Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

UNDERLYING ISSUES ASSOCIATED WITH VALIDATION AND VERIFICATION OF DYNAMIC DATA DRIVEN SIMULATION

John G. Michopoulos
Samuel G. Lambrakos

Center of Computational Material Science
Naval Research Laboratory
Washington, DC 20375, U.S.A.

ABSTRACT

This paper presents a brief exposition of three underemphasized issues concerning modeling and simulation as they relate to Dynamic Data Driven Application Systems. One issue concerns the fact that conventional procedures for validation of data-driven models and simulations are unnecessary as they intrinsically contain validation in that they have been constructed according to acquired data. Another issue concerns the inherent coupling between the experimental frame used to measure system response and the system itself. This coupling can lead to an unrealistic simulation of the system in that the data contains the interaction of the system with the experimental frame. The final issue concerns the inherent pluralism of parametric representation and of potential mappings from data space into model space. This inherent pluralism imposes the need for optimal model and data space navigation procedures that are structured for appropriate sampling of data space and specification of faithful model parameterizations.

1 INTRODUCTION

High fidelity simulation of realistic size and complexity systems have motivated studies concerning methodologies for the development of Dynamic Data Driven Application Systems (DDDAS) by various groups. The National Science Foundation has taken a leadership role in fostering research on such efforts through the initiative on DDDAS (Darema 2004).

Recent advances in DDDAS facilitated by the present level of computational achievements, as well as advances in data-driven modeling and simulation, impose the need for a critical evaluation of paradigms underlying Qualification, Validation and Verification (QV&V). Furthermore, there is a plethora of underlying issues associated with the development of DDDAS as articles related to Modeling and Simulation (M&S) that span the analytical, algorithmic, programming and problem solving contexts that practitioners are sensitive to. However, the goal of this paper is to increase the awareness of the research community with regard to three of these issues related to the quality of the final simulations of physical systems in the area of DDDAS.

The first issue concerns the fundamental irrelevance of conventional validation procedures with respect to data-driven models and simulations. This follows since conventional validation procedures are based on paradigms and associated terminology, that are historically biased and related to concepts and implementations that do not necessarily correspond to today's advanced computational practices and possibilities within the context of DDDAS. An initial attempt to explore this issue along with its historical relationship to the various scientific methods has been given in (Michopoulos and Lambrakos 2005).

There are formal aspects of "data-driven system modeling" that are related to methods of "system identification" and "parameter estimation," which are based on systems theory. In the discussions that follow, the concept of system identification has been extended to include that of data-driven modeling. Further, "dynamic" data driven is to be understood as implying system identification within the context real-time input of data, which is considered as having been obtained most recently from sensors. Similarly, the use of most recently obtained data for adaptive simulation control or steering, constitutes the second use of the term "dynamic data driven" as has been discussed elsewhere (Michopoulos, Tsompanopoulou, Houstis, Rice, Farhat, Lesoinne, and Lechenault 2003, Michopoulos, Farhat, Houstis, Tsompanopoulou, Zhang, and Gullaud 2005, Michopoulos, Tsompanopoulou, Houstis, Farhat, Lesoinne, Rice, and Joshi 2005).

A key aspect of our discussion concerns the concept of "weak termination" that implies informally the incomplete character of validation with respect to the predictive capability of a model. A rigorous examination of the mathematical foundations of this concept is beyond the scope of this presentation in that it relates to temporal and ontological formalisms of logical meta-systems (Van Benthem 1949).

1-4244-0501-7/06/$20.00 ©2006 IEEE

In this paper, we first present the definitions and origins of the particular QV&V terms. We subsequently identify how weak termination undermines the fidelity, accuracy and reusability of models. We further demonstrate how the application of data-driven M&S avoids weak termination and enables reusability of models and simulations. We further demonstrate that when an experimental frame is involved in exercising and measuring the physical system then the acquired data contain the aggregate behavior of the coupling between the experimental frame and the system itself thus leading to an inaccurate model creation that in turn forces a refactoring of the internal structure of the system. The last issue discussed is the fact that the data and model spaces are not necessarily one-dimensional and that there is large variety of possibilities corresponding to selection of loading paths and sampling schemes (from the data space) as well as representational models (from the model space). Therefore, the inescapable question for the "best choice" becomes an important issue of research. The effects these issues may have on from a V&V are also discussed. We conclude with a brief description of the three main issues.

2 DATA-DRIVEN SIMULATIONS DO NOT NEED CONVENTIONAL VALIDATION

2.1 Qualification, Validation and Verification

Model fidelity, accuracy, and high confidence in predictability requires that contemporary M&S are subject to various QV&V procedures. There are many descriptions of how QV&V relates to modeling and simulation. Figure 1 represents the modeling and simulation process along with the associated QV&V in terms of logical flow. This logical flow represents a unification of the abstractions defined by many organizations such as the AIAA (AIAA 1998), ASME (ASME-JFE 1993), DoD's Defense Modeling and Simulation Office (DMSO) (DoD) and DOE Defense Program's (DOE/DP) Accelerated Strategic Computing Initiative (ASCI) (Pilch, Truncano, Moya, Froelich, Hodges, and Peercy 2001). The dotted arrows shown in this figure represent human activities that are implemented with various degrees of automation, allowing a transitioning from the physical system to the system's conceptual model (via analysis), next to the computational model (via computational methods), and then back to the physical system (via simulation). A conceptual model is constructed by analysis of the behavioral structure of the physical system within an application context.

