credibility of simulation results is judged with respect to the actual problem definition and involves the assessment of system definition and identification of study objectives. Therefore, model credibility assessment is a subset of credibility assessment of simulation results.

Principle 15: Formulated problem accuracy greatly affects the acceptability and credibility of simulation results.

The correct formulation of a problem can be even more crucial than its solution. The goal of a simulation study should not be to provide a solution, but to provide results that are credible, accepted, and eventually support decision making by stakeholders.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

5. VALIDATION OF THE MIND STEP TOOLBOX

The common framework and indicator system for model validation within the MIND STEP project is guided by concepts and principles presented in the previous sections. Figure 6 provides an illustration of components and linkages for model evaluation. The underlying idea is that the tasks of quality management and validation are entangled, mutually condition themselves but also complement each other. For instance, good and transparent documentation is one crucial requirement for quality management but also necessary for successful validation while the clear allocation of responsibility is important for quality management, but might be less relevant in the realm of validation. In general, Figure 6 emphasizes the point that model evaluation is a process consisting of a set of validation (or testing) techniques – grouped into validation, verification, as well as calibration – and supported or complemented by quality management, and (mostly) dependent on data. Furthermore, a key aspect of model validation within the project is involvement of stakeholders. For instance, two stakeholder workshops where organized within work package 1 to identify relevant policy areas and scenarios (Coderoni et al., 2020; Pérez-Soba et al. 2021). A third workshop to focus on model evaluation will be organized within Task 6.3 in autumn of 2022.

