REPRODUCIBILITY AND REPLICABILITY OF SIMULATION MODELS

Yilin Huang

Section Systems Engineering and Simulation, Faculty of Technology, Policy and Management,
Delft University of Technology, Delft, Netherlands

ABSTRACT

Computer simulation is increasingly complex and popular in virtually every domain. But computer models and experiments are rarely reproduced or replicated by independent researchers. With the goal to form a stronger community for (computationally) reproducible and/or replicable simulation models, and to encourage collaboration on the topic, this paper aims to highlight the values and challenges of simulation Reproducibility and Replicability (R&R), and call for more R&R research. It first reviews the terms R&R, and then presents different views and opinions on the topic. It explains the separation of method and result stages in reproducibility assessment, and discusses typical challenges in conducting R&R studies. Given the current complexity and widespread use of diverse simulation models, this paper argues that the R&R of such models must be explicitly integrated into existing modelling workflows. Researchers who engage in these efforts face numerous methodological challenges, many of which remain under-researched.

Keywords: reproducibility, computational reproducibility, replicability, methdology, open simulation.

1 INTRODUCTION

Computer simulation models are increasingly used in science across different disciplines. The results of simulation experiments shall uphold reproducibility, a fundamental tenet of scientific studies [1, 2]. Despite well-recognized critical necessity of reproducibility, and that all research should be reproducible, there is a growing concern among scientists that too few scientific studies can be reproduced; some even termed this as a "reproducibility crisis" [1, 3, 4, 5]. For example, Baker [3] reported that more than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Many challenges in reproducible research are persistent [6].

There is no consensus in scientific literature on what reproducibility is or should be [3]. A closely related term is replicability. "These terms – and others, such as repeatability – have long been used in relation to the general concept of one experiment or study confirming the results of another. Within this general concept, however, no terminologically consistent way of drawing distinctions has emerged" [7]. Similarly, the term docking is also used, e.g., in [8, 9], referring to the validation & verification (V&V) method used for aligning multiple models (code) implemented based on the same core conceptual model.

In the field of Modelling and Simulation (M&S), reproducibility (and replicability), or R&R in short, is one of the grand challenges discussed by Yilmaz in [10]. They are even less studied compared to that of traditional (non-simulated) experiments [11]. Computer models and experiments are rarely reproduced or replicated by independent researchers [10, 12, 13]. A few such studies exist. For example, Jalali et al. [14] assessed 1,613 articles that applied simulation modelling as a core method in health policy and epidemiology. Almost half of those articles did not report model details. A more in-depth evaluation of 100 these

Proc. of the 2025 Annual Modeling and Simulation Conference (ANNSIM'25)
May 26-29, 2025, Universidad Complutense de Madrid, Madrid, Spain
J.L. Risco-Martín, G. Rabadi, D. Cetinkaya, R. Cárdenas, S. Ferrero-Losada, and A. Bany Abdelnabi, eds.

articles showed that only seven out of 26 evaluation criteria were satisfied by more than 80% of those articles. About two percent of these articles provided modeling code and had reproducibility discussions. Zhang and Robinson [15] searched six prominent journals for articles focused on Agent-based Modelling (ABM), where nine out of 348 resulting articles aimed for replicating an existing model partially or entirely. The works of [16, 17] showed that different implementations (on different modelling platforms) of the same conceptual model gave significantly different results. Existing works suggest that the challenges of R&R for computational modeling may be more persistent than that for traditional experiments [11, 18].

With the goal to form a stronger community in M&S for reproducible and/or replicable simulation models and to encourage collaboration on the topic, this paper aims to highlight the values and challenges of simulation R&R based on literature, and call for more R&R research. It starts with a brief review on the terminology used for R&R. Researchers in different domains do not have to align the use but should be well-aware of the difference so that relevant work can be discovered in literature across domains and disciplines. It is followed by different views and opinions on the reproducibility of simulation studies. The paper then presents a few ways of assessment, and typical challenges in conducting R&R studies. It ends with a discussion argues that given the current complexity and widespread use of diverse simulation models, reproducibility of such models needs to be explicitly addressed operationally in existing modelling workflows by researchers who wish to engage in the efforts, for which many methodological issues are under-researched.

