Methods for Integrated Simulation 10 Concepts to Integrate

Niki Popper[1*,2], Martin Bicher[1], Felix Breitenecker[3], Barbara Glock[1,4], Irene Hafner[1], Miguel Mujica Mota[5], Gašper Mušič[6], Claire Rippinger[1], Matthias Rössler[1], Günter Schneckenreither[1], Christoph Urach[1,4], Matthias Wastian[1,4], Günther Zauner[1,4], Melanie Zechmeister[1,4]

[1]Research Unit of Data Science, TU Wien, Favoritenstraße 9–11, 1040 Vienna, Austria; [*]*nikolas.popper@tuwien.ac.at*
[2]Research Unit of Computational Statistics, TU Wien, Wiedner Hauptstraße 8–10, 1040 Vienna; Austria
[3]Inst. of Analysis and Scientific Computing, TU Wien, Wiedner Hauptstraße 8–10, 1040 Vienna; Austria
[4] dwh Simulation Services, Neustiftgasse 57-59, 1070 Vienna;
[5] Amsterdam University of Applied Sciences, Aviation Academy, 1097 DZ Amsterdam, The Netherlands
[6] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia

SNE 32(4), 2022, 225-236, DOI: 10.11128/sne.32.on.10627
Received: 2022-10-15; Revised: 2022-12-10
Accepted: 2022-12-15
SNE - Simulation Notes Europe, ARGESIM Publisher Vienna, ISSN Print 2305-9974, Online 2306-0271, www.sne-journal.org

Abstract. This note summarises the current status of the work of EUROSIM's and ASIM's Technical Committees "Data Driven System Simulation" - with main emphasis on Big Data integration in simulation. This overview suggests ten developed concepts and methods, which should be considered, implemented and documented in modern simulation studies with Big Data.

Motivation

Diversity and heterogeneity of man-made systems is rapidly increasing and with it the costs spent on them. For a long time, these systems supposedly worked well. Currently, however, we are seeing a number of challenges, be it in the area of energy, mobility, logistics or health systems. Here, costs are rising, there are supply bottlenecks and, above all, the lack of resilience of the systems, i.e. the adaptation to changing conditions, is a major challenge.

Measuring efficiency and effectiveness of such systems is becoming increasingly complex, but is urgently needed. The development of new methods, models and simulations is necessary to support analysis, planning and control. Especially the possibility to calculate "what if" scenarios is an absolute necessity in order to be able to react to changing framework conditions.

The heterogeneity of the systems requires the possibility of integrating different modelling concepts in order to be able to depict the systems in sufficient detail. In addition, the quantity and quality of the available data are increasing strongly and thus facilitate the descrip-tion and analysis of such systems.

On the other hand, this increases the effort to parameterise, calibrate and validate the models and simulations. Bringing together the necessary technologies is thus itself an enormous challenge.

1 Outline

Data-based Demographic models have to be combined with models for the spread of diseases. Dynamic modelling concepts must be parameterised with dynamically changing data sets from various sources.

For system simulation an important aspect is the possibility to implement changes inside the system, like interven-tions within the computer model, and to analyse their effects. As a recent example see Covid-19 Modelling at TU Wien [1]).

Based on the concepts of equations, networks, algorithms and the causal understanding of the world, modelling and simulation have reached a high level in describing systems and processes, e.g. complex technical systems, ecological systems, production and logistics processes or socio-economic systems such as the healthcare system.

On the other hand, Big Data based on sensors and computations to measure our world has gained outstanding importance. Today, there is a range of technologies for building, monitoring and evaluating it. A multitude of activities can be observed in research, development, politics and the media.

Nevertheless, interfaces and methods for linking these technologies need to be intensified. In particular, complex socio-technical systems that link technologies and people should serve the goals of citizens at different levels, from a citizen's personal goals, e.g. in terms of work or mobility, to the management of health care by politicians and decision-makers. A prerequisite for this is the analysis of actual data, the prediction of future behaviour and the calculation of "what if" scenarios using simulation methods.

Next generation tools are needed to make the development, construction monitoring and analysis of such systems easier, faster, more reliable and - most importantly - understandable for decision makers and other stakeholders. The EUROSIM Technical Committee DDSS 'Data Driven System Simulation' [2] aims to support and coordinate combined research in the following areas:

Data. Integration, storage, management and analysis of very large data sets, unstructured data, secure and reproducible data management from sensors, IoT and different data sources such as dynamic databases or unstructured information sources. Development and Operation of Digital Twins and Synthetic Data Interfaces, Data acquisition, interfaces and analysis methods from statistics, machine learning and visual analysis.

Model. Formal, scalable modelling of different systems, heterogeneous modelling of subsystems and integration of these subsystems, development of modelling methods for computationally complex systems, multimethod modelling, including coupling and comparison based on data, system knowledge and application requirements. Development of innovative methods in numerical mathematics, co-simulation, hybrid simulation.

Processes. Linking data and models to simulation tools for complex systems and methods for reproducibility of results. Interfaces and visualisation of simulation results, decision support systems and the future development of Human-Computer Interaction (HCI).

Guidelines. Standardisation of the processes mentioned, modularisation of models, connectivity of simulations, comparability of the tools used as well as international guidelines on privacy, security, how to implement, test and quality assure specific technologies and how to integrate stakeholders.

In this article, the EUROSIM Technical Committee presents a first draft for ten concepts that should be considered for the implementation of modern and adequate simulation models.

