PHILOSOPHICAL TRANSACTIONS A

Phil. Trans. R. Soc. A.
doi:10.1098/not yet assigned

Reliability and reproducibility in computational science: implementing validation, verification and uncertainty quantification *in silico*

Peter V. Coveney[1*], Derek Groen[2] and Alfons G. Hoekstra[3]

[1]*Centre for Computational Science, University College London, Gordon Street, London, UK*
[2]*Department of Computer Science, Brunel University London, UK*
[3]*Institute for Informatics, Science Park 904, University of Amsterdam, 1098 XH Amsterdam, Netherlands*

PVC: https://orcid.org/0000-0002-8787-7256
DG: https://orcid.org/0000-0001-7463-3765
AGH: https://orcid.org/0000-0002-3955-2449

Keywords: computational science, computer science, reproducibility, validation, verification, uncertainty quantification

---

The objectivity of science is its crowning and distinguishing feature. Its stock-in-trade are experimental facts, observations and theories which do not depend on who reports them but rather on the notion that the same findings would be obtained by anyone else performing similar procedures. This is what is meant by scientific reproducibility. That, at least, is the aspiration. In practice, things are often less clear cut. Theories are built on observations and experimental data. They involve a process of logical thought based on mathematical methods. These can sometimes be found to be in error because the empirical data they depend on are subsequently shown to be wrong; or there were errors in the logical development of the theory propounded. And then there are the many differences between a given experimental set up and others used to measure the same thing, so error and uncertainty are inevitable.

In the modern era of science, computers have come to play a central role. Computer simulation is a way of extracting useful information from theories and the models built using them. Such models are typically impossible to analyse without computers. They produce results which may be designed for comparison against existing experimental measurements; but they are also capable of making predictions for which no experimental data are available. Owing to the sophistication of modern science, such calculations often require powerful computers if results of any kind are to be forthcoming. And for situations in which it is thought that the theories and models are sufficiently accurate, one would like to use computer-based simulation in order to make *actionable* predictions—predictions whose credibility is sufficiently great that we can use them to make

*Author for correspondence: p.v.coveney@ucl.ac.uk.
†Centre for Computational Science, University College London, Gordon Street, London WC1H 0AJ, UK

important decisions. Examples of actionable predictions occur in weather forecasts, environmental disasters, climate science, the design of advanced materials, drug discovery and clinical decision making. Are the methods we use today of sufficient reliability that they can generate actionable results?

That question is what this theme issue is about. Three notions inform the assessment of such reliability. First, validation: confirmation that the results are in agreement with experiment, the litmus test for whether a simulation is credible. Second, verification: that the software does what it is supposed to do, and does not contain any errors arising from an incorrect implementation or incorrect numerical methods. Third, uncertainty quantification: identification of the provenance of errors within the model, which stem from two distinct sources, one being systematic errors due to parameter estimation, the other arising from random errors which come from the use of random number generators in the code.

The purpose of the present theme issue is to survey the state of the art in this domain. The emphasis is on seeking to obtain reproducible scientific findings using computers, and to quantify the level of uncertainty in the codes and procedures used in contemporary scientific research. The issue contains a total of fourteen papers, ranging from research articles and opinion pieces to reviews of aspects of the domain. The diversity of the topics underlines it trans-disciplinary character. There are papers by computer and computational scientists, alongside or together with contributions from authors from established scientific domains.

While uncertainty quantification is a well-established field so far as engineers and applied mathematicians are concerned, it is relatively uncommon in other branches of science such as physics and chemistry, life and medical sciences. Within this theme issue, there are papers which look at single scale modelling and simulation methods such as the well-known molecular dynamics approach, while others are concerned with uncertainty quantification within a multiscale context, in which multiple different single-scale methods are connected in order to bridge spatial and temporal domains. This is currently a research frontier in uncertainty quantification; unlike single-scale approaches, which can be implemented in an unobtrusive manner, quantifying uncertainty within multiscale models frequently requires semi-intrusive or more fully intrusive modifications to the existing modelling code base.

The practicalities of reproducing results from computational studies are made evident in the work presented by Krafczyk *et al.*[1] [RSTA-2020-0069.R1]. They examine over three hundred computational studies and attempt to access the code used and recreate the results presented in the articles. Within their self-imposed time limits, they could not fully reproduce the results from any of these papers. In response to this, the authors propose a set of Reproducibility Principles and Guidelines to assist researchers in making their computational results reproducible. Complementing these, the authors outline the structure of a Reproduction Package as a set of documentation and files to allow a simulation code to be reproduced in a straightforward manner. They provide several vignettes describing their efforts to reproduce results from published articles in order to highlight issues that may be overlooked or neglected when reporting on computational research.