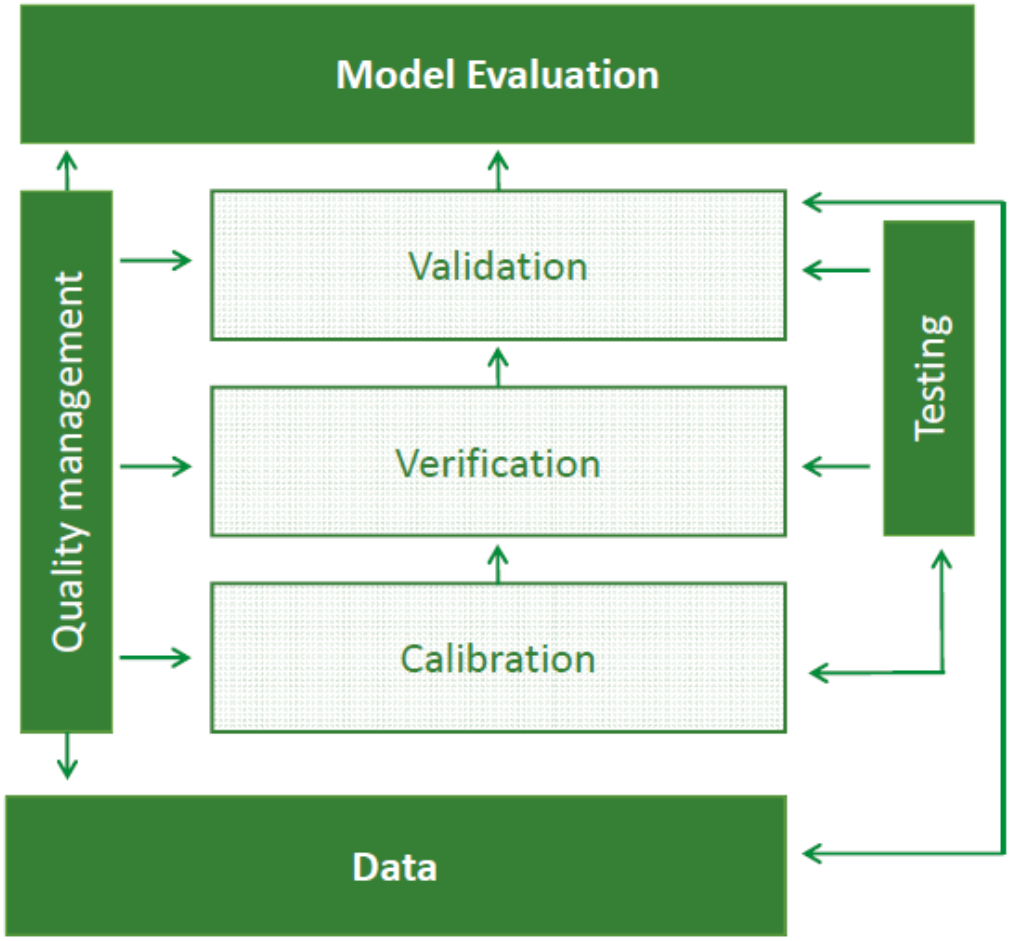

Figure 6: Framework for Model Evaluation within the MIND STEP project.

Because each of these components have different priorities and organization depending on the model and its purpose in MIND STEP, we aim to collect more information from the partners with a survey to identify important concepts and approaches but also gaps or challenges in the project. The survey is based on the "Checklist for the Quality of Models, Datasets and Indicators" by the Wageningen

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Modeling Group (Müller et al., 2021) and the structure of the "Modelling Inventory and Knowledge Management System of the European Commission" (MIDAS). Both sources where synthesized and adapted to specific requirements by the MIND STEP project. The latter concerns for instance the need for modularity (Britz et al., 2021) and the aim for sustainable research software development (Anzt et al., 2021).

The structure of the survey is shown in Table 2 and a more detailed presentation is provided in Appendix 2. In general, the survey includes all entries of the MIDAS database for two reasons: (a) Some of the models that belong to the MIND STEP model were and are already used by the European Commission to support policy making and thus already listed in the database. Hence, there was the aim to be consistent with information that is available and required for MIDAS. (b) The gathering of these information shall help to efficiently transfer new or updated information about the models to the MIDAS database if models or their results are or will be used by the European Commission. Beyond the MIDAS entries, the survey includes detailed questions about validation techniques and the potential for modularity.

The survey designed will be finalized within Task 6.2.1 (Validation of the MIND STEP model toolbox), where also the distribution (in terms of an online questionnaire) and its analysis will be conducted. The aim of the survey is to collect and synthesize important approaches to model evaluation as used within the MIND STEP project. On the one hand, we aim to provide a transparent documentation of validation efforts as it pertains to the individual models. This also includes the identification of challenges and gaps in model evaluation potentially due to lack of data or resources. On the other hand, we seek to provide a guideline on the evaluation of models and results for integrated applications as planned for Task 6.4 (Policy Evaluation) in the last phase of the MIND STEP project. With new or varied application as foreseen in this task, the guideline shall help to identify appropriate and potentially additional requirements to improve model validity and the credibility of the obtained results.

Table 2: Structure of the Survey.

Category	Description
Overview	Summarizes important information and features of the model including the objectives of the model, the modeling approach, the nature of input and output data, its spatial and temporal resolution, (potential) applications, and ownership of the model.
Quality	Concerns detailed information on model calibration, verification, and validation. For instance, if and how uncertainties are considered in the model, whether there is a sensitivity analysis conducted and if so how, and which validation techniques and criteria where used.
Transparency	Describes whether or under which conditions data, code, or results are available to third parties or the public; Provides information on extent of model documentation and whether version control, a management, or development plan are established.
Policy Support	Information on (potential) applications including benefits and potentials for policy assessment, the area of application, and which stage of the policy cycle can be targeted (e.g., policy formulation or implementation).
References	Provides additional resources like model documentation, model code, description of applications, peer-reviewed publications, or studies that use the model or its results.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

6. SUMMARY

With the increasing importance of simulation models in research in general and for the evaluation of policies within impact assessment in particular, there is increasing attention towards the reliability of these models and their result—in other words how credible can these models be used for decision support. Model evaluation, i.e., verification and validation, takes center stage in this question but also, or in combination to that is the necessity of quality management as both models and use cases become increasingly complex.