2 Methods

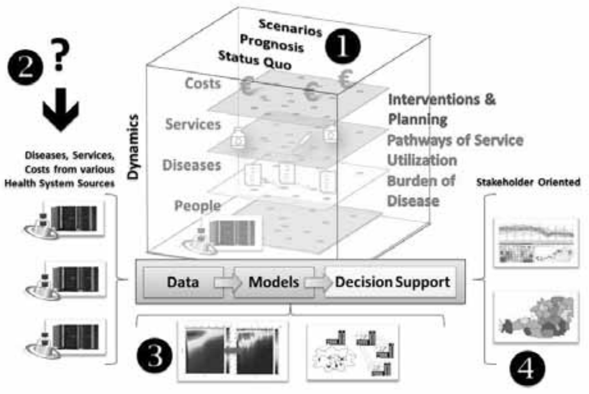

Figure 1: Schematic Overview DEXHELPP Infrastructure & Process (2014).

On basis of experiences of the Austrian DEXHELPP Competence Centre for Decision Support in Health Policy and Planning [3], which started in 2014 a concept was developed how large, interdisciplinary teams can handle these complex processes in the future and what are similarities and differences between health systems and other complex man-made systems. DEXHELPP developed an innovative research infrastructure with (1) a flexible virtualised health system, (2) methods to cope with data, (3) an adaptive analysis and simulation methods pool and (4) stakeholder oriented interfaces to enable researchers and other stakeholders to share data and methods for research and decision making (see Figure 1).

Within the framework of the development of this platform, corresponding methods were developed in six different sub-areas on the one hand, and on the other hand the methods were tested in practice with partners from the Austrian health system. Ten different areas in which there is a need for action were derived from this. These were compared with other disciplines within the framework of EUROSIM and expanded accordingly.

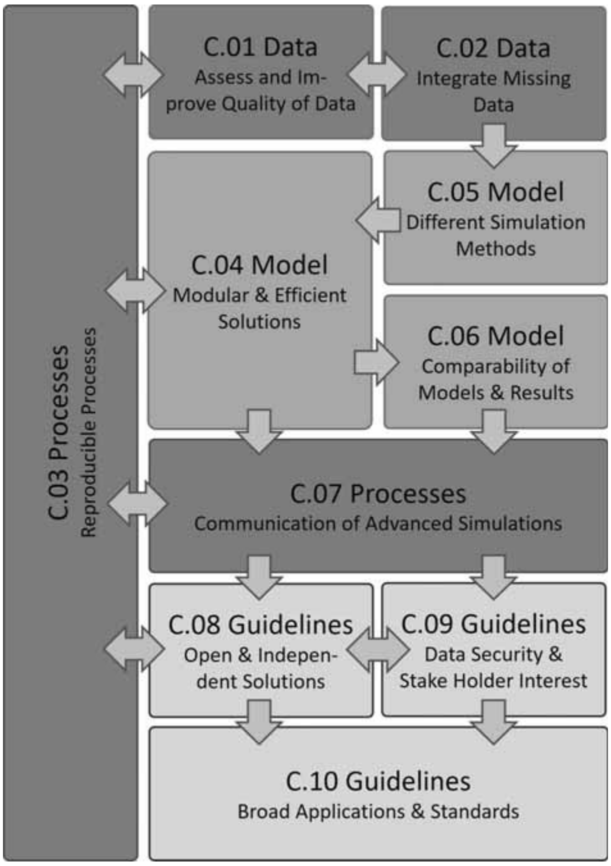
C.01 Data
Assess and Improve Quality of Data
C.02 Data
Integrate Missing Data
C.05 Model
Different Simulation Methods
C.04 Model
Modular & Efficient Solutions
C.06 Model
Comparability of Models & Results
C.03 Processes
Reproducible Processes
C.07 Processes
Communication of Advanced Simulations
C.08 Guidelines
Open & Independent Solutions
C.09 Guidelines
Data Security & Stake Holder Interest
C.10 Guidelines
Broad Applications & Standards

Figure 2: 10 Concepts to Integrate in Simulation Processes.

10 Concepts to Integrate were identified and presented at Invited Talks at University of Rostock [4] in 2018 and University of Stuttgart [5] in 2019 and are under development and extension with input from researchers in other applications and domains. The ten concepts are shown in Figure 2 and Table 1).

The concepts are structured into four thematic categories, which are **"Data"**, **"Model"**, **"Processes"** and **"Guidelines"**. This article does not include basic process mechanisms of modelling and simulation technology, which can be found in the literature. The concepts listed are state of the art and work in progress in their respective research areas and part of current research work. This paper aims to summarise their necessity and value for the development and operation of sustainable simulation studies.

3 Data

In this section challenges, ideas, examples, and benefits in the area of data integration into simulation studies are summarised.

	Concept
C.01	Assess and Improve Quality of Data
C.02	Integrate Missing Data
C.03	Reproducible Processes
C.04	Modular & Efficient Solutions
C.05	Different Simulation Methods
C.06	Comparability of Models & Results
C.07	Communication of Advanced Simulations
C.08	Open & Independent Solutions
C.09	Data Security & Stake Holder Interest
C.10	Broad Applications & Standards

Table 1: Overview Concepts.

Two concepts are included: **C.01 “Assess and Improve Quality of Data”** and **C.02 “Integrate Missing Data”** (shown as “Data” in Figure 2) offer the possibility to find wrong data and correct them, ideally also during the simulation process. For this purpose, methods of interactive visualisation and statistics are used, among others, to pre-process data collected unilaterally, e.g. sensor data or reimbursement data or to link data that are unstructured or have different structures without existing direct linkage. A particular challenge in simulation studies is the fact that not only the data itself can change during the project, but also the quality and structure. Furthermore, it may become necessary to integrate new data sources.

3.1 Challenge

The integration of large and heterogeneous data sets is an enormous challenge for classical simulation models. It is probably a central aspect why classical simulation has meanwhile fallen behind data science and machine learning approaches in some areas. In the areas mentioned, the integration of these data sources is in the foreground and is thus "thought of" from the beginning of a project. Classic simulation projects often focus on the development of methods and only later on data integration. Especially the lack of flexibility in data integration is often a challenge that is difficult to solve.

