SANDIA REPORT

SAND2010-6474
Unlimited Release
Printed September 2010

QMU as an Approach to Strengthening the Predictive Capabilities of Complex Models

Genetha A. Gray, Paul T. Boggs, and Matthew D. Grace

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

National Nuclear Security Administration

SAND2010-6474
Unlimited Release
Printed September 2010

QMU as an Approach to Strengthening the Predictive Capabilities of Complex Models

Genetha A. Gray
Quantitative Modeling & Analysis, Dept. 8954
Sandia National Laboratories
P.O. Box 969, MS9159 Livermore, CA 94551-0969
gagray@sandia.gov

Paul T. Boggs
Quantitative Modeling & Analysis, Dept. 8954
Sandia National Laboratories
P.O. Box 969, MS9159 Livermore, CA 94551-0969
ptboggs@sandia.gov

Matthew D. Grace
Scalable & Secure Systems Research, Dept. 8961
Sandia National Laboratories
P.O. Box 969, MS9158 Livermore, CA 94551-0969
mgrace@sandia.gov

Abstract

Complex systems are made up of multiple interdependent parts, and the behavior of the entire system cannot always be directly inferred from the behavior of the individual parts. They are nonlinear and system responses are not necessarily additive. Examples of complex systems include energy, cyber and telecommunication infrastructures, human and animal social structures, and biological structures such as cells. To meet the goals of infrastructure development, maintenance, and protection for cyber-related complex systems, novel modeling and simulation technology is needed. Sandia has shown success using M&S in the nuclear weapons (NW) program. However, complex systems represent a significant challenge and relative departure from the classical M&S exercises, and many of the scientific and mathematical M&S processes must be re-envisioned. Specifically, in the NW program, requirements and acceptable margins for performance, resilience, and security are well-defined and given quantitatively from the start. The Quantification of Margins and Uncertainties (QMU) process helps to assess whether or not these safety, reliability and performance requirements have been met after a system has been developed. In this sense, QMU is used as a sort of check that requirements have been met once the development process is completed. In contrast, performance requirements and margins may not have been defined a priori for many complex systems, (i.e. the Internet, electrical distribution grids, etc.), particularly not in quantitative terms. This project addresses this fundamental difference by investigating the use of QMU at the start of the design process for complex systems. Three major tasks were completed. First, the characteristics of the cyber infrastructure problem were collected and considered in the context of QMU-based tools. Second, UQ methodologies for the quantification of model discrepancies were considered in the context of statistical models of cyber activity. Third, Bayesian methods for optimal testing in the QMU framework were developed. This completion of this project represent an increased understanding of how to apply and use the QMU process as a means for improving model predictions of the behavior of complex systems.

Acknowledgments

The authors would like to thank and acknowledge the following people for their contributions to this project:

- Jim Ringland
- Youssef Marzouk
- John Clem

Contents

1 Motivation **7**

2 The Cyber Problem **9**

3 Model Discrepancy **13**

4 Optimal Testing Strategy **15**

5 Future Work **17**

References **18**

Chapter 1

Motivation

The Quantification of Margins and Uncertainties (QMU) supports risk-informed decision making, where risk refers to the likelihood of a system failing or not achieving performance requirements. To date, QMU has played an important role in the nuclear weapon life cycle for which formal requirements have been long established. In this project, we considered the role that QMU might play in the design of complex systems that will support our future national security infrastructure. Specifically, we focused on the cyber infrastructures where acceptable performance margins have yet to be defined. The cyber infrastructure is particularly complicated in that it does not fit the policy or budgetary constraints of traditional infrastructures (i.e. roads, water systems, electrical grid); one cannot rely on "guns, guards, and gates" for security.

One of the main tools of the QMU process is uncertainty quantification (UQ) which provides the statistical rigor in the analysis of the performance requirements, the model predictions, and the comparison between the two. Expanding the UQ methodologies developed in the NW program is nontrivial as the foundational components of the NW program are based on robust physical equations. In contrast, the cyber world is based on observed data, heuristics, and anecdotal evidence. Evaluating cyber security includes assessing probabilities of events that have never been seen before and cannot necessarily be predicted using physical laws.

