Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

DATA FARMING IN SUPPORT OF NATO OPERATIONS - METHODOLOGY AND PROOF-OF-CONCEPT

Gary Horne

MCR Federal Systems
901 North Stuart Street, Suite 603
Arlington, VA 22203, USA

Stephan Seichter

Bundeswehr Planning Office
Einsteinstr. 20
Ottobrunn, BY 85521, GERMANY

ABSTRACT

Data Farming is a process that has been developed to support decision-makers by answering questions that are not currently addressed. It uses an inter-disciplinary approach that includes modeling and simulation, high performance computing and statistical analysis to examine questions of interest with large number of alternatives. Data Farming allows for the examination of uncertain events with numerous possible outcomes and provides the capability of executing enough experiments so that both overall and unexpected results may be captured and examined for insights. In 2010, the NATO Science and Technology Organization started the three-year Task Group "Data Farming in Support of NATO" to assess and document the data farming methodology to be used for decision support. Two case studies were performed as proof-of-concept explorations to demonstrate the power of Data Farming. The paper describes the Data Farming methodology as an iterative process and summarizes the results of the case studies.

1 INTRODUCTION

Data Farming is an iterative and inter-disciplinary process that has been developed in the context of simulation-based support for decision-makers with the goal of answering questions that are not currently addressed. The Data Farming concept combined the six realms of Data Farming – Model Development, Rapid Scenario Prototyping, Design of Experiments, High Performance Computing, Analysis and Visualization and Collaboration - to examine questions of interest with large number of alternatives. Harnessing the power of data farming to apply it to our questions is essential to providing support not currently available to NATO decision-makers. This support is critically needed in answering questions inherent in the scenarios we expect to confront in the future as the challenges our forces face become more complex and uncertain.

Data farming is an iterative team process. Figure 1 shows the iterative process as a loop of loops with five of the six realms of data farming depicted. The sixth, collaboration, underlies the entire process and emphasizes the importance of the team aspect of data farming.

Since the term was coined in Horne (1997), the essence of data farming is that it is first and foremost a question-based approach. The basic question repeatedly asked in different forms and in different contexts is: What if? Data farming enables a refinement of questions as well as obtaining answers and insight into the questions. From 1998 to 2006, data farming developed along with a project funded by the US called Project Albert which quickly grew into an international effort and where the iterative nature of data farming was documented over these years (Horne 1999, Horne and Leonardi 2001, Horne and Meyer 2004, Horne and Meyer 2005). Development of data farming continued after Project Albert officially ended through sponsored work using the methods and the Data Farming Community has met for

workshops that continue to be held about twice a year as can be seen, for example, in the proceedings in Meyer and Horne (2007, 2013).