Concrete challenges in the area of data acquisition and data processing are the bias of collected data. Whether it is sensor data in technical applications or data collected in the case of the health system, e.g. in billing systems, this data must always be seen in the context of the purpose for which it was collected.

For example, in the health system this means that data collected for the purpose of billing often has a bias in that more expensive services are billed more often than they actually occur. In case of doubt, it is highly likely that, given two possible types of documentation, the one that is of greater benefit to the person, who documents will be chosen. However, if this data is used to estimate the burden of disease, the input parameters of the model are bound to be incorrect. In addition, pre-processing by different stakeholders (i.e. hospitals with different hospital information systems or doctors' practices) or the need to anonymise the data at an early stage of processing for data protection reasons pose further challenges. The handling of these points should be summarised and described in simulation projects in **C.01 "Assess and Improve Quality of Data"**. Furthermore, the data used in large simulation studies will usually not come from only one source. The question of how different data sets, which may be used, for example, in agent-based models for the parameterisation of the individual agents, must be brought together should be at the centre of considerations from the outset. Special attention should be paid to the fact that new data sources may be added in the course of projects lasting several years. An example of this is the development of a new therapy that is recorded in a new registry. This should be documented in a section on **C.02 "Integrate Missing Data"** and monitored continuously during the simulation study.

3.2 Implementation & Examples

For implementing **C.01**, a number of innovative methods are available. In DEXHELPP, the use of Explorative Visual Computing -Visual Analytics and Computational Statistics has proven to be an important building block in the implementation of the projects. The use of such methods makes it possible to identify outliers in data at an early stage in the simulation study, which have arisen through collection or preprocessing. A major limitation is that the reason for the bias is not directly apparent and the process view should be added to classical quantitative methods. In DEXHELPP, many projects were implemented with methods developed at VRVis (Centre for Virtual Reality and Visualisation). One example is [6]. In order to be able to evaluate a large number of time series with regard to their data quality, it is necessary on the one hand to automate the processes, but on the other hand to use necessary meta-information about the semantics of both the time series and the plausibility checks in order to structure and summarise the results of data quality checks in a flexible way. Already in this phase it is important to implement a comprehensive task analysis with domain experts and to derive processes from it in order to link quantitative methods with system knowledge. An example for statistical methods for the analysis of compositional data is [7]. Compositional data analysis refers to analysing relative information, based on ratios between the variables in a data set. In contrast to the interpretation as absolute information, it can be shown that already in data preprocessing not only different input parameters can be generated by means of univariate as well as multivariate statistical analyses, but that also important interpretations of these data for the modelling process arise.

Regarding the concept **C.02**, the necessary methods can be divided into two areas. On the one hand, suitable possibilities for the integration and linkage of (new) data sources must be embedded into the data processes, and on the other hand, parameterisation and calibration must be implemented with suitable methods for the modelling. Often, no unique identifiers are available for the record linkage of data. Agents can therefore not be parameterised appropriately. Therefore, the development of deterministic [8] and stochastic linkages is necessary [9]. Furthermore, the sustainable integration and combination of basic demographic data, structural socio-economic data and survey data is an example of a typical challenge in the health system. Optimisation and allocation algorithms are used in [10] to construct a structured population with corresponding temporal close-proximity interaction network from this data. This is particularly important for developing modular and reusable models. An example for linking such statistical population data with other data can also be found in [11]. Here, epidemiological data that are also used for the population model is combined with election data, e.g. voter turnout or arrival times of voters. Other necessary methods include the integration of AI methods to parameterise models based on historical data sets. In [12], for example, historical data of railway operations is used to parameterise a delay prediction model with a special focus on feature selection.

It is of particular interest that the two steps of integration and record linkage of data sources on the one hand and parameterisation and calibration on the other hand are in a direct interplay. I.e. the integration of new available data must be possible, but this directly results in new necessities as to how the model is parameterised and calibrated. Conversely, when the simulation is implemented with the domain experts, knowledge will continuously be gathered that leads to the necessity of integrating new data sources. Even if the problem is, of course, well known and described in principle, special attention must be paid to efficiency (automation) and feasibility with large and heterogeneous data sources as well as complex models during method development.

3.3 Benefit

The briefly mentioned methods are examples of which tasks in this area must not only be implemented, but also documented in a reproducible and comprehensible way in order to be prepared for the next points. A concrete goal (and evaluation criterion) is the possibility to identify wrong data sets and to exclude them from the further modelling process. With the potential to identify possible causes for the errors, a first benefit of the modelling process would be added. Furthermore, it must be possible to change wrong data sets, even over the time of the simulation project and after the data sets have been integrated and to include new data sets and data structures. We can process data to use it in subsequent simulation studies. This means that suitable methods are available to parameterise, calibrate and validate the model.

4 Model

To develop **C.04 “Modular & Efficient Solutions”** using **C.05 “Different Simulation Methods”** and maintain **C.06 “Comparability of Models & Results”** (summarised as “Model” in Figure 2) includes sustainable, modular models that can be quickly adapted to new problems and concepts for comparing, combining and linking models (qualitatively and quantitatively). The basic idea is that there is not only one methodology for modelling. The process of which method was chosen should be clearly presented, and the possibility of comparing or coupling models should be discussed if applicable and usable. Qualitative and quantitative comparison to analyse limitations of modelling approaches and implementation is possible, as well as methods like parameter transformation between models. The models should be modular in the case of usability in other areas so that, starting from data integration, the modules can be reused with minimal effort.

4.1 Challenge

Often, modelling methods proposed in the literature for dealing with the given questions are used for simulation projects - in line with good scientific practice. However, due to the change in available data described in the chapter "Data", emerging system knowledge and adopted research questions [13] there is often the possibility to try other modelling approaches. This potential should be used, but there is a risk of getting bogged down in the task: We need transparent, "simple" models.