2 REPRODUCIBILITY AND REPLICABILITY IN SIMULATION

The language and conceptual framework of research reproducibility are non standard [4]. Different scientific disciplines and institutions use the terms of reproducibility and replicability in inconsistent or sometimes even contradictory ways; some use them interchangeability [1, 3, 7, 15, 19, 20]. In M&S, replications are also used referring to repeated runs of simulation models with different seeds of the random-number generators but otherwise the exact same model configurations [21]; they are known as independent replications (or runs) of simulation.

For clarity, in this paper, reproducibility narrowly refers to the "computational reproducibility" of simulation models; we make the following distinction of R&R in M&S based on [7, 22].

- *Reproducibility* is obtaining identical results using the same input data; computational steps, methods, and code; and conditions of analysis.
- *Replicability* is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data and/or uses different code.

According to this distinction, simulation studies that use the same core conceptual model but have different computational implementations (as those reported in, e.g., [15, 19, 23]) are replication studies, regardless if the studies use the same input data. Modellers commonly distinguish model conceptualization from model implementation. The former, or a conceptual model, is often a textual, mathematical and/or diagrammatic description of model characterization and processes of interaction, based on a real system of modelling interest [11]. A conceptual model is often not executable, thus it has some ambiguities in how to compute model inputs to outputs [24]. The latter, or model operationalisation, or simply a simulation model, is a computational formalization of a conceptual model into an executable computer program where numerical output can be produced by executing (i.e., running) the implemented model [11].

Reproducing simulation studies aims to demonstrate that a computer simulation experiment's findings are repeatable and were not an exception [12, 24]. Without verifying claimed results through reproducible simulation experiments, it is possible that published findings were incorrect due to, e.g., programming errors, mistakes in the analysis or reporting of results, or misrepresentation of the simulation experiment [13]. Consistent results from model replications can build confidence in the simulation mechanism used [25].

3 REPRODUCIBLE RESEARCH: WHY AND WHY NOT

There are different views and opinions on the reproducibility of simulation studies in diverse research areas and due to personal preferences which can be equally interesting or at least open to debate. For example, reproducibility can be hard or practically impossible to achieve for certain research areas such as military and critical infrastructure, and for some topics in medicine and other industries due to business interests, security, privacy or ethical reasons among others. These simulation applications often require non-disclosure agreements or are demanded by laws or regulations that limit or prevent the publication of information that is necessary to reproduce the simulation [10, 26].

Making simulation models (and results) R&R demands dedicated time and efforts from the original researcher. This is also true for the researcher or team other than the original researcher who tries to reproduce the work, hereafter simply the second researcher(s). As some put it, "publication is already a grueling process, why would we increase our workload [26]?" There is no direct reward or consequence for researchers (hence little incentive) to undertake what they view as additional work to make their study more open and accessible [26].

There are also voices that criticise R&R in research. For example, Drummond [27] raises concerns about the strong influence the reproducible research movement is having on which papers get published; in addition, widening the responsibilities of peer reviews adds extra workload to reviewers, and does not recognize the broad role the scientific community plays (at the post-publication stage) in determining the value of an idea. The same author [28] states that reproducible research in some fields also requires open source code, which is a narrow interpretation of how science works; the effort necessary to meet the aim, and the general attitude it engenders would not serve well any of the research disciplines. Fanelli [29] in his PNAS article argues that the rapidly growing scientific literature uncritically endorses a new "science is in crisis" or "reproducibility crisis" narrative, which is not only empirically unsupported and unreliable, but also quite counterproductive and might foster cynicism and indifference in younger generations.

No matter what the views are or where the truth lies, clearly, not all models need or can be made R&R. It is a value and (voluntary) community service being placed on good science by some researchers. For them, including the author of this paper, reproducing and/or replicating a computational model can contribute to the scientific community in many different ways that cumulatively consolidate science. Commonly discussed in literature are, e.g., developing shared understanding; obtaining an improved sense of the accuracy, robustness and range of plausibility of model results; points of difference could allow empirical evidence to discriminate between the models [4, 9, 15, 19, 24].