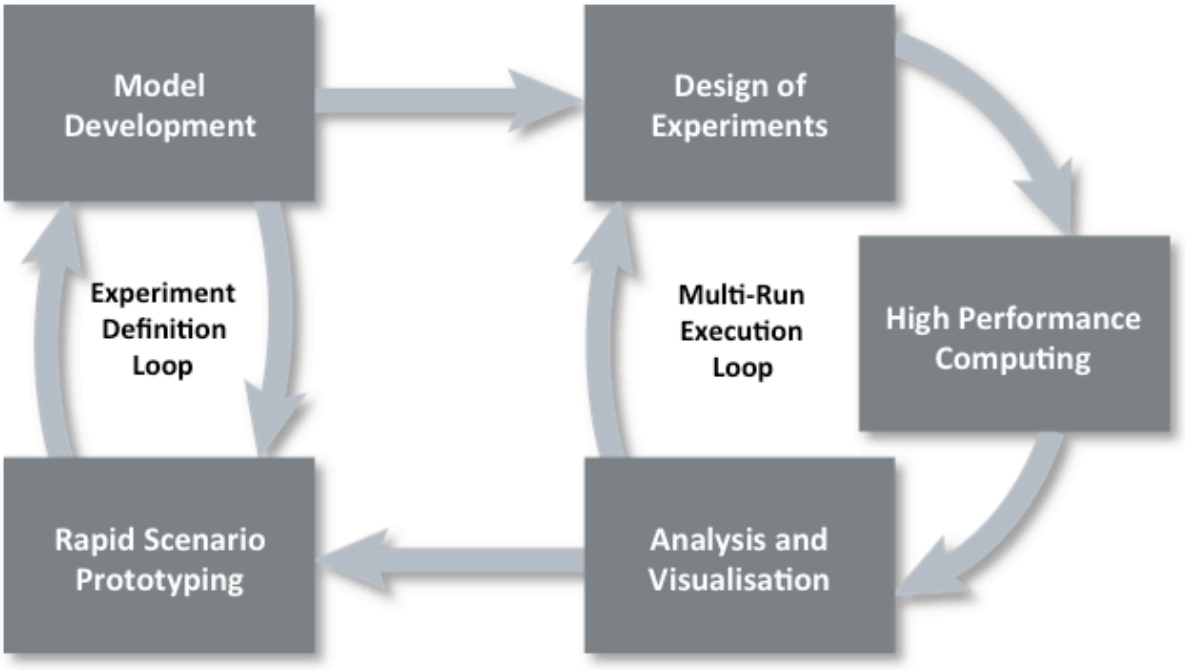

Figure 1. Data farming "loop of loops."

In 2010, the NATO Science and Technology Organization started the three-year Modeling and Simulation Task Group "Data Farming in Support of NATO" to assess and document the data farming methodology to be used for decision support (Horne 2010). This paper summarizes the completed work of this task group, designated MSG-088 (Horne et al. 2013). In the next 8 sections we will use Horne et al. (2013) to describe the six realms of data farming and the two case studies performed during the course of MSG-088.

2 RAPID SCENARIO PROTOTYPING

The model development and the rapid prototyping realms together make up the experiment definition loop in Figure 1. As such, they work hand-in-hand with each other and we could choose either realm to begin our detailed description of data farming. Thus the rapid scenario prototyping process is a good place to start our discussion, although starting with model development realm would also be appropriate.

As with the data farming process in general, the rapid scenario prototyping should always be within the context of the questions to be answered. These questions have to be prepared in such a way that simulation can help to find answers and to get insights. The most important step here is to define measurements to be collected by means of simulation together with required input and output data for the simulation. In most cases this step already requires some rough ideas about the scenario settings. Thus, this realm simply represents the initial formation of the basics of a scenario to be simulated.

The analysis team should make several decisions on the specifics and the resolution of the required simulation model. The analysis team should consider which kind of data is required for the analysis and how to collect these data. Many abstractions and assumptions within the modelling process have to be made and documented. A simulation model then must be chosen and if necessary, adapted to the requirements of the specific analysis. If a suitable simulation model is not available, a new model has to

be developed. All of the above is, as shown in Figure 2, is the starting point of the actual rapid scenario prototyping process, which starts with drafting a more detailed description of the scenario settings together with all the assumptions made so far. Once the scenario is drafted, it can be instantiated into a simulation model and the realm of model development is described in the next section. (Horne et al. 2013) The process of rapid scenario prototyping evaluates the plausibility of the implemented scenario by either checking the defined parameters and settings or the model itself. The resulting product of this process is a tested and documented base case scenario, e. g. the scenario, in which the questions will be answered.