In this respect, concepts should be developed to be able to implement modular simulation parts. Their reusability also serves to increase quality and sustainability. This is also necessary and helpful insofar as in many cases no established simulators can be used in practice for runtime reasons, but the solutions are programmed out fundamentally after a rapid prototyping and proof of concept phase. Therefore, the solutions should be kept as complicated but also as simple as possible. Especially in the case of complex processes, such as the use of buildings over time, it will make sense to couple existing models, i.e. not to develop new solutions that have already been tested and validated. However, this poses the challenge that (at least) interfaces and runtime behaviour have to be validated and documented again. These issues should be reviewed and addressed in item **C.04 "Modular & Efficient Solutions"**.

Based on this point, it becomes clear that there is a need to implement different levels of detail and different issues with different approaches. Therefore, it is necessary to define clear processes according to which criteria a model concept was selected for implementation. In this respect, simulation theory is now intensively concerned with the question of how models can be compared at all and how the "right" model can be selected: Based on this development, corresponding steps in the modelling process should be documented. An example would be the sensible representation of an epidemic spread by agents, if concrete interventions such as school closures are to be simulated and these cannot be represented in a comparable differential equation approach. Conversely, for the analysis of a basic system behaviour, a differential equation model (and the existing methods for analysis) should be used whenever possible. A possible combination of model parts for different (sub)systems should be considered and checked. The motivation for the choice and limitation of the chosen method, which arises from the data situation, the research question or the system knowledge, should be presented clearly and evidence-based in **C.05 "Different Methods for Different Research Questions"**. Under no circumstances should personal preferences of the modeller play a role here.

Last but not least, **C.06 "Comparability of Models & Results"** is by no means only about the obvious Comparability of Models & Results resulting from C.05, in the case that a question was calculated with two different models, but also about the possibility of making the model processes themselves comparable. This raises questions such as how parameter sets can be exchanged between microscopic and macroscopic models. Furthermore, it should be possible to better work out the limitations of individual approaches by comparing models. This will not be possible in every simulation study, but should be considered as a fundamental possibility.

4.2 Implementation & Examples

An approach for sustainable use of individual model parts, as should be fulfilled in **C.04**, is the development of a generic population model within the framework of DEXHELPP. The Generic Population Concept (GEPOC) has been developed since 2014 [14] and makes it possible to map different countries, flexibly integrate different sets of input parameters and use different modelling techniques (Agent Based, Discrete Modelling, System Dynamics). The population model is an example of the usefulness of modularisation, as the model cannot only be used in one application area, but the effects of interventions on the population play an important role in many areas. Examples of this are, in addition to the intervention and supply analysis in the health sector, also the use in the modelling of new mobility concepts or in the area of energy supply.

The current implementation for Austria simulates the population of Austria between 1998 and 2100 in such a way that historical data match the data of Statistics Austria, but also the forecasts match the respective assumptions of the national statistical authority up to a defined (small) error. The standard model is basically without interaction, but capable of it - i.e. this aspect is also optional for reasons of efficiency. In microscopic modelling, for example, agent properties are date of birth, sex and place of residence (latitude/longitude). Here, the link to the chapter "Data" is also established, since the possibility of being able to parameterise an existing, modular model again and again with different data is of great advantage here. The model was used, among other things, to advise the Austrian federal government during the Covid-19 crisis [15] and has proven itself to enable fast, flexible and quality-assured model implementation.

Different model derivations are managed on Git Hub as individual branches. Modules can be added step by step. In the case of Covid-19 modelling, these are a variety of different aspects, such as exact place of residence, or immunisation status. A fundamental example is the contact module, which is important for modelling dispersal. This was developed as part of an influenza modelling project starting in 2010 and implements the contact networks based on the POLYMOD study, a large survey of infection-related contact patterns, on the characteristics of 97,904 contacts recorded with 7,290 participants. Two aspects should be emphasised here: on the one hand, it is necessary to appropriately extrapolate the data-based contacts using statistical methods. On the other hand - based on model assumptions, as in the case of Covid-19 through contact restrictions, lock-downs and other measures - the modelled contact networks change. This must therefore be possible and as efficient as possible in the implementation. Accordingly, the model can also be used for effects in other areas, up to the analysis of possible couplings with other modelling approaches in multimethod modelling [16] or development of new interdisciplinary approaches [17].

Dealing with different models **(C.05)** was a starting point for the DEXHELPP platform. A rather simple example of comparing ODEs, PDEs, difference equations and CAs [18] was extended over the years by agent-based models and the respective modelling process parts such as parameterisation and cross model validation.

The aforementioned GEPOC model can also be used to illustrate how a model comparison **(C.06)** can be carried out [19]. Thereby, the methodological possibility of comparison forms the basis to make competent decisions why certain decision support should be implemented with concrete methods [20] or to what extent models can be combined to hybrid approaches [21]. In the course of the Covid-19 crisis, model comparisons were used in Austria in advising the federal government by implementing three different model approaches and comparing them on a weekly basis [23] , a current example of the comparison of models used (also internationally) is the ECDC Covid-19 Scenario Hub [24] The approaches mentioned are not limited to the level of population modelling, an example of which is [22]. The microscopic behavioural aspects like motion and proliferation of ‘pigment cells‘of the human skin are implemented using basic principles of agent-based simulation whereas the complex geometry of the microphysiological environment of melanocytes is modeled using the techniques of differential geometry. The combination of a small-scale behavioural model and the interaction with the complex environment allows to simulate and reproduce the growth of melanocytic skin lesions in silico.