The present deliverable reports on options for quality management and model evaluation based on a comprehensive literature review. The report shows that there is a shift in the last couple of years in both of these areas. On the one hand, quality management does not only comprise certain requirements related to data quality, coding conventions or documentation of models, but also that specific expectations can be met that concern the design of models (modularity) and the maintenance of them (in terms of providing sustainable research software). On the other hand, the focus of model validation did also change. While empirical validation, i.e., confronting model structure and output with observational evidence, is still the most important aspect of validation, there is increasing emphasis on comprehensive and efficient interaction between modelers and stakeholders.

Considering these developments and based on existing frameworks for quality management and model evaluation, Task 6.1 developed the common framework and indicator system for model validation for the MIND STEP toolbox. The aim is to support validation across the work packages in MIND STEP and to provide a guideline and identification of indicators to assess model validity on different scales. Given the wide variety of different models and potential applications, the challenge was to provide a sufficiently flexible but also detailed framework to provide a reasonable source of information. The framework will serve as basis for an online survey conducted in Task 6.2 to transparently document model quality and validation among the partners, focusing on each individual model used or developed within the MIND STEP project.

7. ACKNOWLEDGEMENTS

This report is compiled for the H2020 MIND STEP project which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

8. REFERENCES

Addor, N., & Melsen, L. A. (2019). Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models. *Water Resources Research*, *55*(1), 378-390. https://doi.org/https://doi.org/10.1029/2018WR022958

Anderson, J. R. (1974). Simulation: methodology and application in agricultural economics. *Review of Marketing and Agricultural Economics*, *42*(430-2016-31038), 3-55.

Ang, F., & Kerstens, P. J. (2016). To mix or specialise? A coordination productivity indicator for English and Welsh farms. *Journal of Agricultural Economics*, *67*(3), 779-798.

Anselin, L. (1988). Model validation in spatial econometrics: a review and evaluation of alternative approaches. *International Regional Science Review*, *11*(3), 279-316.

Anzt, H., Bach, F., Druskat, S., Löffler, F., Loewe, A., Renard, B., Seemann, G., Struck, A., Achhammer, E., Aggarwal, P., Appel, F., Bader, M., Brusch, L., Busse, C., Chourdakis, G., Dabrowski, P., Ebert, P., Flemisch, B., Friedl, S., . . . Weeber, R. (2021). An environment for sustainable research software in Germany and beyond: current state, open challenges, and call for action [version

2; peer review: 2 approved]. *F1000Research*, *9*(295). https://doi.org/10.12688/f1000research.23224.2

Appel, F., & Loewe, A. (2021). Research software - Sustainable development and support. *IAMO Policy Brief*, *42*.

Axelrod, R. (1997). *Advancing the art of simulation in the social sciences*. Springer.

Axtell, R., Axelrod, R., Epstein, J. M., & Cohen, M. D. (1996). Aligning simulation models: A case study and results. *Computational & mathematical organization theory*, *1*(2), 123-141.

Balci, O. (1998). Verification, validation, and testing. *Handbook of simulation*, *10*(8), 335-393.

Balci, O. (2003). Verification, validation, and certification of modeling and simulation applications. Winter simulation conference,

Balci, O. (2004). Quality assessment, verification, and validation of modeling and simulation applications. Proceedings of the 2004 Winter Simulation Conference, 2004.,

Baldos, U. L. C., & Hertel, T. W. (2013). Looking back to move forward on model validation: insights from a global model of agricultural land use. *Environmental Research Letters*, *8*(3), 034024.

Bankes, S. (1993). Exploratory modeling for policy analysis. *Operations research*, *41*(3), 435-449.

Barde, S. (2017). A practical, accurate, information criterion for nth order markov processes. *Computational economics*, *50*(2), 281-324.

Barlas, Y., & Carpenter, S. (1990). Philosophical roots of model validation: two paradigms. *System Dynamics Review*, *6*(2), 148-166.

Beisbart, C. (2019). What is validation of computer simulations? Toward a clarification of the concept of validation and of related notions. In *Computer Simulation Validation* (pp. 35-67). Springer.

Bergdahl, M., Ehling, M., Elvers, E., Földesi, E., Körner, T., Kron, A., Lohauß, P., Mag, K., Morais, V., Nimmergut, A., Saeboe, H. V., Timm, U., & Zilhao, M. J. (2007). *Handbook on Data Quality Assessment Methods and Tools*. Eurostat.