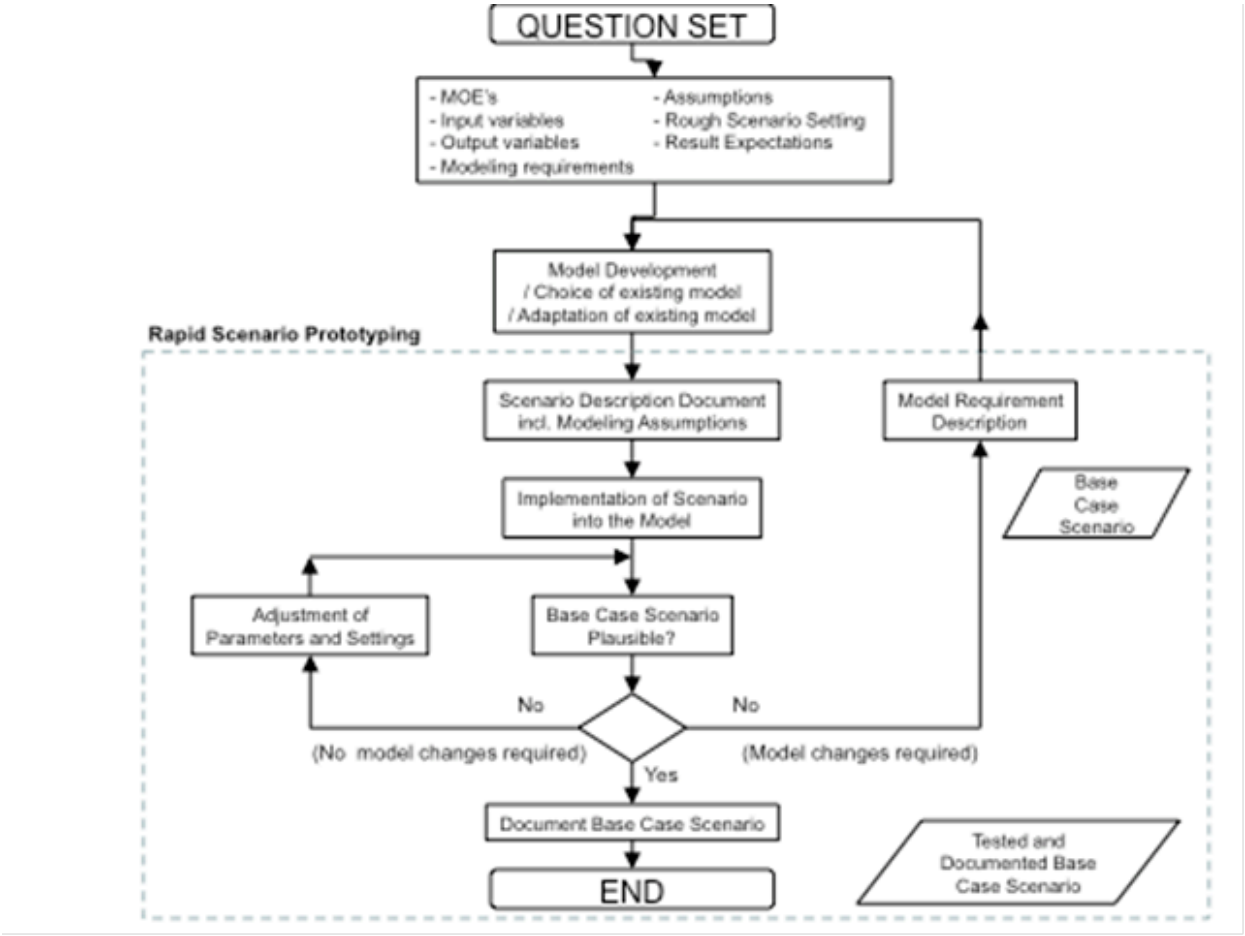

Figure 2. Rapid scenario prototyping process (Horne et al. 2013).

3 MODEL DEVELOPMENT

As stated in the previous section, the model development realm works hand-in-hand with the rapid scenario prototyping realm in the experiment definition loop on the left side of Figure 1. The fundamental output of this loop is a scenario instantiated in a working model that captures the essence of a question and that can be sent to the multi-run execution loop of the data farming process. Of course, more insight into the question, refinement of the question, and/or deeper examination of the question may be enabled later through a return to the experiment definition loop later in the process.

The model development subgroup of MSG-088 pursued the task of providing basic characteristics of data farmable simulation systems, such as general technical requirements on simulation systems that are used for data farming. We investigated possible application areas of data farmable simulation systems and

studied technical concepts within modelling. The group documented some of the most important system contributions made by each member nation. Furthermore, we documented existing model practices for data farming obtained from experiments with applications within each nation. In addition the group identified and documented the overall scope of applications and the real world domains that can be addressed using data farming methodology with the existing models. Space constraints prohibit us from discussing all of this work, but here in the remainder of this section we will summarize our recommendations regarding model development in data farming applications.