A prime example of good practice in both evaluating the performance of a numerical model and adherence to reproducibility standards is provided by Clementi and Barba[2] [RSTA-2020-0068.R1] in the field of nanoscale electrostatics. They commence by attempting to replicate the results for the resonance modes of silicon carbide obtained from two studies in the literature and extend this to validation of their model against experimental data presented in one of the studies. They successfully achieve replication up to fundamental differences between their modelling approach and those in the comparison papers; validation was also successful. The application of reproducibility packages provides readers with access to all the digital artefacts needed to create the results presented in the study - including source code, input files and post-processing scripts.

Numerical simulations have become a cornerstone of research in many fields of science and engineering. With this prevalence, ensuring the reproducibility of simulation studies is key to maintaining confidence in such work. Although many such models are deterministic, understanding the sensitivity of outputs to input variation is of central importance. Volodina and Challenor[3] [RSTA-2020-0071.R1] seek to overcome the computational expense of gaining such insight through multiple (i.e. ensemble) simulations by capturing the characteristics of a complex model with a cheaper Gaussian process emulator. Through demonstration of their methodology using a simple one-dimensional function and a climate model of cloud behaviour, they illustrate how uncertainty characteristics of complex deterministic models can be assessed and interpreted.

In "Towards validated multiscale simulations for fusion"[4] [RSTA-2020-0074.R1], Luk *et al.* apply uncertainty quantification to modelling nuclear fusion for energy production using coupled multiscale simulations. Time scale bridging requires standardised procedures to determine scale separation and the existence of a steady state within the fastest evolving model. In the case of nuclear fusion, the turbulence model associated with the plasma instabilities needs to reach a steady state. The authors discuss and compare existing and newly introduced time-scale bridging methods by means of sensitivity analysis. Furthermore, quantitative probabilistic metrics are used to assess the validity of the predictions of the multiscale model by comparison with experimental data using the Hellinger distance, Jensen-Shannon divergence, and Wasserstein metric.

Wan *et al.*[5] [RSTA-2020-0082.R2] discuss the quantification of uncertainty in simulations that are based on classical molecular dynamics. The paper addresses simulations in a wide range of applications from binding affinity calculations for drug discovery to properties prediction within condensed matter and materials. Valuable insights are provided concerning the intrinsic stochasticity of molecular systems due to their chaotic nature, whose resulting uncertainty is a dominant factor contributing to the uncertainty of individual trajectories. The authors show that ensemble methods provide statistically reliable results and that the distributions predicted from such simulations are often non-Gaussian in nature.

Suleimenova *et al.*[6] [RSTA-2020-0077.R1] investigate how human migration modelling depends on reliable handling of the many parameters which pervade such computer-based studies. They integrate the use of

sensitivity analysis into the development of new simulation rule sets. Based on an agent-based simulation of migration, they use Sobol's method for sensitivity analysis to identify the most sensitive assumptions. They refine the rule set with the aim of making these assumptions more detailed (e.g. through parameter splitting) and are able to reduce the sensitivity of these assumptions. Their development approach is potentially more robust than conventional ones as developers do not directly aim for error reduction but instead optimise for balanced sensitivity across the assumptions made within the models.

In "Uncertainty Quantification Patterns for Multiscale Models", Ye *et al.*[7] [RSTA-2020-0072.R2] present a conceptual framework of computing patterns that support the analysis of uncertainty in coupled models, irrespective of their source domain. The paper presents the basic templates for each uncertainty quantification pattern (UQP) and introduces the notion of semi-intrusive UQ, where sub-models are treated as black boxes but UQ algorithms are applied to the coupling between the individual sub-models. They showcase their implementation through two applications, each of which has been coupled using the Multiscale Coupling Library and Environment (MUSCLE3).

Daub *et al.*[8] [RSTA-2020-0076.R1] provide a form of tutorial which introduces a surrogate-model based uncertainty quantification approach applied to an earthquake rupture simulator. It uses the mogp_emulator package to perform model calibrations, combining it with the FabSim3 automation toolkit to automatically execute and curate the large number of surrogate model executions required on remote resources. The tutorial has been successfully performed in various workshops, is fully open, and can be readily undertaken by readers within a matter of hours.

Jansson *et.* al[9] [RSTA-2020-0073.R1] in "Assessing uncertainties from physical parameters and modelling choices in an atmospheric LES model" apply a range of modern UQ methods to investigate uncertainties in a large eddy simulation. To do this, they use the stochastic collocation scheme with the EasyVVUQ package to calculate the Sobol indices for a range of parameters. They identify uncertainties caused by small random initial state perturbations and find that the chosen advection scheme has a major influence on the resulting quantities of interest.

Coveney and Highfield[10] [RSTA-2020-0067.R1] discuss the importance of reproducibility in science and more particularly computational science. They assess where this can impact data analysis and simulation. The paper describes existing initiatives as well as new suggestions for ensuring increased trust in computer-based predictions. The authors draw attention to various limits in the applicability of computer simulation methods and raise concerns about the lack of transparency of many artificial intelligence methods which are often applied as "black boxes" to solve complex problems without a clear understanding of their inherent limitations.