Bert, F. E., Rovere, S. L., Macal, C. M., North, M. J., & Podestá, G. P. (2014). Lessons from a comprehensive validation of an agent based-model: The experience of the Pampas Model of Argentinean agricultural systems. *Ecological modelling*, *273*, 284-298.

Bharathy, G. K., & Silverman, B. (2013). Holistically evaluating agent-based social systems models: a case study. *Simulation*, *89*(1), 102-135.

Bousquet, F., & Le Page, C. (2004). Multi-agent simulations and ecosystem management: a review. *Ecological modelling*, *176*(3-4), 313-332.

Britz, W., Ciaian, P., Gocht, A., Kanellopoulos, A., Kremmydas, D., Müller, M., Petsakos, A., & Reidsma, P. (2021). A design for a generic and modular bio-economic farm model. *Agricultural systems*, *191*, 103133.

Burton, R. M. (2003). Computational laboratories for organization science: Questions, validity and docking. *Computational & Mathematical Organization Theory*, *9*(2), 91-108.

Burton, R. M., & Obel, B. (1995). The validity of computational models in organization science: From model realism to purpose of the model. *Computational & mathematical organization theory*, *1*(1), 57-71.

Canova, F., & Sala, L. (2009). Back to square one: Identification issues in DSGE models. *Journal of Monetary Economics*, *56*(4), 431-449.

Destatis. (2021). *Qualitaetshandbuch der Statistische Aemter des Bundes und der Laender (Destatis)*. https://www.destatis.de/DE/Methoden/Qualitaet/qualitaetshandbuch.html

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134-144.

Eker, S., Rovenskaya, E., Obersteiner, M., & Langan, S. (2018). Practice and perspectives in the validation of resource management models. *Nature communications*, *9*(1), 1-10.

Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A., & Roventini, A. (2019). Validation of agent-based models in economics and finance. In *Computer Simulation Validation* (pp. 763-787). Springer.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Finger, R., Droste, N., Bartkowski, B., & Ang, F. (2022). A note on performance indicators for agricultural economic journals. *Journal of Agricultural Economics*, *73*(2), 614-620.

Fresco, L. O., Geerling-Eiff, F., Hoes, A.-C., van Wassenaer, L., Poppe, K. J., & van der Vorst, J. G. (2021). Sustainable food systems: do agricultural economists have a role? *European Review of Agricultural Economics*, *48*(4), 694-718.

Garcia, R., Rummel, P., & Hauser, J. (2007). Validating agent-based marketing models through conjoint analysis. *Journal of Business Research*, *60*(8), 848-857.

Giannone, D., Reichlin, L., & Sala, L. (2006). VARs, common factors and the empirical validation of equilibrium business cycle models. *Journal of Econometrics*, *132*(1), 257-279.

Gocht, A., Neuenfeldt, S., Yang, X., Müller, M., Helming, J., Roerink, G., Sander, J., Oudendag, D., Kremmydas, D., & Brouwer, A. (2021). *A guide/handbook to build an interface for accessing the data in the project required by partners* (Deliverable 2.2 of the Project "Modelling Individual Decisions to Support the European Policies Related to Agriculture" (Grant Agreement No. 817566, CALL H2020-RUR-2018-2), Issue.

Happe, K., Kellermann, K., & Balmann, A. (2006). Agent-based analysis of agricultural policies: an illustration of the agricultural policy simulator AgriPoliS, its adaptation and behavior. *Ecology and society*, *11*(1).

Huber, R., Bakker, M., Balmann, A., Berger, T., Bithell, M., Brown, C., Grêt-Regamey, A., Xiong, H., Le, Q. B., & Mack, G. (2018). Representation of decision-making in European agricultural agent-based models. *Agricultural systems*, *167*, 143-160.

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, *21*(5), 602-614.

Janssen, S., & Van Ittersum, M. K. (2007). Assessing farm innovations and responses to policies: a review of bio-economic farm models. *Agricultural systems*, *94*(3), 622-636.