When developing models, both modelling and subject matter experts should be present. Rapid scenario prototyping provides model requirements for model development. When developing models for the questions to be answered it could be shown that creating aggregated models that combine simple models instead of building single monolithic models was more effective with respect to flexibility and performance. It was found that the more independent the models are from each other, the better the potential results were achieved, especially in the process of rapid scenario prototyping, because these small models could be interchanged more easily by other ones meeting better the requirements. Thus, one needs to encourage modularization and clear separation of different models, including development practices for using models of different aggregation level and scope.

Reusability of models is also an important topic. To achieve good reusability, models should be loosely coupled and be interoperable. We need to make models interoperable with other models and easily data farmable. Interoperability is achieved, when input and output variables of a model are properly exposed and documented. Existing standards of the modelling and simulation community should therefore be applied wherever applicable.

Furthermore, model calculations and results should be exactly repeatable. For example, any random number generators in models should have their seed values exposed as input variables, so that simulations can be repeated. Good standards require appropriate validation of models. To be useful they need to reflect reality at the correct level of approximation. In addition, data validation should be properly documented and provided.

User interfaces should be clearly separated from calculation engines. This makes it easier to reuse the models. For example, in high performance computer applications, simulation systems are often used without a graphical interface. Also, model verification should be made as easy as possible. To ensure that the models work properly, they should have an extensive test suite that can be run through. In case of problems, simulation systems should provide transparent state of their inner workings to make investigation and problem fixing easy.

Whenever possible, it is recommended to provide supporting software with the simulation systems. Complex models, especially those dealing with complex input parameters, need supporting software. This supporting software should also be provided with the simulation systems, using similar good software practices. Because even the most accurate and efficient model is useless without information on how to use it, documentation of models and their validation has to be done properly. And, finally, openness should be encouraged, the source code should be provided with the model when possible given other constraints (Horne et al. 2013).

4 DESIGN OF EXPERIMENTS

Design of experiments is one of the three realms of data farming in the multi-run execution loop. Along with the realms of high performance computing and analysis and visualization, the realm of design of experiments allow us to perform multiple runs to gain simulation results over a wide landscape of possibilities. The full MSG-088 report describes the methodology in design of experiments related to data farming and documents currently available designs in this area, but here we simply give a broad overview of design of experiments.

Simulation models have many inputs or parameters (factors) that can be changed to explore alternatives. A designed experiment is a carefully chosen set of combinations of these inputs, called design points, at which the simulation model will be run.

Changing the factors all at once limits your insights. It will allow you to see whether or not this changes the responses, but you will not be able to tell why the changes occur. For example, if mission effectiveness improves when you equip a squad with better sensors and better weapons, you will not know whether it is the weapon or the sensor that has the most impact.

Changing the factors one at a time also limits your insights. If the squad gets a very small improvement from a better weapon, a very small improvement from a better sensor, but a large improvement from both, you will not be able to identify this interaction (or synergistic effect) if the experimental design does not involve factors for both the weapon and the sensor.

Changing the factors in a brute force way, by looking at all possible combinations, is impractical or impossible, except for extremely simplistic simulations with only a handful of factors. If you have 100 sensors, each of which can be turned on or off, there are 2^{100} possible sensor configurations. Even printing these alternatives would take millions of years on the world's fastest supercomputers.

Design of experiments helps overcome the curse of dimensionality, while letting you achieve a broad variety of insights about your simulation model's performance. It provides smarter ways of setting up the experiment that facilitate follow-on analysis and visualization of results in a reasonable amount of time. The type of design used in an experiment dictates the output data that will be generated and collected in a simulation experiment. It also impacts the analysis and visualization methods that can be used in the analysis of simulation output data (Horne et al. 2013).

5 HIGH PERFORMANCE COMPUTING

The main task of the high performance computing (HPC) subgroup of MSG-088 was to document best practices and the lessons learned by the member nations in their pursuit of implementing an HPC environment for data farming. In addition, the subgroup documented those individual member nations' environments. This documentation appears in the full MSG-088 report. Here we will summarize the realm of high performance computing within the loop of loops that make up the data farming process.

HPC consists of both hardware and software resources. HPC systems can be configured as a single supercomputer with thousands of processors, as a network of clustered computers, or even as a single powerful desktop computer with multi-core processors. The hardware on these systems includes such things as processors, memory, networking hardware, and disk storage. HPC software includes, among other things: the operating system; underlying or supporting software which provide the environment to execute the model; and the data farming software, which enables running instances of the model across the HPC systems' "compute units". By generating and managing each of the model runs over a set of design points or input sets, the data farming software provides the infrastructure "glue" that "sticks together" the model, its set of inputs, the design, and the HPC resources.