In addition, reproducing and/or replicating a computational model can be a good way to have a first assessment of the reusability of reported simulation models [26, 30]. The knowledge and data embodied by the simulation model is available to be utilized by model re-users as a tool to advance their own research agenda [24]. Simulation studies often have two main audiences: methodological researchers and applied researchers [31]. The former reads a study to gain an overview of a method's uses, limitations and potential improvements. The latter reviews a study with the main aim of applying the method or result used to their own research problem. Reuse of simulation models can benefit both types of researchers. Model reuse allows testing the wider parameter space of the existing simulations [17]. It can also facilitate multi-modelling and hybrid simulation, i.e., combining different models for the application of complex systems analysis [32, 33].

4 ASSESSMENT OF SUCCESS

For M&S, it is beneficial to examine R&R from two aspects of simulation: the methods (i.e., the computational procedure of the simulation) and the simulated results. This distinction is inline with the "method reproducibility" and "result reproducibility" discussed in [4]. The former means that the computer simulation is methodologically reproducible in theory and practice; the latter means that the simulated results can be quantitatively reproduced using the same computational method. Clearly, method reproducibility proceeds (and is necessary for) result reproducibility.

Method Reproducibility. When the original simulation model (i.e., code and input data) is accessible to the second researcher, then the method reproducibility can be directly assessed. This typically means that the second researcher follows the simulation computational procedure, as exactly as possible, using the same code, tools and data, based on the original documentation and publication. If the original model is not accessible, method reproducibility has to be assessed by newly implementing the simulation model based on the original publication and the model conceptualization discussed within. This is typically known as replicating a simulation model. This, of course, can also be done independently when the original code and data are accessible. The actions needed to assess model reproducibility may appear straightforward; however, there are many associated challenges. These are discussed in Section 5.

Result Reproducibility. After method success, if the reproduced or replicated simulation generates outputs sufficiently similar to that of the original model, the reproduction or replication as a whole can be considered successful [12]. The quantitative measure of the similarity of results (i.e., result reproducibility), however, is neither straightforward [34]. For example, Muradchanian et al. [35] reported on the difficulties in comparing multiple Frequentist and Bayesian measures because there is no established standard on the type of metrics used. Another complicating factor in the comparison is different levels of publication bias [35]. In relation to that, a broader categorization of results reproducibility is the so called "standards of equivalence" or "replication standards". This is referenced in several studies [11, 12, 15] and first discussed in the replication work of [9]. The work includes three general categories of model equivalence (from strict to loose): numerical, distributional and relational equivalences, as follows.

1. Numerical equivalence (or identity) refers to the reproduction of exact reported results. It typically is not expected for stochastic simulations unless information on random seeds were specified.
2. Distributional equivalence is determined by showing that two studies produce distributions of results that cannot be distinguished statistically. This is determined by statistical test of null hypotheses.
3. Relational equivalence means that two models can be shown to produce the same internal relationship among their results (i.e., inputs, parameters and outputs). For example, two models show that a particular output variable is a quadratic function of time, or that a measure on a population decreases monotonically with the population size [9]. This is the least demanding comparison, but for some theoretical purposes, it may suffice.

The separation of method and result reproducibility is particular useful for simulation studies because R&R studies of complex models can be divided into two stages that are more manageable. The simulation method itself can be examined and tested or replicated first for methodological soundness, which is an important contribution of a simulation study. This stage also verifies the alignment of the conceptual model with the computational model and the experimental scenarios (and conditions). It tests that with the stated computational steps, whether the computational method executes as intended. In the second stage, the results are compared, which is typically performed in a traditional (non-simulated) replication study. For a simulation study of stochastic models, the numerical equivalence is expanded to distributional and relational equivalence, which is a more realistic and reasonable assessment depending on the particular goal of the individual simulation study.