Karplus, W. J. (1983). The Spectrum of Mathematical Models. *Perspectives in Computing*, *3*(2), 4-13.

Kaye-Blake, W., Schilling, C., & Post, E. (2014). Validation of an agricultural MAS for Southland, New Zealand. *Journal of Artificial Societies and social simulation*, *17*(4), 5.

Kleijnen, J. P. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, *82*(1), 145-162.

Kremmydas, D., Athanasiadis, I. N., & Rozakis, S. (2018). A review of agent based modeling for agricultural policy evaluation. *Agricultural systems*, *164*, 95-106.

Lamperti, F. (2018). An information theoretic criterion for empirical validation of simulation models. *Econometrics and Statistics*, *5*, 83-106.

Louie, M. A., & Carley, K. M. (2008). Balancing the criticisms: Validating multi-agent models of social systems. *Simulation Modelling Practice and Theory*, *16*(2), 242-256.

Macal, C. M., & North, M. J. (2005). Validation of an agent-based model of deregulated electric power markets. North American Computational Social and Organization Science (NAACSOS) 2005 Conference.

Marks, R. E. (2013). Validation and model selection: Three similarity measures compared. *Complexity Economics*, *2*(1), 41-61.

McCallum, I., & Subash, A. (2021). *Prototype of the data services and download services* (Deliverable 7.6 of the Project "Modelling Individual Decisions to Support the European Policies Related to Agriculture" (Grant Agreement No. 817566, CALL H2020-RUR-2018-2), Issue.

McCarl, B. A. (1984). Model validation: an overview with some emphasis on risk models. *Review of Marketing and Agricultural Economics*, *52*(430-2016-31544), 153-173.

Mitchell, P. (1997). Misuse of regression for empirical validation of models. *Agricultural systems*, *54*(3), 313-326.

Moss, S. (2008). Alternative approaches to the empirical validation of agent-based models. *Journal of Artificial Societies and social simulation*, *11*(1), 5.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Mueller, M., Schaefer, D., Britz, W., Sckokai, P., Carpentier, A., Femenia, F., Offermann, F., Wang, S., Ang, F., Oude-Lansink, A., & Pahmeyer, C. (2021). *Specification of model requirements - Protocols for code and data* (Deliverable 3.1 of the Project "Modelling Individual Decisions to Support the European Policies Related to Agriculture" (Grant Agreement No. 817566, CALL H2020-RUR-2018-2), Issue.

Müller, M., Schäfer, D., Britz, W., Sckokai, P., Carpentier, A., Femenia, F., Offermann, F., Wang, S., Ang, F., Oude-Lansink, A., & Pahmeyer, C. (2021). Specification of model requirements - Protocols for code and data. Deliverable 3.1 of the Project ”Modelling Individual Decisions to Support the European Policies Related to Agriculture. In: Grant Agreement No. 817566, CALL H2020-RUR-2018-2.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge University Press.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641-646.

Pachepsky, L., Haskett, J., & Acock, B. (1996). An adequate model of photosynthesis—I Parameterization, validation and comparison of models. *Agricultural systems*, *50*(2), 209-225.

Panhans, M. T., & Singleton, J. D. (2017). The empirical economist’s toolkit: from models to methods. *History of Political Economy*, *49*(Supplement), 127-157.

Power, M. (1993). The predictive validation of ecological and environmental models. *Ecological modelling*, *68*(1-2), 33-50.

Rand, W., & Rust, R. T. (2011). Agent-based modeling in marketing: Guidelines for rigor. *International Journal of research in Marketing*, *28*(3), 181-193.

Reidsma, P., Janssen, S., Jansen, J., & van Ittersum, M. K. (2018). On the development and use of farm models for policy impact assessment in the European Union–A review. *Agricultural systems*, *159*, 111-125.

Robinson, S. (1999). Simulation verification, validation and confidence: a tutorial. *Transactions of the Society for Computer Simulation*, *16*(2), 63-69.

Rykiel, E. J. (1996). Testing ecological models: the meaning of validation. *Ecological modelling*, *90*(3), 229-244.