Chapter 2

The Cyber Problem

The cyber world is expanding quickly. Communication, entertainment, transportation, shopping, and medicine are just a few of the regular activities of daily life that rely on information that is stored on computers attached to the internet. Individuals, businesses, and the government all rely on cyber security measures to protect information and financial transactions and to prevent, detect, and respond to attacks. However, analyzing the robustness of a cyber system and taking protective action against threats is immensely complicated. There are a wide variety of attack scenarios, and even for recognized threats, controlled experiments are typically infeasible. Cyber networks and sensors collect an enormous amount of data, and this data must be efficiently and effectively aggregated. Decisions regarding security may have to be made quickly or with incomplete information, or both. In order to develop appropriate UQ/QMU methods for strengthening the predictive capabilities of cyber-related models, it is important to understand the basic characteristics of the underlying security problem. Therefore, one aspect of this project included gathering the information, both internally and externally, needed to best characterize the problem. This section summarizes the lessons learned.

Security can be defined by the following five basic issues:

1. **Confidentiality**: Assurance that information is never revealed to parties who aren't supposed to see it
2. **Integrity**: Ensuring that data are not altered in an unauthorized way during transmission
3. **Authentication**: Confirmation that communication is indeed with the expected party
4. **Non-Repudiation**: Prevention of false denial of actions or false commitment of illegal actions from others
5. **Availability**: Guarantee of network availability when required

Therefore, cyber infrastructures can be assessed in terms of these five issues. QMU methods can be applied to models to assess the degree to which each of these five issues is met. For example, it may be the case that QMU can be used to give a statistical distribution of integrity of a system. Then, the decision maker can use this information to determine whether or not system changes need to be made in order to meet stricter integrity requirements. Moreover, the QMU-based methods can elucidate the best case scenario in terms of security expectations.

Methods of defense are evaluated using the following five principles:

1. **Level of security**
2. **Functionality**

3. **Methods of operation**
4. **Ease of implementation**
5. **Performance**

In the QMU methodology, the results of these evaluations could be used to illustrate the effect of a method on an issue. For example, the method of operation of a cyber defense tool may have a profound effect on the availability of a system. And, the availability resulting from the method of operation of a given defense may be unacceptable in terms of performance requirements. If this is the case, the defense tool must be amended.

The nation's R&D program in the area of cyber is becoming increasingly "forward thinking" and is no longer just reactive, *i.e.*, just focused on fixing problems. Two current issues of concern are

- The complexity of cyber systems is counter to current security practices which were developed for much simpler systems.
- When considering data, security is often viewed as equivalent to privacy. Privacy is an expectation of the individual, but in fact is not the totality of security.

Moreover, the biggest weaknesses in most cyber-security systems are the system users. This is a result of lack of education, carelessness, and misinformation, not of malicious intent. Therefore, security must be multi-disciplinary and include behavioral ideals. To respond to some of these issues, Deloitte created the cyber threat intelligence framework [9]. The goal of this framework is to assess the existing capabilities of a legacy organization by collecting and analyzing relevant data regarding internal threats, external threats, and proactive surveillance. The main challenge is using the data collected to make better decisions and develop actionable security plans. Such a data set is ripe for the application of UQ methods. For example, Bayesian statistical emulation, wherein data sets are used to train a statistical model and the model is used to draw inferences, is one approach that allows full estimation of uncertainty.

As natural and man-made systems become increasingly interconnected, instrumented, and "smart," new security considerations are emerging. These include:

- Geographical dispersion of sensors and clients remote from data centers are more susceptible to attacks and disruptions.
- Heterogeneous technologies and proprietary protocols between sensors and devices are not always standardized and secure.
- Endpoints are often built in embedded systems with non-traditional operating systems where normal security functions may not exist.
- Cost-conscious endpoint vendors often cut corners at the expense of security.

These issues are indicative of the fast-evolving cyber infrastructure. Appropriate QMU methods must be general enough to "keep up." Business processes must be a part of the infrastructure models and security analytics must become a part of the decision making process.