The main purpose of HPC in the context of data farming is to provide the means to execute a data farming experiment. Other purposes of HPC are for use in analysis and visualization of the output and for generating designs used in future data farming experiments. Given the large number of model runs made in a typical data farming experiment, HPC facilitates conducting the experiment in a timely manner as well as supporting the storage and analysis of huge volumes of output. From a purely computational perspective, there are six elements involved in a data farming experiment:

1. A "data farmable" model (we use the term "model" generically; it can refer to any computational model or simulation).
2. A set of model inputs, generically called the "base case".
3. A specification of your experiment (the set of factors in your design and a mechanism for finding and setting those in the set of model inputs).

4. A set of HPC resources, both software and hardware, needed to execute a model "instance".
5. The data farming software.
6. A set of model outputs.

The first five elements are required to begin execution of the data farming experiment; the final element is the product or the results of the data farming experiment. Basically, the process proceeds as follows: for each "design point" in the design, the data farming software creates and executes a compute "task" or "job", where that task consists of creating a set of model inputs using the base case as a template; executing the model with that modified input set; and collecting and storing the model output for that design point. Other tasks may include collecting and staging the raw output for further analysis and visualization, additional post-processing of the output, or automated analysis of the output (Horne et al. 2013).

6 ANALYSIS AND VISUALIZATION

We define analysis as the process of examining data that is produced by data farming processes using statistical, summarization and presentation techniques to highlight useful information, extract conclusions, and support decision-making. Visualization is a collection of graphical and visual analysis techniques used to optimize and speed the process of exploring data, conveying understanding, and presenting in data farming processes. Much of the current usage of analysis and visualization in the data farming process has been the analytic examination of multiple replicate and excursion model output and we describe this usage in the full report. Here we will give some of the high level conclusions regarding the realm of analysis and visualization from MSG-088:

In order to exploit the potentially huge data output from the high performance computing execution of the design of experiments, highly effective analysis techniques must be employed. Statistical analysis and visualisation can be used to discern whether data may has useful meaningful value and aid in the translation of data into information useful in making progress in understanding possible answers to the questions at hand.

Visualization consists of analyzing the simulation output data using appropriate techniques as well as presenting the results to the decision-making authorities. Even with a smart design of experiments, simulation experiments can create huge volumes of multi-dimensional data that require sophisticated data analysis and visualization techniques.

The ability to use multiple techniques gives us the ability to explore, investigate, and answer the questions posed. Every technique has strengths and limitations, therefore, especially for highly-dimensional data sets, use of a family of techniques is preferable to use of a single technique.

As stated earlier, data farming gives us the ability to map the landscape of possibilities and in the process discover outliers. These outliers should always be considered and only be eliminated for appropriate reasons. Using various analysis and visualization techniques these outliers can also be investigated as a separate cohort of the data. The full MSG-088 report describes analysis and visualization techniques and technologies that have been used in this pursuit of both examination of the full landscape of possibilities as well as discovering the surprises that can often lead to important additional support to decision makers (Horne et al. 2013).

7 COLLABORATION

The spirit of collaboration is the key tenet of data farming. It underlies the loop of loops in Figure 1 and holds within it much of the power of data farming. Throughout the development of data farming and the formation of the data farming community, people have openly shared experiences and expertise. One focus for collaborative efforts has been and continues to be the international workshops. The first international workshop took place in 1999 at the Maui High Performance Computing Center. The first 4 workshops were methodology driven, dealing with complex adaptive systems modeling and agent based

representation, with statistical experiment design and experiment evaluation. The subsequent workshops were application driven, focusing on the development of simulation models, the implementation of scenarios of interest into the models, and the use and further development of computer clusters to run the models large numbers of times.

The real work is in making progress on important questions and the real secret is the combination of military subject matter experts and highly knowledgeable and multi-disciplinary scientists. This special mix of personnel has been the hallmark of the international workshops and this mix has promoted much networking opportunity. It has been a dynamic combination to have data farming work teams headed up by a person who really knows and cares about the question (e.g. a military officer who knows that the answers may have an impact on both mission success and lowering casualties) and supported by men and women with technical prowess who can leverage the tools available.