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of simulation*, *7*(1), 12-24.

Schlesinger, S. (1979). Terminology for model credibility. *Simulation*, *32*(3), 103-104.

Schwanitz, V. J. (2013). Evaluating integrated assessment models of global climate change. *Environmental Modelling & Software*, *50*, 120-131.

Segerson, K. (2015). The role of economics in interdisciplinary environmental policy debates: opportunities and challenges. *American Journal of Agricultural Economics*, *97*(2), 374-389.

Smajgl, A., House, A. P., & Butler, J. R. (2011). Implications of ecological data constraints for integrated policy and livelihoods modelling: An example from East Kalimantan, Indonesia. *Ecological modelling*, *222*(3), 888-896.

Stevens, W. P., Myers, G. J., & Constantine, L. L. (1974). Structured design. *IBM Systems Journal*, *13*(2), 115-139.

Swinton, S. M. (2018). Why Should I Believe Your Applied Economics? *American Journal of Agricultural Economics*, *100*(2), 381-391.

Topping, C. J., Dalkvist, T., & Grimm, V. (2012). Post-hoc pattern-oriented testing and tuning of an existing large model: lessons from the field vole.

Troost, C. (2015). Agent-based modeling of climate change adaptation in agriculture: a case study in the Central Swabian Jura.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

Trucano, T. G., Swiler, L. P., Igusa, T., Oberkampf, W. L., & Pilch, M. (2006). Calibration, validation, and sensitivity analysis: What's what. *Reliability Engineering & System Safety*, *91*(10-11), 1331-1357.

Utomo, D. S., Onggo, B. S., & Eldridge, S. (2018). Applications of agent-based modelling and simulation in the agri-food supply chains. *European Journal of Operational Research*, *269*(3), 794-805.

van Vliet, J., Bregt, A. K., Brown, D. G., van Delden, H., Heckbert, S., & Verburg, P. H. (2016). A review of current calibration and validation practices in land-change modeling. *Environmental Modelling & Software*, *82*, 174-182.

Werker, C., & Brenner, T. (2004). *Empirical calibration of simulation models*.

Windrum, P., Fagiolo, G., & Moneta, A. (2007). Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and social simulation*, *10*(2), 8.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

APPENDIX 1: TECHNIQUES FOR MODEL EVALUATION

Verification, Validation and Testing Techniques

Informal	**Static**	**Dynamic**	**Formal**
Audit	Cause-Effect Graphing	Acceptance Testing	Induction
Desk Checking	*Control Analysis*	Alpha Testing	Inductive Assertions
Documentation Checking	Calling Structure Analysis	Assertion Checking	Inference
Face Validation	Concurrent Process Analysis	Beta Testing	Lambda Calculus
Inspections	Control Flow Analysis	Bottom-Up Testing	Logical Deduction
Reviews	State Transition Analysis	Comparison Testing	Predicate Calculus
Turing Test	Data Analysis	Compliance Testing	Predicate Transformation
Walkthroughs	Data Dependency Analysis	Authorization Testing	Proof of Correctness
	Data Flow Analysis	Performance Testing	
	Fault/Failure Analysis	Security Testing	
	Interface Analysis	Standards Testing	
	Model Interface Analysis	Debugging	
	User Interface Analysis	*Execution Testing*	
	Semantic Analysis	Execution Monitoring	
	Structural Analysis	Execution Profiling	
	Symbolic Evaluation.	Execution Tracing	
	Syntax Analysis	Fault/Failure Insertion Testing	
	Traceability Assessment	Field Testing	
		Functional (Black-Box)Testing	
		Graphical Comparisons	
		Interface Testing	
		Data Interface Testing	
		Model Interface Testing	
		User Interface Testing	
		Object-Flow Testing	
		Partition Testing	
		Predictive Validation	
		Product Testing	
		Regression Testing	
		Sensitivity Analysis	
		Special Input Testing	
		Boundary Value Testing	
		Equivalence Partitioning	
		Extreme Input Testing	
		Invalid Input Testing	
		Real-Time Input Testing	
		Self-Driven Input Testing	
		Stress Testing	
		Trace-Driven Input Testing	
		Statistical Techniques	
		Structural (White-Box)Testing	
		Branch Testing	
		Condition Testing	
		Data Flow Testing	
		Loop Testing	
		Path Testing	
		Statement Testing	
		Submodel/Module Testing	
		Symbolic Debugging	
		Top-Down Testing	
		Visualization/Animation	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