5 CHALLENGES IN REPRODUCING OR REPLICATING A SIMULATION STUDY

Different domains and disciplines are often of distinct nature, thus resulting in models of different nature. They can be, e.g., with different levels of abstraction and details, and various stochastic characterization representing uncertainty [11]. Unlike in many physical sciences and engineering domains, systems that have less clear or agreed-upon “ground truth”, such as social systems or value systems, have many degrees of freedom in model conceptualization. See, e.g., models mentioned in [36] used for hardware development in industries such as automotive and healthcare versus models in [37] used for incorporating justice considerations in energy transitions. The conceptualization of latter types of systems, often termed as socio-technical systems [38], are often subject to disparate or even inconsistent interpretations. In such cases, simulation R&R studies by other researchers are particularly hard, even when the original code and data are available.

Many reported that, communications (sometimes extensive personal interactions) between the researchers and the original author(s) of the work are helpful for simulation R&R studies [11, 15, 24]. When such communications are not possible, e.g., when the (first) author(s) left the field of work, a R&R study has to be based on the original publication and the original code and data. Thus typical challenges reported in literature concern these two categories: (1) the reporting of the original simulation studies, and (2) the documentation and quality of corresponding models and associated data (if they are available). These challenges, briefly discussed in the next two paragraphs, impede the process and results of (computationally) reproducing or replicating a *simulation study*, should one wish to do so [8, 13, 15, 23]. Note that here we do not refer to contributing factors such as increased complexity, cognitive biases, publication biases, intellectual property rights, lack of incentives and funds, among others [1, 3, 39].

Many reproducibility challenges are caused by incomplete or ambiguous reporting and documentation of simulation studies [11, 15, 40, 41]. Sometimes not enough information could be obtained regarding the conceptual model, computational model and/or experimental conditions. Sometimes the referred documents in the original publication can not be retrieved. Studying and understanding the conceptual model is often the most significant step [15] where ambiguity in communicating a model and its experimental conditions can result in varying interpretations including that of assumptions, mathematical processes and mechanisms [11, 15]. When simulation models include stochastic processes, the experimental conditions are often not explicitly mentioned or lack sufficient details, which make numerical or distributional equivalence difficult to achieve, and the method’s strengths and limitations poorly understood [15, 42].

The availability of the original study’s source code and associated data is helpful for R&R [15, 41]. In many cases, however, open is not enough for reproducibility [30]. With accessibility, still many challenges can lay in model implementation software and data, sometimes even hardware, regarding their executability and comprehensibility [43, 44]. Software includes, e.g., simulation environments (a.k.a., platforms), libraries, toolkits and programming languages used for constructing a model implementation, and that for data management and statistics. The choices can influence how models can be represented and interact, possibly yielding different outcomes. Cross-platform and cross-language replications, portability and consistency of algorithms, different workflows, and performance constrains can be potential sources of significant variability between model results [17, 30, 44]. Sometimes, the published descriptions were explicit but have incorrect or inconsistent source code and/or data in relation to the descriptions [11, 15]. They can be caused by translation from conceptual model to model implementation (or vice versa), or alignment of the report to conceptual model or implementation [11].

6 DISCUSSIONS AND FUTURE WORK

Simulation is a useful research method empowered by computers. Unlike field experiments or laboratory experiments, simulation models generate results relying on computational routines. Thus their computational reproducibility is a basis for meaningful analysis, corroboration and further use of the results. Scientists and grant agencies have spent a large amount of time and funds on projects that develop new simulation models – while the efforts have been indispensable and fruitful, they often do not explicitly address computational reproducibility, despite the fact that many regard it as critical for scientific simulation [15, 19, 45].

At the same time, simulation models are becoming increasingly complex and more widely used. Recent developments in M&S also included more functionalities that use artificial intelligence, particularly machine learning. The domains of application have expanded beyond physical sciences and engineering to social and behavioural sciences, etc. In addition, those who use and develop simulation models are well-trained in their domains but not necessarily in the software aspect of computational methods. All these make simulation reproducibility a complex task to pursue and fulfil. As a community of model users and developers, we need to recognize that reproducibility is extremely challenging in practice, and that the steps needed to tackle it are not going to be acceptable for everyone [26, 28, 45].