The collaboration subgroup of MSG-088 documented the following aspects of the collaborative processes in data farming: defining the characteristics and dimensions of collaboration in data farming, collaboration within and between the realms in data farming, collaboration of the people, application of collaboration tools. This information can be found in the full report as well as information on the current status of data farming in the attending nations and ideas about the future development of data farming (Horne et al. 2013).

8 HUMANITARIAN ASSISTANCE/ DISATER RELIEF (HA/DR) CASE STUDY

Trends and current military missions ask for new capabilities. Modelling and simulation (M&S) makes an essential contribution to support military decision makers when developing and evaluating conceptual fundamentals regarding tactical and operational proceedings. In that context the NATO Modelling and Simulation Group MSG-88 conducted case studies to illustrate the benefits of the experimentation method data farming. In the case study *Humanitarian Assistance / Disaster Relief (HA/DR)* the simulation model Sandis, which was developed by the Finnish Defense Forces Technical Research Centre, was used in conjunction with the data farming process to explore medical logistics and casualty evacuation questions for an earthquake scenario in a coastal region. Data farming was used here as an analysis process, where thousands of simulations were conducted to test a variety of potential improvement ideas for practices as well as resources.

The following questions were explored in this case study:

- How do the logistical networks, evacuation chains, and distribution of materials affect the loss of life?
- Where can the response be improved and where are the bottlenecks?
- What are the probability distributions for different triage classes over time under various conditions?
- What are the effects of changes in coordination, capacity, and resource distribution on triage classes/loss of life?
- How would better allocation of transportation resources affect the performance measures?
- What if improved ship-to-shore assets are available? What are the implications regarding this greater capacity on coordination, evacuation/treatment, and kinds of resources available?

As documented in the full MSG-088 report, the six realms of data farming were employed and allowed for an exploration of these questions. Medical facilities were embarked on ships and different courses of action and many alternatives were simulated in a joint NATO environment in response to the casualties caused by the earthquake. Variables included the number and capacity of treatment facilities (ships, hospitals, collection points), the arrival times of the ships, the capacity of transports to ships, and the speed of ground transportation. Nine different iterations of the data farming multi-execution loop were performed in the course of the work, which together included simulations of thousands of alternatives.

The remaining details are in the final MSG-088 report, but the basic conclusion was that data farming is a good process to model highly variable HA/DR situations and test a wide variety of potential

improvements ideas for practices as well as resources. This capability may be quite useful as several large recent disasters have demonstrated that significant improvement is needed in HA/DR planning and procedures, e.g. transporting aid to Haiti following the 2010 earthquake. NATO, with its common role as a coordinating agency, is in a position to make a significant impact in HA/DR practice and data farming may be quite useful in this regard (Horne et al. 2013).

9 FORCE PROTECTION CASE STUDY

A second case study allowed for the examination of questions in the area of *Force Protection.* In this case study, the agent-based simulation model PAXSEM, which was developed for the German Bundeswehr to support procurement and answering operational questions, was used in conjunction with the data farming process to find a robust configuration of a combat outpost in different kinds of threat scenarios. Data farming was used here and 241.920 simulation runs were conducted on a German High Performance Computing (HPC) Cluster with 512 nodes (processors) to check assumptions, to gain new insights, and to obtain more robust statements on opportunities and risks of specific combat outpost configurations.

This case study, also documented in Kallfass et al. (2013) , shows a successful implementation of the data farming process for a realistic operational question set to support operational decision making in an Armed Forces Staff. The work was comprised of an integrated team of subject matter experts with experiences and specific knowledge in the fields of modelling and simulation, design of experiments and military operations.

The overall question was "*In order to effectively protect a Combat Outpost (COP), which tactics / equipment are most robust against different kinds of threats?"* This question was answered via the analysis of the results of a large amount of simulated configurations in a tactical scenario that develops over time. The relevant input parameters as well as the necessary measurements of effectiveness were determined and a newly developed experimental design helped to decrease the overall number of possible configurations to a manageable size. Within the given parameter ranges of all possible COP configurations, two different classes of COP configurations were identified to be effectively robust against the different kinds of threats.