APPENDIX 2: SURVEY OF QUALITY MANAGEMENT AND VALIDATION

Item	Explanation	MIDAS export
Overview		
Main Purpose	Brief description of the objective(s) of the model/database.	yes
Summary	The summary should mention the area of application (regional focus or generic) as well as the theoretical and methodological framework based on the underlying paradigms (positive or normative approach) and core assumptions.	yes
Model Type	Please specifiy the type of the model, e.g., partial or general equilibrium, ABM, (non-)linear programming	yes
Ownership	Who is the owner of the model	yes
Licence	e.g., open source	yes
Homepage		yes
Details On Model Structure And Approach	Highlighting the conceptual model and model structure with the main modules (e.g., flow charts) and processes, i.e., the major stages in excecuting the model.	yes
Modularity	Is the model integrated or linked to another model?	no
	If yes, specify if input from other models is used, the output of the model will be feed into (an)other model(s) or if the model is or can be integrated into a (meta)model.	no
The parameters, variables, inputs to and output of the model are described		no
Model Inputs	List of model input with description, unit of measurement, and data source	yes
Model Outputs	List of model output with description and unit of measurement.	yes
Model Spatial-Temporal Resolution And Extent	state: Spatial Extent/Country Coverage, Spatial Resolution, Temporal Extent, Temporal Resolution	yes
Calibration of parameters	Has the model been calibrated? Are criteria for the goodness of the calibration been described?	no

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Item	Explanation	MIDAS export
Quality		
Model uncertainties	Models are by definition affected by uncertainties (in input data, input parameters, scenario definitions, etc.). Have the model uncertainties been quantified? Are uncertainties accounted for in your simulations?	yes
	Please specify	yes
Sensitivity analysis	Sensitivity analysis helps identifying the uncertain inputs mostly responsible for the uncertainty in the model responses. Has the model undergone sensitivity analysis?	yes
	Please specify	yes
Model verification	The technical environment is documented	no
	The model is tested	no
	Please specify	no
Model validation	Has model validation been done? Have model predictions been confronted with observed data (ex-post)?	yes
	Brief summary of procedures that were used as detailed below.	yes
Animation	The model's operational behaviour is displayed graphically as the model moves through time. For instance, a grid of plots may represent the fields in a region and colors illustrate its current use while a change of the color represents land use change during the simulation.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Comparison to other models	The output of the (simulation) model is compared to the results of other (validated) models.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Cross validation/data splitting	Division of the available data into two data sets: an estimation and a prediction data set. The estimation data set is used to estimate model parameters and to assess the replicative validity of the resulting model. The prediction data set is used exclusively for predictive validation.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Data relationship correctness	Requires data to have the proper values regarding relationships that occur within a type of data and between different types of data. For example, are the input-output levels for a given production activity correctly presented.	no