The Open Science initiatives taking place in many parts of the world have brought positive changes in more transparent and accessible science. But for those who wish to engage in reproducibility of simulation models, openness alone is not sufficient. There is still a big gap to jump over towards reproducibility and possible reuse of simulation models. With the reproducible research movement, many authors have advocated for more structural and cultural change in institutions and research communities. For example, good institutional practice, appropriate incentive and evaluation systems, funders' policies, journal guidelines and standards, training programs, among many others [26, 45, 46]. Besides those, another crucial question we need to ask ourselves, which is not yet sufficiently addressed, is how to make reproducibility more *operational*, for both model builders and model replicators?

To create more reproducible models, and to reproduce the methods and results of others' models, researchers face related but different challenges, thus we need different types of skills and supports. For both cases, how to make the "extra" time and effort worthwhile and implementable for researchers who primarily focus on original research and already have high workloads? There are no simple solutions. We briefly discuss two initial thoughts that could work well with efforts in simulation models. The first is to tie reproducibility in M&S more closely to model reuse. Model reusability is challenging by itself, while reproducibility is a promising first step towards model reuse, which can be a good incentive for some researchers. It would be interesting to create resources, e.g., registers of reproducible simulation models, for domain specific corroboration and potential reuse. There is no lack of online code and model repositories and versioning systems, but how reproducible are the simulation models indexed therein is highly unclear.

The second is to develop processes, methods, and supports – benchmarks, guidelines, tooling, etc. – that could integrate well with existing model development (or reuse) cycles and workflows. While R&R is a socio-economic problem [22], it is also a methodological issue. Exsiting works of R&R in different disciplines often have their particular focus. For example, the ACM Conference on Reproducibility and Replicability (ACM REP, inaugurated in 2023) heavily focuses on computational issues (https://acm-rep.github.io/). But simulation has a particular focus on the imitation of systems changing over time. For that, workflow steps, software dependencies, data, the trajectory of state changes and results, explanations, all these should be captured in an iterative M&S-based research lifecycle [30]. Moreover, because the model conceptualization phase employs simplification and abstraction based on many assumptions given a certain goal of the simulation study, modellers shall be facilitated to capture such conceptual conditions in a more clarifying and methodological way. There are good works or recommendations towards this direction, e.g., metadata structures, model and experiment description languages, and reporting guidelines

[12, 30, 40, 43, 44, 47]. We also need a better understanding of the needs and workflows of researchers to design methods and supports for wider adoption of model reproducibility practices. Furthermore, being able to identify the type and level of complexity of simulation models could also be a good help to assess the time and efforts needed for reproducible simulation.

Reproducible simulation is a process rather than a destination, hence it shall not be treated as an end in itself or as an afterthought at the end of an scientific endeavour [1, 4, 30]. It calls for a way of working and thinking about how we can build and treat our scientific simulations. Some might think that we need openness, transparency and wide adoption of good practices to make simulation reproducible. While these are all true, this paper also argues that the how of the good practices in reproducible simulation is still largely unknown for both model developers and model users. We need implementable methods and supportive tools to bring reproducibility in computer simulation forward.

REFERENCES

[1] J. M. Alston and J. A. Rick, "A beginner's guide to conducting reproducible research," *The Bulletin of the Ecological Society of America*, vol. 102, no. 2, p. e01801, 2021. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/bes2.1801

[2] R. D. Peng, "Reproducible research in computational science," *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011.

[3] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, p. 452–454, 2016. [Online]. Available: https://doi.org/10.1038/533452a

[4] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis, "What does research reproducibility mean?" *Science Translational Medicine*, vol. 8, no. 341, pp. 341ps12–341ps12, 2016. [Online]. Available: https://www.science.org/doi/abs/10.1126/scitranslmed.aaf5027

[5] F. Fidler and J. Wilcox, "Reproducibility of Scientific Results," in *The Stanford Encyclopedia of Philosophy*, Summer 2021 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021.