4.3 Benefit

The aspects of modelling mentioned in the section are fundamentally linked to the data processes described in the section before. Assuming that a stable and validated modelling has been implemented, we can assume that with reasonable effort - if the requirements are not adapted - a change in the model strategy and a change in the parameterisation will not lead to any significant advantage. In contrary with changing requirements the model can be flexible improved if necessary. The possibility to represent the heterogeneity of the system under consideration is sufficiently given and the level of detail of the modelling is justified. This is supported by the possibility (and ideally implementation) of quantitative as well as qualitative comparison of different, methodically cleanly comparable methods. Differences in the results of different models are reasonable - as different model assumptions, aggregations or focus are set - and explained. Parameters can be transferred between diverse implementations and models. They can either be combined through multi-method modelling or suitably coupled through co-simulation if required. The implemented processes can clearly show the limitations of modelling and implementation.

5 Processes

C.03 "Reproducible Processes" and **C.07 "Communication of Advanced Simulations"** (Blocks 3 and 7 in Figure 2) are essential to guarantee the credibility and usability of the models and are decisive for the impact of decision-support models. Tools to manage and share data (e.g. [30] and concepts to communicate not only the simulation results, but also the modelling process and model construction are used. The reproducibility of processes and Data Citation Principles applied on all data sources is indispensable to be able to repeat the simulation studies at any time and thus increase the credibility of simulation technology itself in the medium term. Modelling steps such as the implementation of stochasticity, coupling of model parts and others are clearly documented. Selection and presentation of the results is justified, the conclusions are clearly presented and are related to the simulation results, especially when outcomes are relative (e.g. prioritisation of interventions) or qualitative. Improving the comprehensibility of the modelling process, simulation use and results of different categories through Data Representation and Human Computer Interfaces (HCI) and other strategies is essential to achieve the real purpose of simulation in the field of decision support, namely sound change management.

5.1 Challenge

From modelling practice we know that the data landscape is usually heterogeneous and that the data situation changes continuously in the modelling process and during the use of simulations. Different models with regard to different time scales, granularity of the representation of system variables and outputs as well as different properties make it difficult to keep simulation studies as reproducible as possible. In addition, the necessary and important regulations and standards regarding data security, personal protection and confidentiality must be observed. Nevertheless, in order to achieve credibility, it is essential that simulation studies in the future - like real experiments - should be reproducible worldwide in the laboratory, starting with the documentation of input data and parameters, through the systemic and strategic assumptions, the modular model parts and their documentation, to the specification of stochastic process assumptions. Corresponding methods must be developed and used in **C.03 "Reproducible Processes"**. Furthermore, projects still fail to reduce the model and simulation complexity in the direction of the decision-makers. On the one hand, complicated models are built precisely in order to do justice to the heterogeneity and dynamics of modern socio-technical processes, the reduction of which may not be (sensibly) possible. On the other hand, these models are not acceptable to the decision-makers because they are not comprehensible. In this respect, possibilities must be created (or used) in **C.07 "Communication of Ad-**

vanced Simulations", which can present models in a suitable way, can clearly present relevant mechanisms of action and break down the results of dynamic processes in a suitable way. Need for Change Management & Interdisciplinarity.

5.2 Implementation & Examples

As described in chapter "Data", one challenge is to deal with the changing data situation. New data is added, which not only fills existing structures with new data, but also changes the structures themselves **(C.03)**. Specific subsets of data are selected that fit the research question at hand. In this respect, we need methods to identify versions of a subset. Recommendations for this were developed, for example, by the RDA Working Group on Dynamic Data Citation (WGDC) in [25] with versioned data, timestamps and a query-based mechanism for subset formation. We also need fair data use, including concepts for managing the life cycle of research data that can be machine-processed with Data Management Plans (maDMPs), as in [26]. Simulation models can also be used to fill in missing data or generate new data. Data farming [27] enables the use of simulation models to generate data that can also be used for other methods such as machine learning. Simulation thus serves as a complement to observational data, the connection of which requires innovative methods to couple simulation, real-time data sources and the traditional historical data to verify and validate models [28]. But the convergence of physical and virtual worlds now goes far beyond this in cyber-physical systems (CPS), of course. The digital twin represents the maximum challenge, serving as a digital image of the physical world from the planning phase through strategic planning to operational use. It consists of a set of adaptive models that mimic the behaviour of a physical system and are updated along its life cycle by real-time data [29]. The digital twin has both simulation capabilities that can approximate the behaviour of the real system and emulation capabilities that allow the digital twin to synchronise with the real system and thus duplicate and mimic the physical system in the real world. The digital twin therefore offers more accurate replication compared to the simulation model and represents a new paradigm in modelling and simulation [29].

Last but not least, this also involves reproducibility at the "other end" of the simulation process, namely the use of the results of large agent-based models as a synthetic data source, as for example presented in the Covid-19 crisis here [30]. Other aspects include the conflicting goals of protecting and controlling sensitive data on the one hand, while allowing third party access on the other, as described in [31]. This is a major challenge especially for simulation studies, as the choice of modelling method is also affected here. The better the mechanisms are implemented in the data selection, the more detailed models can be applied. Here, the close connection of data, i.e. parameterisation, calibration and validation, to the model structure becomes apparent. Specifically for agent-based modelling in archaeology, this is outlined in [32], where the issue is up to clearly defining what stochastic range needs to be achieved in corresponding ones (see [33]) and how this can be mapped in the reproducibility discussion.

On the way to credibility and usability, achieving (and communicating) reproducibility is only one pillar. In addition, the communicability of the model results, the modelling process and the simulation study itself is equally decisive **(C.07)**. The importance of this was already shown in 2017 in [34] by means of an analysis of the extent to which people are more willing to be vaccinated if the benefits of vaccination are presented to them more clearly. This is a crucial aspect for the fruitful use of simulation models, as recently became apparent in the Covid-19 crisis. Even in phases in which the benefits of certain interventions were quite provable (in other phases these benefits were quite controversial), it became increasingly difficult to communicate these benefits widely. Three aspects are currently being researched intensively in this context: First, corresponding models must be clearly documented and communicated. Black box models whose structure is not comprehensible or understandable will justifiably not generate any benefit in the future, be it in the analysis of climate change or all other questions. Secondly, their use must be clearly documented and transparently implemented. In the field of health systems research, there are established processes for how questions are defined, how they are processed and how they are finally evaluated. In the case of the vaccine evaluation of Covid-19 in Austria, this was implemented and published in [35] including the involvement of a steering board. Last but not least, it is about clear communication of the results, as they are researched in the visualisation community. Here, not only are clear concepts for evaluation implemented, but these are evaluated themselves, as in [36]. Simulation research is still some steps away from this status.