Finally, it should be noted that the application of the cyber system may affect its security. For example, retail stores are worried about brand damage if their customers' credit card numbers are stolen. On the other hand, hospitals do not need to worry about brand damage and instead focus on public welfare and their ability to respond in an emergency. These difference should be fleshed out in the security concerns and should appear as an end result in the performance requirements of a QMU methodology.

Chapter 3

Model Discrepancy

Model discrepancy measures the error between the actual system and the assumed statistical model. It reflects the uncertainty that cannot be resolved merely by varying the model parameters and represents the missing or structurally deficient aspects of the model. In fact, model discrepancy is the dominant source of error and uncertainty in statistical models, not measurement error. Since the cyber infrastructure will be studied using statistical models, it is important to understand how to quantify model discrepancy as part of the UQ/QMU process. Therefore, we investigated Bayesian approaches to model discrepancy that were successful for applications with a high level of uncertainty in future projections. These applications included climate, health, systems biology, and infectious diseases. In this section, we review the interesting characteristics of these problems and mention the Bayesian methodologies applied to handle uncertainties.

Health Economics: Health economics models are used to make decisions about the allocation of health-care resources. The cost and efficacy of a treatment are combined into one measure called the net-benefit of the treatment. However, since there is not enough data to strictly observe the net-benefit, economic models are constructed. There are uncertainties in the model inputs and thus the corresponding uncertainties in the model output must be communicated to decision makers. To study the discrepancy, the model is decomposed into a series of linked submodels. Then, the discrepancy of each submodel is considered. Then, variance-based sensitivity analysis is applied to order the model inputs with respect to their contribution to the overall uncertainty. Moreover, sensitivity analysis is focused at the "boundaries of difficult choices." This work is described in detail in [22].

Climate Change: The uncertainties investigated here include emissions projections utilized in the climate change model and the actual modeling uncertainty that is the result of an incomplete understanding of the climate processes. In this study, twelve different models of the same events resulted in twelve different results. To take advantage of all the models, errors common to all models were compared to errors found in single models. The results gave decision makers three different scenarios for climate change based on emissions control. Current work in this area is striving for "seamless" predictions which join the approaches for short and long term predictions. Results of this work can be found in [14, 21, 20].

Systems Biology: Some work in this area studies cellular processes. The difficulties of this problem arise from the stochastic nature of these processes. Many cellular processes are dependent on the amount of proteins present and this value changes when random reactions occur. The uncertainties in models of these biological processes include noise in the biological system, complex stochastic measurements, cell-to-cell variability, uncertain initial conditions, and model inadequacy. To address these issues, researchers relay on Bayesian methods such as likelihood free Markov chain Monte Carlo [10], which updates one observation at a time. Complete information about the use of Bayesian models in biology can be found in [24].

Infectous Diseases: The modeling of infectious disease has shifted from deterministic, ordinary-differential-equation-based, compartmental models to complex individual-based stochastic models. Although these

models are more suitable, they require robust calibration techniques. Work is underway to study to efficacy of combining emulation and approximate Bayesian computation (ABC). ABC techniques are applied to summary data (*i.e.*, population mean) to make broad inferences. They reduce the overall computational time and are especially useful when evaluating likelihood is computationally prohibitive or impossible. ABC methods are primarily used in genetics, but are receiving increased attention in the analysis of complex and stochastic dynamical systems [18, 1]. By combining them with full statistical emulation, experts seek to both decrease computational requirements and increase the predictive value of models. These ideas are being studied at the London School of Hygiene & Tropical Medicine.

Many of the problem characteristics found in these problems also can be translated to issues in the cyber world. For example, just as in health economics, decisions need to be made which balance the effectiveness and cost of defense strategies. Similar to climate change projections, scenarios to control cyber crime must rely on other uncertain projections such as network usage. Moreover, there are often multiple statistical models of cyber activity. Cyber processes share stochasticity with systems biology as many aspects of activity are also random. Thus, Bayesian techniques which eliminate the likelihoods may play an important role in cyber model analysis. The use of ABC methods is attractive for cyber as much data exists to describe cyber traffic patterns. However, full Bayesian emulation is attractive to reduce uncertainty. Thus, as in infectious disease studies, some hybrid of the two is likely most appropriate. It is also interesting to note that the drive to eliminate disease in humans is not unlike the drive to eliminate computer viruses. Eliminating a disease does not necessarily require curing all the current infections. For example, the polio vaccine eradicated the disease without curing it. This is the same with computer viruses where good virus software can eliminate the negative effects of identified computer viruses.