Overall, all six realms of data farming were integrated: collaborative processes, rapid scenario prototyping, model development, high performance computing, design of experiments, and data analysis and visualization showing the possibilities as well as the requirements for success when applying this approach as described in the previous chapters. This case study has proven the data farming concept and also supports our recommendation to military leaders to consider the support of data farming analyses for their decisions (Horne et al. 2013).

10 CONCLUSION

The objective of MSG-088 was to document and assess the data farming capabilities that could contribute to the development of improved decision support to NATO forces. The six realms of data farming were documented and assessed. Also, proof-of-concept explorations in the form of two case studies were undertaken. The results of both the assessment and case study explorations indicate the potential high value of data farming to NATO decision-makers.

Harnessing the power of data farming to apply it to our questions is essential to providing support not currently available to NATO decision-makers. This support is critically needed in answering questions inherent in the scenarios we expect to confront in the future. Thus we recommend implementing data farming methods as codified in the MSG-088 report in NATO modelling and simulation contexts and we recommend undertaking specific efforts to apply data farming to NATO questions.

REFERENCES

Horne, G. 1997. *Data Farming: A Meta-Technique for Research in the 21st Century*. Newport, Rhode Island: USA Naval War College.

Horne, G. 1999. "Maneuver Warfare Distillations: Essence not Versimilitude". In *Proceedings of the 1999 Winter Simulation Conference*, edited by A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 1147-1151. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Horne, G., and M. Leonardi. 2001. *Maneuver Warfare Science 2001*. Quantico, Virginia: Marine Corps Combat Development Command.

Horne, G., and T. Meyer. 2004. "Data Farming: Discovering Surprise". In *Proceedings of the 2004 Winter Simulation Conference*, edited by R. Ingalls, M. D. Rosetti, J. S. Smith, and B. A. Peters, 171-180. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Horne, G., and T. Meyer. 2005. "Data farming: Discovering Surprise". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 1082-1087. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Horne, G. 2010. "Program of Work MSG-088 Data Farming in Support of NATO." NATO Science and Technology Organization (STO), Modeling and Simulation Coordination Office, Paris, France.

Horne, G. et al. 2013. "MSG-088 Data Farming in Support of NATO." Final Report. NATO Science and Technology Organization (STO), Modeling and Simulation Coordination Office, Paris, France.

Kallfass, D., T. Schlaak, S. Seichter, and S. Mayer. 2013. "MSG-088 Data Farming in Support of NATO: Case Study Force Protection." NMSG Multi-Workshop Paper 17. Sydney, Australia.

Meyer, T., and G. Horne. 2007. *Scythe, Proceedings and Bulletin of the International Data Farming Community*. Issue 1, Workshop 13. The Hague, Netherlands.

Meyer, T., and G. Horne. 2013. *Scythe, Proceedings and Bulletin of the International Data Farming Community*. Issue 12, Workshop 25. Istanbul, Turkey.

AUTHOR BIOGRAPHIES

GARY HORNE is a Research Scientist at MCR Federal Systems with a doctorate in Operations Research from The George Washington University. During his career in defense analysis, he has led data farming efforts examining questions in areas such as humanitarian assistance, convoy protection, anti-terrorist response, and cyber security. He chaired the NATO Modeling and Simulation Task Group MSG-088, "Data Farming Support to NATO" and this group has completed documentation of the data farming process. He continues work with NATO as co-chair of the follow-on task group "Developing Actionable Data Farming Decision Support for NATO."

STEPHAN SEICHTER is Staff Officer in the German Bundeswehr. After he graduated as a Master of Science in Operations Research from U.S. Naval Postgraduate School Monterey, USA in 2005 he has been working in the fields of analysis methodologies with the focus on simulation based analysis methods for Military Operations Research. He currently is Section Chief IT-Support, Experimentation and Bundeswehr Simulation and Testing Environment at the Bundeswehr Planning Office. He co-chaired NATO MSG-088, "Data Farming Support to NATO" and continues work with NATO as co-chair of the follow-on task group "Developing Actionable Data Farming Decision Support for NATO."