5.3 Benefit

When implementing the above concepts, the results is not only a well-parameterised model with reasonable data sources, but also a simulation study that finds acceptance in the respective field of application. This can never be completely described technically, but two points are covered as well as possible: Firstly, the credibility of the simulation study was implemented with suitable documentation of the data used, the selected model modules, the implementation of the studies and all other framework conditions in such a way that the respective state of the art of reproducibility valid at the time of implementation was achieved. The experiment is reproducible. Secondly, the study can be appropriately communicated to those experts and the general public in accordance with the current state of the art. This also applies to the model structure, the implementation of the simulation study and the results. The study is comprehensible. Once this has been done to a sufficient extent, there is still the question of the connection to international standards, how a balance can be found between the needs of the clients (who also provide the data) and transparency and open access, as well as the question of domain-specific and interdisciplinary standards. These will be described in the next section.

6 Guidelines

Last but not least, guidelines, standards that go beyond the concrete implementation are crucial (Blocks 8-10 in Figure 2). Here, the concepts of **C.08 “Open & Independent Solutions”**, **C.09 “Data Security & Stake Holder Interest”** and **C.10 “Broad Applications & Standards”** are crucial. The possibility of publication is limited, for example, by (justified) economic or data protection interests, which, however, leads to a lack of comparability of different models and thus jeopardises quality. This requires fundamental regulations such as those addressed in the General Data Protection Regulation and Data Governance Act. Clear and transparent processes are necessary for every project (even before the start of a simulation development) as well as the reuse of models is necessary to ensure quality and sustainability over time. Standards for different domains have to be established and - as a vision - should be valid for simulation studies in all domains.

6.1 Challenge

The basic "possibility" of achieving credibility (as described in the previous chapter through reproducibility and comprehensibility) is currently often limited in reality for "external reasons", i.e. not due to the technical implementation of the simulation study. Data, model structure and documentation of the simulation study are often not published due to data protection, interests of data owners or clients. At the same time, a lack of comparability of (published) models, simulations and results leads to a lack of credibility of the discipline itself. Exaggeratedly formulated, one could write that in the case of simulation research, the lack of transparency leads to the discrediting of the entire discipline, as this lack is not attributed to a concrete implementation, but to the concept of simulation. Furthermore, non-publication prevents the further improvement of the quality of standardised model modules. For this reason, **C.08 "Open & Independent Solutions"** is an even more important concept and rules and guidelines regarding this will play an important role. This leads directly to the challenge of how to (pre-)define and guarantee stakeholder interests. This is the only way to clarify justified (or unjustified) objections at an early stage and to establish clear guidelines on the extent and aggregation of results that may be published. This is also the motivation for the concept **C.09 "Priority for Data Security and Stakeholder Interests"**, because only through this prioritisation will the necessary willingness to receive data and to be able to implement the publications to a sufficient extent be achieved. Of course, legal frameworks are still necessary and important. On the one hand, these must keep data protection in mind (GDPR) and, on the other hand, enable the necessary publication in the "public interest" in order to enable transparency in decision-making processes and efficient control of systems and processes. Through publication and the associated possibility of reusing models interdisciplinarily, on the one hand resources for the new and further development of models can be better used worldwide, and on the other hand the quality of the models can be better and more sustainably ensured. Examples of this would be the establishment of population models that follow the same standards worldwide and on which interventions can be simulated. This is an urgent challenge because - taking climate change and successful possible interventions as an example - domain boundaries are already history and individual aspects can no longer be considered singularly. Currently, the development of standards in sub-domains is at different stages of progress. Interdisciplinary standards are an urgent challenge that should be solved as soon as possible. Solutions should be developed in **C.10 "Broad Applications & Standards"** and implemented or referenced in individual simulation studies.

6.2 Implementation & Examples

The consequences of non-transparent, non-reproducible processes are twofold, for example in the case of concrete decision support in the area of the health system. Lack of credibility of results on the one hand and lack of availability of data or reproducible model assumptions (in a more general sense) on the other. The second aspect seems to be more cause than effect, but it turns out as follows: the knowledge that results of calculations are not shared and further used in a quality-assured way leads to irreconcilable differences between stakeholders as to who should provide which data and how they are processed. Especially in systems like the Austrian health system, where resources and decisions are shared between several stakeholders (in the case of the Austrian health system government, federal states and social insurance), lack of process quality are good arguments for not participating in a common analysis strategy. The result is diffuse or contradictory bases for decision-making.

Concepts **C.08** and **C.09** should therefore be considered together. On the one hand, open and transparent processes are needed; ideally, both the data sources and the processes and results should be published **(C.08)**. Open access journals, corresponding data platforms (described in the previous chapters) and GitHub servers are suitable for publishing the source code. Corresponding access points have been and are being continuously developed in the respective communities and should be sustainably linked to activities of EUROSIM and other simulation societies. In the context of concrete political decision-making processes, one should go one step further. In the analysis of the current immunisation against Sars-CoV-2, an up-to-date, model-based evaluation of this immunisation was published monthly within the framework of DEXHELPP [38], as concrete discussion processes were also continuously accompanied on the basis of this assessment.