Fursin presents a new framework (Collective Knowledge, or CK) to decompose projects into reusable components[11] [RSTA-2020-0211.R1]. Among other things, the approach facilitates the assembly of portable workflows, and helps to reproduce, compare, and reuse research techniques from existing publications. To showcase the added value and generality of CK, the author applies it to six exemplary use cases, many of which are directly informed by industrial needs.

In the paper entitled "VECMAtk: A Scalable Verification, Validation and Uncertainty Quantification toolkit for Scientific Simulations", Groen *et al.*[12] [RSTA-2020-0221.R2] introduce a toolkit that helps users to gain access to a wide range of methods by means of which to scrutinize and assess all kinds of scientific simulations. It facilitates the efficient and straightforward execution of substantial sensitivity analysis and uncertainty quantification (UQ) investigations using remote supercomputers. The authors present applications across six different scientific domains, each of which highlight different aspects of the toolkit in terms of advanced UQ algorithm support, code coupling with uncertainty taken into account, efficient execution of tens of thousands of ensemble-based simulation jobs, and the automated calculation of key sensitivity and uncertainty measures.

In "The case for free and open source software in research and scholarship", Fortunato and Galassi[13] [RSTA-2020-0079] explore the close relation between Free and Open Software (FOSS) and academia at large. They resolve a range of common misconceptions among academics about free software and open source software, and introduce a primer to FOSS suitable for researchers in any field. The paper includes a case study about the GNU Scientific Library project which demonstrates among other things how seemingly minor misconceptions about the openness of underlying libraries can give rise to existential and far-reaching problems in reproducing scientific results.

In his paper entitled "A Fundamental View on Reproducibility" Odd Erik Gundersen provides a survey of the literature on reproducibility and a clarification on its meaning in a computer science context[14] [RSTA-2020-0210.R1]. Through the use of the scientific method, Gundersen identifies four types of transparency that enable reproducible software and distinguishes between two types of reproducibility: output reproducible and analysis reproducible. Overall, transparency and openness are identified as key drivers for reproducibility, which in turn promotes more fast-paced and assured scientific progress.

Additional Information

Authors' Contributions
All authors contributed equally.

Competing Interests
The authors have no competing interest.

Funding Statement

Phil. Trans. R. Soc. A.

The authors are grateful for funding from the European Commission for the VECMA grant (number 800925) and from the Alan Turing Institute in London that enabled us to run an event in January 2020 under a similar title to this theme issue.

Acknowledgments
We are grateful to Dr Apostolos Evangelopoulos for his important support in the development of this theme issue, and with Dr Hugh Martin, for his contribution to organising the event at the Alan Turing Institute.

References

1. Krafczyk M, Shi A, Bhaskar A, Marinov D, Stodden V. Learning from Reproducing Computational Results: Three Reproducibility Principles and the Reproduction Package. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0069.R1.
2. Clementi N, Barba L. Reproducible Validation and Replication Studies in Nanoscale Physics. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0068.R1.
3. Volodina V, Challenor P. The importance of uncertainty quantification in model reproducibility. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0071.R1.
4. Luk O, Lakhlili J, Hoenen O, Scott B, Coster D. Towards validated multiscale simulations for fusion. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0074.R1.
5. Wan S, Sinclair R, Coveney P. Uncertainty Quantification in Classical Molecular Dynamics. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0082.R2.
6. Suleimenova D, Arabnejad H, Edeling W, Groen D. Sensitivity-driven simulation development: A case study in forced migration. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0077.R1.
7. Ye D, Veen L, Nikishova A, Edeling W, Luk O, Krzhizhanovskaya V, et al. Uncertainty Quantification Patterns for Multiscale Models. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0072.R2.
8. Daub E, Arabnejad H, Mahmood I, Groen D. Uncertainty Quantification of Dynamic Earthquake Rupture Simulations. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0076.R1.
9. Jansson F, Edeling W, Attema J. Assessing uncertainties from physical parameters and modelling choices in an atmospheric LES model. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0073.R1.
10. Coveney P V., Highfield RR. When We Can Trust Computers (and When We Can't). Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0067.R1.
11. Fursin G. Collective Knowledge: organizing research projects as a database of reusable components and portable workflows with common APIs. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0211.R1.
12. Groen D, Arabnejad H, Jancauskas V, Edeling W, Jansson F, Richardson R, et al. VECMAtk: A Scalable Verification, Validation and Uncertainty Quantification toolkit for Scientific Simulations. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0221.R2.
13. Fortunato L, Galassi M. The case for free and open source software in research and scholarship. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0079.

14. Gundersen OE. A Fundamental View on Reproducibility. Philos Trans R Soc A Math Phys Eng Sci. 2020;this issue:RSTA-2020-0210.R1.