[6] Special Collection, "Challenges in irreproducible research." Nature, 2018. [Online]. Available: https://www.nature.com/collections/prbfkwmwvz

[7] National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press, 2019. [Online]. Available: https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science

[8] S. M. Niaz Arifin, G. J. Davis, and Y. Zhou, "Verification & validation by docking: a case study of agent-based models of anopheles gambiae," in *Proceedings of the 2010 Summer Computer Simulation Conference*, ser. SCSC '10. San Diego, CA, USA: Society for Computer Simulation International, 2010, p. 236–243.

[9] R. Axtell, R. Axelrod, J. M. Epstein, and M. D. Cohen, "Aligning simulation models: A case study and results," *Computational & Mathematical Organization Theory*, vol. 1, no. 2, pp. 123–141, 1996. [Online]. Available: https://doi.org/10.1007/BF01299065

[10] S. J. E. Taylor, A. Khan, K. L. Morse, A. Tolk, L. Yilmaz, J. Zander, and P. J. Mosterman, "Grand challenges for modeling and simulation: simulation everywhere—from cyberinfrastructure to clouds to citizens," *SIMULATION*, vol. 91, no. 7, pp. 648–665, 2015. [Online]. Available: https://doi.org/10.1177/0037549715590594

[11] B. G. Fitzpatrick, "Issues in reproducible simulation research," *Bulletin of mathematical biology*, vol. 6, no. 81, pp. 1–6, 2019.

[12] L. Yilmaz and T. Ören, *Toward Replicability-Aware Modeling and Simulation: Changing the Conduct of M&S in the Information Age*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 207–226. [Online]. Available: https://doi.org/10.1007/978-3-642-31140-6_11

[13] R. Axelrod, "Advancing the art of simulation in the social sciences," *Japanese Journal for Management Information System*, vol. 12, no. 3, 2003, special Issue on Agent-Based Modeling.

[14] M. S. Jalali, C. DiGennaro, A. Guitar, K. Lew, and H. Rahmandad, "Evolution and reproducibility of simulation modeling in epidemiology and health policy over half a century," *Epidemiologic Reviews*, vol. 43, no. 1, pp. 166–175, 09 2021. [Online]. Available: https://doi.org/10.1093/epirev/mxab006

[15] J. Zhang and D. T. Robinson, "Replication of an agent-based model using the replication standard," *Environmental Modelling & Software*, vol. 139, p. 105016, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815221000591

[16] K. Bajracharya and R. Duboz, "Comparison of three agent-based platforms on the basis of a simple epidemiological model," in *Proceedings of the Symposium on Theory of Modeling & Simulation - DEVS Integrative M&S Symposium*, ser. DEVS 13. San Diego, CA, USA: Society for Computer Simulation International, 2013.

[17] E. Donkin, P. Dennis, A. Ustalakov, J. Warren, and A. Clare, "Replicating complex agent based models, a formidable task," *Environmental Modelling & Software*, vol. 92, pp. 142–151, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815216310088

[18] B. G. Fitzpatrick, D. M. Gorman, and C. Trombatore, "Impact of redefining statistical significance on p-hacking and false positive rates: An agent-based model," *PLOS ONE*, vol. 19, no. 5, pp. 1–18, 05 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0303262

[19] K. Luijken, A. Lohmann, U. Alter, J. Claramunt Gonzalez, F. J. Clouth, J. L. Fossum, L. Hesen, A. H. J. Huizing, J. Ketelaar, A. K. Montoya, L. Nab, R. C. C. Nijman, B. B. L. Penning de Vries, T. D. Tibbe, Y. A. Wang, and R. H. H. Groenwold, "Replicability of simulation studies for the investigation of statistical methods: the RepliSims project," *Royal Society Open Science*, vol. 11, no. 1, p. 231003, 2024. [Online]. Available: https://doi.org/10.1098/rsos.231003

[20] O. E. Gundersen, "The fundamental principles of reproducibility," *Philosoplycal Transactions of the Royal Society A*, vol. 379, no. 20200210, 2021. [Online]. Available: http://doi.org/10.1098/rsta.2020.0210

[21] A. M. Law, *Simulation Modeling and Analysis*, 4th ed. McGraw-Hill, 2007.