It should also be noted that the classic hierarchy tree that serves as the system-of-systems model discussed in the next section was also explored with an eye toward model discrepancy. In that case, a variable (η) is included to represent the "unmeasurable" error of a node. This can also be interpreted as the discrepancy or the error in the model.

As shown above, there is a lot of exciting work being done in the area of UQ for complex systems. Because the characteristics of the problems studied in this project are similar to those of cyber systems, the next obvious step is to apply the Bayesian techniques for model discrepancy discussed here to the models of cyber activity being developed at Sandia. The results will can be used as part of an overall QMU process to help inform system designers and decision makers of the bounds of system performance. This will in turn assist in the process of putting reasonable system requirements in place.

Chapter 4

Optimal Testing Strategy

Bayesian techniques, along with a modern probability approach, have also been used to develop techniques for estimating the reliability of complex, hierarchical systems. The result is a statistical distribution of the reliability rather than a single number. This means that we can answer many questions concerning the confidence we have in the estimate; it also provides a way to specify optimal points for subsequent tests [3]. Because of the obvious applicability of this approach to improved model predictability, additional studies were completed.

Consider a classic hierarchy tree as a basic model of a "system-of-systems" network, where each node represents a component system (which may itself consist of a set of sub-systems). For this general composite system, we developed a technique for computing the optimal testing strategy, which is based on Bayesian decision analysis [2]. In previous work, we developed a Bayesian approach for computing the distribution of the reliability of a system-of-systems structure that uses test data and prior information. This allows for the determination of both an estimate of the reliability and a quantification of confidence in the estimate. Improving the accuracy of the reliability estimate and increasing the corresponding confidence require the collection of additional data. However, testing all possible sub-systems may not be cost-effective, feasible, or even *necessary* to achieve an improvement in the reliability estimate. To address this sampling issue, we formulate a Bayesian methodology that systematically determines the optimal sampling strategy under specified constraints and costs that will maximally improve the reliability estimate of the composite system, e.g., by reducing the variance of the reliability distribution. This methodology involves calculating the "Bayes risk of a decision rule" for each available sampling strategy, where risk quantifies the relative effect that each sampling strategy could have on the reliability estimate. A general numerical algorithm is developed and tested using an example multicomponent system. The results show that the procedure scales linearly with the number of components available for testing. Complete details of this work can be found in [4].

We accomplish this by defining a suitable loss function, which, given a particular action (e.g., "accept" or "reject") and state of nature, quantifies the cost or loss associated with that action, e.g., the cost of accepting a device although it is not performing within specifications. Next, we introduce a decision rule that, given additional data, selects the action that minimizes the expected loss. The loss function and the decision rule are combined in the Bayes risk of a decision rule, which then allows us to calculate relative risk for each possible sampling strategy. In this scenario, risk quantifies the relative effect that each sampling strategy could have on the reliability estimate.

Chapter 5

Future Work

As a result of this work, we identified the need for additional basic research in the area of UQ for cyber security problems. In particular, time must be taken to develop a mathematically sound foundation for the UQ required to best understand the nature of cyber threats and probabilistically define various anomalous events. Therefore, we prepared a regular FY11 LDRD proposal for HSD which focused on informing decisions related to cyber security through the use of UQ. Although this project was not funded, we received encouraging feedback. This prompted us to resubmit the idea to the newly formed Cyber Investment Area. In the resubmission, we narrowed our focus to supervisory control and data acquisition cyber security as it is used for the smart electrical grid.

In addition, we are working with staff member from 6000 to apply these analysis techniques to multi-dimensional threat security systems. The goal is to identify a model which could be used to illustrate the effectiveness of the QMU-based techniques discussed in this paper.