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Item	Explanation	MIDAS export
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Degenerate tests	The degeneracy of the model's behaviour is tested by appropriate selection of values of the input and internal parameters. For instance, does the number of farms in a region decrease if the farm exit rate is larger than the farm entry rate.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Event validity	A comparison between the model and system is made of the occurrence, timing and magnitude of simulated and actual events. For instance, farm exit might be triggered by profitability or generation change.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Extreme condition test	The model structure and outputs should be plausible for any extreme and unlikely combination of levels of factors in the system.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Face validity	Expert assessment if the model and its behavior are reasonable, i.e., whether the model logic and input-output relationships appear reasonable 'on the face of it' given the model's purpose.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Historical data validation	Evaluate the model performance (results and dynamic behaviour) under given specifications against the historical system behaviour	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Internal validity	Several replications (runs) of a stochastic model are made to determine the amount of (internal) stochastic variability in the model.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Multistage validation	Validation methods are applied to critical stages in the model building process: (1) developing the model on theory, observations, and general knowledge; (2) validating the model's assumptions where possible by empirically testing them; and (3) comparing (testing) the input–output relationships of the model to the real system.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Item	Explanation	MIDAS export
Predictive validation	The model is used to forecast the system behavior and comparisons are made to determine if the system's behavior and the model's predictions are the same. The system data may come from data sets not used in model development or from future observations of the system. The strongest case is when the model output is generated before the data are collected.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Sensitivity analysis:	The (systematic) change of input values and internal parameters of a model to determine the effect upon the model's behaviour or output. Parameters that are sensitive, that is, cause significant changes in the model's behaviour or output (qualitatively and/or quantitatively), should be made sufficiently accurate prior to using the model.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Structured walkthrough	The entity under review is formally presented usually by the developer to a peer group to determine the entity's correctness. An example is a formal review of computer code by the code developer explaining the code line by line to a set of peers to determine the code's correctness.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Trace	The behavior of specific variables is traced through the model and through simulations to determine if the behavior is correct and if necessary accuracy is obtained.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Turing tests	Knowledgeable individuals are asked if they can discriminate between system and model outputs.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Statistical validation	A variety of tests performed during model calibration and operation. Three cases occur most often: (1) the model produces output that has the same statistical properties as the observations obtained from the real system; (2) the error associated with critical output variables falls within specified or acceptable limits; (3) several models are evaluated statistically to determine which test fits the available data.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Item	Explanation	MIDAS export
Visualization techniques	The dynamical behaviours of performance indicators are visually displayed as the (simulation) model runs through time to ensure that the performance measures and the model are behaving correctly. Such visualization can form the basis for comparisons between system and model and a subjective statement concerning the visual goodness of fit.]	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Others	Were other validation procedures (not listed above) applied?	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Data validity	Data are (judged) appropriate, accurate, and sufficient data are available. Data transformations are correct. Appropriate validation procedures (e.g., as listed above) are used to determine data validity.	no
	Brief description/justification of procedure and validation criteria (e.g., objective indicator or subjective judgement)	no
Transparency		
Availability of the underlying data	Is the model underlying database (i.e. the database the model runs are based on) publicly available?	yes
	Please specify	yes
Availability of model outputs	Can model outputs be made publicly available?	yes
	Please specify	yes
Model documentation	Is the model transparently documented (including underlying data, assumptions and equations, architecture, results) and are these documents available to the general public?	yes
	The references to model documentation are provided by …	yes
Accessibility to the model source code	Is the model source code publicly accessible or open for inspection?	yes
	Please specify	yes
Development plan	There is a development plan?	no
	Please specify	no
Version control system	There is a version control system?	no
	Please specify	no
Management plan	Is there a management plan?	no
	Please specify	no
POLICY SUPPORT		
Policy Role	Brief description (benefit and potential for policy assesment).	yes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Item	Explanation	MIDAS export
Policy Cycle	This model contribute to the following policy cycle: ANTICIPATION (foresight and horizon scanning), FORMULATION (impact assessments), IMPLEMENTATION (with monitoring), EVALUATION (ex-post evaluations)	yes
Policy Areas	This model can contribute to the following policy areas: Agriculture and rural development; Banking and financial services; Business and industry; Climate action; Competition; Consumers; Customs; Digital economy and society; EU enlargement; Economy, finance and the euro; Education and training; Employment and social affairs; Energy; Environment; European neighbourhood policy; Humanitarian aid and civil protection; Institutional affairs; International cooperation and development; Maritime affairs and fisheries; Public health; Regional policy; Research and innovation; Single market; Taxation; Trade; Transport	yes
Impact Assessments	Brief description with links to resources/publications	yes
References		
Studies That Uses The Model Or Its Results	Please enter references	yes
Peer Review For Model Validation	Please enter references	yes
Model Documentation	Please enter link or references	yes
Other Related Documents	Please enter link (website, GitHub etc) or references	yes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.