On the other hand, it must already be clear before the start of a simulation project which partners provide which data under which boundary conditions **(C.09)**. In the context of governance, it must be clarified which stakeholders and scientific partners have access to which data aggregates, how and where these are linked and processed with which methods, and which results are published in which resolution by which stakeholder. These aspects must be clarified legally, technologically and formally and documented in writing. The two aspects C.08 and C.09 are directly related and it is short-sighted to think that the data issue can be considered in isolation from the methodology. A proof of concept for the whole process was implemented as an integrated solution, the DEXHELPP Research Server, within the COMET project DEXHELPP [37].

Last but not least, modularisation for the purpose of a cross-domain use of reusable, quality-assured models as well as guidelines and the standardisation of methods and their use are important steps for the future use of simulation in decision support **(C.10)**. In health systems research, the SMDM/ISPOR Modeling Good Research Practices [39] should be mentioned as an example, which define quality .and selection criteria for methods. It will also become necessary to define these standards and guidelines across domains. Just as energy consumption, energy transfer, energy storage and other aspects must be considered together, this also applies to simulation studies and models of the future. From the health system to mobility and climate, integrated models will be needed in the future in order to be able to depict the heterogeneity and dynamics appropriately.

6.3 Benefit

In the final chapter, the areas of open access and public domain were briefly outlined, especially in connection with the protection and planning of stakeholder interests, as well as the question of modularisation, standardisation and guidelines. In a simulation study, clear rules are documented in advance as to which results are published in which aggregation and which are not. The non-publication is justified, as are the legitimate interests of all parties involved. Methods are available to reuse as many model parts as possible under regulated framework conditions, either on one's own or in the research network.

7 Summary

There are many developments in the respective communities. At this point, an attempt has been made to briefly outline - in a first statement - which aspects are specifically crucial for modern simulation studies and which points are recommended for attention for specific projects. The aim is to show a spectrum of possibilities that should be considered, tested and ideally integrated as far as possible.

Some of the major challenges (and weaknesses) are exemplary: Centralised data bunkers, are too inflexible to keep up with ongoing changes. Models that are not scalable according to data, new system knowledge or research focus - and that are not comparable - weaken credibility and usability. Lack of interfaces and willingness to cooperate between models and other methods prevent optimal solutions.

We therefore need **Flexibility**, through decentralised (and secure) storage and documented, professional processes. **Sustainability** through modular models and simulations as well as **Appreciation** between methods from data science, mathematics and simulation.

References

[1] Covid-19 Simulation in Austria, http://www.dexhelpp.at/en/appliedprojects/tmp/covid-19/, [Online; accessed 8.9.2022]

[2] EUROSIM Technical Committee "Data Driven System Simulation", https://www.eurosim.info/tcs/tc-ddss/, [Online; accessed 8.9.2022]

[3] DEXHELPP, http://www.dexhelpp.at/en/project-description/state-of-the-art/, [Online; accessed 8.9.2022]

[4] Abstract Colloquium Talk 14.6.2018 "Sharing Data, Methods, a. Simulation Models - New Opportunities for Digital Health Care", https://www.informatik.uni-rostock.de/veranstaltungen/detailansicht-des-events/n/kolloqiumsvortrag-von-nikolas-popper-director-dexhelpp-wien-coordinator-centre-for-computational-complex-systems-tu-wien-46337/ [Online; accessed 8.9.2022]

[5] Abstract Faculty Talk 25.3.2019 "Integrated Processes for Modelling & Simulation", https://www.sfbtrr161.de/newsandpress/events_sfbtrr161/pastevents/, https://www.sfbtrr161.de/newsand press /downloads/Stuttgart_Faculty_Talk _20190325 _PrintVersion_s.pdf [Online; accessed 8.9.2022]

[6] Arbesser C, Spechtenhauser F, et al. Visplause: Visual Data Quality Assessment of Many Time Series Using Plausibility Checks. *IEEE Transactions on Visualization and Computer Graphics* 23, Nr. 1 (Januar 2017): 641–50. DOI 10.1109/TVCG.2016.2598592.

[7] Mert MC, Filzmoser P. Endel G, Wilbacher I. Compositional Data Analysis in Epidemiology. *Statistical Methods in Medical Research* October 6, 2016, 096228021667153. doi:10.1177/0962280216671536.

[8] Popper N, Glock B, et al. Deterministic Record Linkage of Health Data as Preparatory Work in Modelling and Simulation-Use Case: Hospitalizations in Austria. *Proceedings of the 6th International Workshop on Innovative Simulation for Health Care*. pp. 44-49 (2017).

[9] Glock B, Endel,F. et al. Challenges and Results with the Record Linkage of Austrian Health Insurance Data of Different Sources. *Informatics for Health Conference* 2017 (24 – 26. April, Manchester, UK)

[10] Schneckenreither G, Popper N. Dynamic multiplex social network models on multiple time scales for simulating contact formation and patterns in epidemic spread. *2017 Winter Simulation Conference (WSC)* 2017, pp. 4324-4336. DOI 10.1109/WSC.2017.8248138

[11] Weibrecht N, Roessler M, et al. How an election can be safely planned and conducted during a pandemic: Decision support based on a discrete event model. *Plos one* 2021, 16(12).

[12] Leser D, Wastian M, et al. Comparison of Prediction Models for Delays of Freight Trains by Using Data Mining and Machine Learning Methods. *SNE Simulation Notes Europe* 2019, 29(1), 45–47. DOI 10.11128/sne.29.sn.10467

[13] Popper N. Comparative modelling and simulation a concept for modular modelling and hybrid simulation of complex systems *Phd Thesis* 2015, TU Wien Repos. DOI 10.34726/hss.2015.21448

[14] Bicher M, Urach C, Popper, N. GEPOC ABM: A generic agent-based population model for Austria *in Proceedings of the 2018 Winter Simulation Conference* 2018, pp. 2656-2667. DOI 10.1109/WSC.2018.8632170

[15] Bicher M, Rippinger C, Urach C, Brunmeir D, Siebert U, Popper N. Evaluation of Contact-Tracing Policies against the Spread of SARS-CoV-2 in Austria: An Agent-Based Simulation. *Medical Decision Making, 41-8*. pp. 1017-1032 (2021).