References

[1] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[2] James O. Berger. *Statistical decision theory and Bayesian analysis*. Springer series in statistics. Springer, New York, NY, 2nd edition, 1985.

[3] Paul T. Boggs et al. Practical reliability and uncertainty quantification in complex systems: Final report. Technical Report 2009-6072, Sandia National Laboratories,, 2009.

[4] Paul T. Boggs, Matthew D. Grace, Philippe P. Pébay, and James T. Ringland. Determining the Bayesian optimal sampling strategy in a hierarchical system. Technical report, Sandia National Laboratories, 2010.

[5] Paul T. Boggs, Youssef M. Marzouk, Philippe P. Pébay, John Red-Horse, James T. Ringland, Kathleen Diegert, and Rena Zurn. Bayesian methods for estimating the reliability in complex hierarchical networks (interim report). Technical Report SAND2007-2758, Sandia National Laboratories, 2007.

[6] Paul T. Boggs, Youssef M. Marzouk, James T. Ringland, Philippe P. Pébay Matthew D. Grace, John Red-Horse, James T. Ringland, Kathleen Diegert, and Rena Zurn. Determining the reliability distribution in a time-dependent hierarchical system. Technical report, Sandia National Laboratories,, 2009.

[7] R. H. Cameron and W. T. Martin. The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals. *Ann. Math.*, 48(2):385–392, 1947.

[8] Thierry Crestaux, Olivier Le Maître, and Jean-Marc Martinez. Polynomial chaos expansion for sensitivity analysis. *Rel. Eng. & Sys. Safety*, 94(7):1161–1172, 2009.

[9] T. DeZabala and R. Baich. Cyber crime: a clear and present danger. Technical report, Deloitte, 2010.

[10] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman & Hall/CRC Press, New York, NY, 1996.

[11] F. James. Monte Carlo theory and practice. *Rep. Prog. Phys.*, 43(9):1145–1189, 1980.

[12] Youssef M. Marzouk and Habib N. Najm. Dimensionality reduction and polynomial chaos acceleration of bayesian inference in inverse problems. *J. Comp. Phys.*, 228(6):1862–1902, 2009.

[13] Youssef M. Marzouk, Habib N. Najm, and Larry A. Rahn. Stochastic spectral methods for efficient Bayesian solution of inverse problems. *J. Comp. Phys.*, 224(2):560–586, Jun 20007.

[14] James Murphy, David Sexton, et al. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430:768–772, Aug 2004.

[15] P. Pébay, M. J. Rojas, and D. Thompson. Optimization and np_r-completeness of certain fewnomials. In *3rd International Workshop on Symbolic-Numeric Computation*, Kyoto, Japan, Jun 2009.

[16] P. Pébay, D. Thompson, and J. M. Rojas. Computational algebraic geometry for statistical modeling FY09Q2 progress. Technical report, Sandia National Laboratories, Apr 2009.

[17] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical recipes: The art of scientific computing*. Cambridge University Press, Cambridge, UK, 2nd edition, 1992.

[18] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. T. Feldman. Population growth of human y chromosomes: A study of Y chromosome microsatellites. *Mol. Biol. Evol.*, 16(12):1791–1798, 1999.

[19] P. K. Sarkar and M. A. Prasad. A comparative study of pseudo and quasi random sequences for the solution of integral equations. *J. Comp. Phys.*, 68(1):66–88, 1987.

[20] D. M. Smith, S. Cusack, A. W. Colman, et al. Improved surface temperature prediction for the coming decade from a global climate model. *Science*, 317(5839):796–799, 2007.

[21] D. A. Stainforth, T. Aina, et al. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, 433:403–406, Jan 2005.

[22] M. Strong, H. Pilgrim, J. Oakley, and J. Chilcott. Structural uncertainty in health economic decision models. Technical Report ISBN 1 900752 27, ScHARR (University of Sheffield), 2009.

[23] N. Wiener. The homogeneous chaos. *Amer. J. Math.*, 60:897–936, 1938.

[24] Darren J. Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC, Cambridge, UK, 2006.

Sandia National Laboratories