[22] K. Hinsen, "Reproducibility and replicability of computer simulations," Keynote ACM REP'24 in Rennes, France., 2024. [Online]. Available: https://hal.science/hal-04621140/

[23] B. Edmonds and D. Hales, "Replication, replication and replication: Some hard lessons from model alignment," *Jasss-The Journal Of Artificial Societies And Social Simulation*, vol. 6, pp. U227–U253, 2003.

[24] U. Wilensky and W. Rand, "Making models match: Replicating an agent-based model," *Journal of Artificial Societies and Social Simulation*, vol. 10, no. 4, p. 2, 2007. [Online]. Available: https://www.jasss.org/10/4/2.html

[25] B. Edmonds and D. Hales, "Computational simulation as theoretical experiment," *The Journal of Mathematical Sociology*, vol. 29, no. 3, pp. 209–232, 2005. [Online]. Available: https://doi.org/10.1080/00222500590921283

[26] S. J. E. Taylor, T. Eldabi, T. Monks, M. Rabe, and A. M. Uhrmacher, "Crisis, what crisis – does reproducibility in modeling & simulation really matter?" in *2018 Winter Simulation Conference (WSC)*, 2018, pp. 749–762.

[27] C. Drummond, "Reproducible research: a minority opinion," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 30, no. 1, p. 1 – 11, 2018, cited by: 12; All Open Access, Bronze Open Access. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85038128735&doi=10.1080%2f0952813X.2017.1413140&partnerID=40&md5=d866922999a02862c1184e29541da6ff

[28] ——, "Is the drive for reproducible science having a detrimental effect on what is published?" *Learned Publishing*, vol. 32, no. 1, pp. 63–69, 2019. [Online]. Available:

https://onlinelibrary.wiley.com/doi/abs/10.1002/leap.1224

[29] D. Fanelli, "Is science really facing a reproducibility crisis, and do we need it to?" *Proceedings of the National Academy of Sciences (PNAS)*, vol. 115, no. 11, pp. 2628–2631, 2018. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1708272114

[30] X. Chen, S. Dallmeier-Tiessen, R. Dasler, S. Feger, and et al., "Open is not enough," *Nature Physics*, vol. 15, no. 2, pp. 113–119, 2019. [Online]. Available: https://doi.org/10.1038/s41567-018-0342-2

[31] A. Lohmann, O. L. O. Astivia, T. P. Morris, and R. H. H. Groenwold, "It's time! ten reasons to start replicating simulation studies," *Frontiers in Epidemiology*, vol. 2, 2022. [Online]. Available: https://www.frontiersin.org/journals/epidemiology/articles/10.3389/fepid.2022.973470

[32] Y. Huang and I. Nikolic, "Towards a multi-model infrastructure for integrated decision-making in energy transition," Sep. 2024, international Multidisciplinary Modeling & Simulation Multiconference : I3M 2024 ; Conference date: 18-09-2024 Through 20-09-2024. [Online]. Available: https://www.msc-les.org/i3m2024/

[33] S. C. Brailsford, T. Eldabi, M. Kunc, N. Mustafee, and A. F. Osorio, "Hybrid simulation modelling in operational research: A state-of-the-art review," *European Journal of Operational Research*, vol. 278, no. 3, pp. 721–737, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221718308786

[34] R. Heyard, S. Pawel, J. Frese, B. Voelkl, H. Würbel, S. K. McCann, L. Held, K. E. Wever, H. Hartmann, L. Townsin *et al.*, "A scoping review on metrics to quantify reproducibility: a multitude of questions leads to a multitude of metrics," *MetaArXiv*, 2024, center for Open Science.

[35] J. Muradchanian, R. Hoekstra, H. Kiers, and D. van Ravenzwaaij, "How best to quantify replication success? a simulation study on the comparison of replication success metrics," *Royal Society Open Science*, vol. 8, no. 5, p. 201697, 2021. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsos.201697

[36] S. L. Shrestha, S. A. Chowdhury, and C. Csallner, "Replicability study: Corpora for understanding simulink models & projects," in *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2023, Conference Proceedings, pp. 1–12.