[16] Glock B, Popper N, Breitenecker F. Various Aspects of Multi-Method Modelling and its Applications in Modelling Large Infrastructure Systems like Airports. *The 27th European Modeling and Simulation Symposium* 2015, (pp. 197-206). Proceedings of the European Modelling and Simulation Symposium 2015.

[17] Heinzl B, Rößler M, Popper N, et al. Interdisciplinary strategies for simulation-based optimization of energy efficiency in production facilities. *UKSim 13th International Conference on Computer Modelling and Simulation* 2013, pp. 304-309, IEEE.

[18] Schneckenreither G, Popper N, Zauner G, Breitenecker F. Modelling SIR-type epidemics by ODEs, PDEs, difference equations and cellular automata – A comparative study. *Simulation Modelling Practice and Theory* 2008, 16(8), 1014-1023.

[19] Bicher M, Glock B, Miksch F, Popper N, Schneckenreither G. Definition, validation and compari-son of two population models for Austria. *Int. Journal of Business and Technology* 2015, 4(1), 7.

[20] Miksch F, Jahn B, Espinosa KJ, Chhatwal J, Siebert U, Popper N. Why should we apply ABM for decision analysis for infectious diseases? An example for dengue interventions. *PloS one* 2019, 14(8).

[21] Emrich S, Breitenecker F, Zauner G, Popper N. Simulation of influenza epidemics with a hybrid model-combining cellular automata and agent based features. *ITI 30th International Conf. on Information Technology Interfaces* 2008, (pp. 709-714). IEEE.

[22] Schneckenreither G, Tschandl P, Rippinger C, et al. Reproduction of patterns in melanocytic proliferations by agent-based simulation and geometric modeling. *PLOS Computational Biology* 2021, 17(2): e1008660. DOI 10.1371/journal.pcbi.1008660

[23] Bicher M, Zuba M, Rainer L, et al. *Supporting Austria through the COVID-19 epidemics with a forecast-based early warning system.* 2020, medRxiv, 2020-10. DOI 10.1101/2020.10.18.20214767

[24] ECDC Covid-19 Scenario Hub, https://covid19scenariohub.eu/ [Online; accessed 4.11.2022]

[25] Rauber A, Asmi A, Van Uytvanck D, Proell S. Identification of reproducible subsets for data citation, sharing and re-use. *Bulletin of IEEE Technical Committee on Digital Libraries* 2016, Special Issue on Data Citation, 12(1), 6-15.

[26] Blumesberger S, Gänsdorfer N, Ganguly R, et al. FAIR Data Austria – Abstimmung der Implementierung von FAIR Tools und Services. *Mitteilungen Der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 2021, 74(2), 102–120.
DOI 10.31263/voebm.v74i2.6379

[27] Sanchez SM. Data farming: Methods for the present, opportunities for the future. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 2020, 30(4), 1-30. https://doi.org/10.1145/3425398

[28] Scala P, Mota M, Blasco-Puyuelo J, Garcia-Cantu O, Blasco C. A novel validation approach for validating the simulation model of a passengers' airport terminal: case study Palma de Mallorca airport. *EMSS 2022* 2022, Rome, Italy.

[29] Semeraro C, Lezoche M, Panetto H, Dassisti M. Digital twin paradigm: A systematic literature review. *Computers in Industry* 2021, 130, 103469.

[30] Popper N, Zechmeister M, BrunmeirD, et al. Synthetic Reproduction and Augmentation of COVID-19 Case Reporting Data by Agent-Based Simulation. *Data Science Journal, 20(1),* pp. 16ff (2021).

[31] Weise M, Kovacevic F, Popper N, Rauber A. OSS-DIP: Open Source Secure Data Infrastructure and Pro-esses Supporting Data Visiting *Data Science Journal,* 2022, 21(1).

[32] Popper N, Pichler P. *Reproducibility. In Agent-based Modeling and Simulation in Archaeology* 2015, pp. 98, Springer, Cham.

[33] Bicher M, Wastian M, Brunmeir D, Popper N. *Review on Monte Carlo Simulation Stopping Rules: How Many Samples Are Really Enough?* 2022, Simulation Notes Europe, 32(1), 1-8.
DOI 10.11128/sne.32.on.10591

[34] Betsch C, Böhm R, Korn L, Holtmann C. *On the benefits of explaining herd immunity in vaccine advocacy.* 2017, Nature human behaviour, 1(3), 1-6.

[35] Jahn B, Sroczynski G, Bicher M, Rippinger C, Mühlberger N, Santamaria J, et al. *Targeted covid-19 vaccination (tav-covid) considering limited vaccination capacities—an agent-based modeling evaluation.* 2021, Vaccines, 9(5), 434.

[36] Isenberg T, Isenberg P, Chen J, Sedlmair M, Möller T. *A systematic review on the practice of evaluating visualization.* 2013, IEEE Transactions on Visualization and Computer Graphics, 19(12), 2818-2827.

[37] Popper N, Endel F, Mayer R, Bicher M, Glock B. Planning Future Health: Developing Big Data and System Modelling Pipelines for Health System Research. *Simulation Notes Europe* 2017, 27(4), 203-208.
DOI 10.11128/sne.27.tn.10396

[38] Bicher M, Rippinger C, Schneckenreither G, Weibrecht N, Urach C, Zechmeister M, et al. *Model based estimation of the SARS-CoV-2 immunization level in Austria and consequences for herd immunity effects.* 2022, Scientific Reports, 12(1), 1-15.

[39] Caro JJ, Briggs AH, Siebert U, Kuntz KM *Modeling good research practices — overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force.* 2012,
Medical Decision Making, 32(5), 667-677.