[37] A. Sundaram, Y. Huang, E. Cuppen, and I. Nikolic, "Operationalizing justice in models used as decision-support tools in local and regional energy transition planning," 2024, 12th International Workshop on Simulation for Energy, Sustainable Development & Environment : SESDE 2024 ; Conference date: 18-09-2024 Through 20-09-2024. [Online]. Available: https://www.cal-tek.eu/proceedings/i3m/2024/sesde/

[38] Y. Huang, G. Poderi, S. Šćepanović, H. Hasselqvist, M. Warnier, and F. Brazier, *Embedding Internet-of-Things in Large-Scale Socio-technical Systems: A Community-Oriented Design in Future Smart Grids*. Springer International, 2019, pp. 125–150. [Online]. Available: https://doi.org/10.1007/978-3-319-96550-5_6

[39] M. Munafò, B. Nosek, D. Bishop, K. Button, C. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. Ware, and J. Ioannidis, "A manifesto for reproducible science," *Nature Human Behaviour*, vol. 1, p. 0021, 01 2017.

[40] T. Monks, C. S. M. Currie, B. S. Onggo, S. Robinson, M. Kunc, and S. J. E. Taylor, "Strengthening the reporting of empirical simulation studies: Introducing the stress guidelines," *Journal of Simulation*, vol. 13, no. 1, pp. 55–67, 2019. [Online]. Available: https://doi.org/10.1080/17477778.2018.1442155

[41] J. Navarro, A. Deruyver, and P. Parrend, "A systematic survey on multi-step attack detection," *Computers & Security*, vol. 76, pp. 214–249, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404818302141

[42] C. Williams, Y. Yang, M. Lagisz, K. Morrison, L. Ricolfi, D. I. Warton, and S. Nakagawa, "Transparent reporting items for simulation studies evaluating statistical methods: Foundations for reproducibility and reliability," *Methods in Ecology and Evolution*, vol. n/a, no. n/a, 2024. [Online].

Available: https://doi.org/10.1111/2041-210X.14415

[43] M. L. Blinov, J. H. Gennari, J. R. Karr, I. I. Moraru, D. P. Nickerson, and H. M. Sauro, "Practical resources for enhancing the reproducibility of mechanistic modeling in systems biology," *Current Opinion in Systems Biology*, vol. 27, p. 100350, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452310021000445

[44] B. Antunes and D. R. Hill, "Reproducibility, replicability and repeatability: A survey of reproducible research with a focus on high performance computing," *Computer Science Review*, vol. 53, p. 100655, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157401372400039X

[45] A. M. Uhrmacher, S. Brailsford, J. Liu, M. Rabe, and A. Tolk, "Panel—reproducible research in discrete event simulation—a must or rather a maybe?" in *2016 Winter Simulation Conference (WSC)*. IEEE, 2016, pp. 1301–1315.

[46] C. Begley, A. Buchan, and U. Dirnagl, "Robust research: Institutions must do their part for reproducibility," *Nature*, vol. 525, pp. 25–7, 09 2015.

[47] V. Grimm, S. F. Railsback, C. E. Vincenot, U. Berger, C. Gallagher, D. L. DeAngelis, B. Edmonds, J. Ge, J. Giske, J. Groeneveld, A. S. Johnston, A. Milles, J. Nabe-Nielsen, J. G. Polhill, V. Radchuk, M.-S. Rohwäder, R. A. Stillman, J. C. Thiele, and D. Ayllón, "The odd protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism," *Journal of Artificial Societies and Social Simulation*, vol. 23, no. 2, p. 7, 2020. [Online]. Available: http://jasss.soc.surrey.ac.uk/23/2/7.html

AUTHOR BIOGRAPHIES

YILIN HUANG is Assistant Professor at Faculty of Technology, Policy and Management at Delft University of Technology in the Netherlands. She received her PhD on the topic of Automated Simulation Model Generation from the same university. Her research mainly concerns Modelling and Simulation (M&S) theory and methodology, with particular interests in data-driven methods, model conceptualization, and reproducibility of simulation models. She has worked on national and international simulation projects of large-scale complex socio-technical systems in application domains such as transportation, logistics, smart energy systems, and sustainability transitions. Her email address is y.huang@tudelft.nl.