The conceptual model can be a set of partial differential equations (PDEs) representing conservation laws with appropriate constitutive equations. Typically, this type of mathematical representation is known as an analytical model that encapsulates the conceptual model. This model reproduces the behavior of the physical system and belongs to the class of models designated as "explicit" or "physics-aware" (Michopoulos 2004).

Another form of conceptual model is that of a "physics-agnostic" model such as rule-based (e.g. Cellular Automata or Genetic Algorithms) or input-output associator technologies (e.g. Neural Nets, Perceptrons, Support Vector Machines). The corresponding analytical or mathematical model introduces errors associated with its underlying idealized assumptions.

The computational model is encapsulated by the software that implements the conceptual model within the computational infrastructure. It is constructed by a variety of programming techniques based on various degrees of computational automation ranging from manual to automated software generation that exploits the ability of commercial design tools. The computational model justifies the need for verification because it introduces additional uncertainty that is associated with space and time discretization, i.e. errors associated with the discrete representation of differential operators and machine truncation.

The formal definitions of the QV&V terms are as follows (AIAA 1998):

- *Qualification* is the process of determining that a conceptual model implementation represents correctly a real physical system.
- *Verification* is the process of determining that a computational model implementation represents correctly a conceptual model of the physical system.
- *Validation* is the process of determining the degree to which a computer model is an accurate representation of the physical system from the perspective of the intended uses of the model.

Both validation and qualification attempt to establish the representational fidelity of the conceptual (qualification) and computational (validation) models relative to the physical system. For this reason, qualification may be considered another form of validation. This consideration explains why most of the bibliography concerns verification and validation (V&V). Verification attempts to determine the error and accuracy between two models (conceptual and computational). The modeling of uncertainty has to address the error originating from various sources extending from sensor device noise to algorithmic approximation and computational implementation of mathematical primitives by particular hardware architectures.

Shown in Figure 1 are QV&V concepts in terms of a comparison between the behavior of the physical system (via experimentation), and the conceptual and computational systems (via simulation). In addition to complexity, the most critical issue associated with this data-driven behavioral comparison among the various system representations

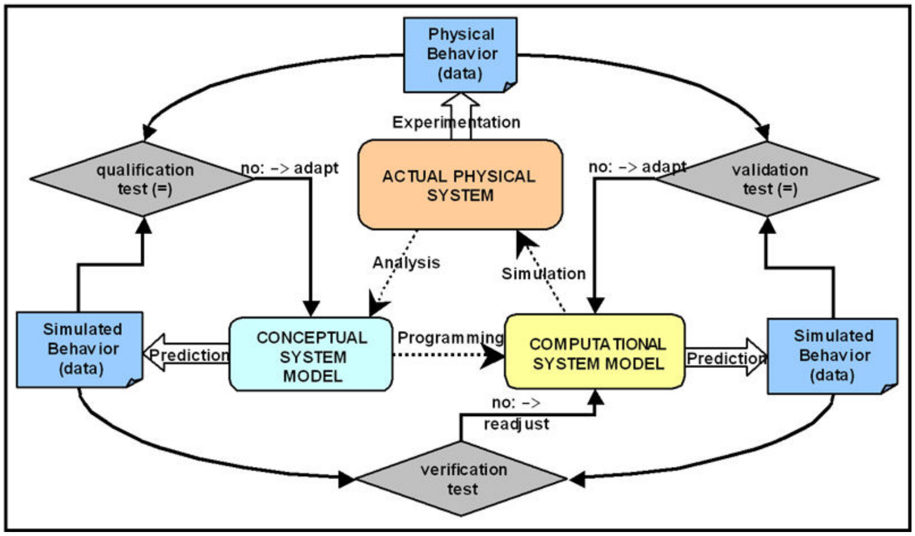

Figure 1: Flow Chart of the Traditional QV&V for the M&S Process

is "weak termination". Unfortunately, termination of the recursive process of minimizing the differences between model predicted data sets and data sets generated by the physical system, is not guaranteed for the user defined accuracy (as a measure of fidelity of simulation) in a finite time, nor is it guaranteed that there is a unique model prediction that converges to the physical system behavior with a desirable speed. This is represented clearly by the fact that as long as the comparison modules in Figure 1 evaluate to "no," the methodology will continue to iterate. This situation forces the user in general, to employ "engineering approximations." That is to say, attitudes of accepting models within "acceptable bounds" for specific applications, which consequently lead to a plethora of low-confidence models that vary according to the personal modeling assumptions of the specific user.

2.2 Embedded Validation of Modeling and Simulation

Addressing weak termination as well as moving towards a realistic approach that guarantees relative fidelity in a model requires considering an alternative approach. This approach involves implementing an M&S methodology that can utilize: (1) *data-streams* of controlled continuous system behavior (stimulus-response pairs of all observables regardless of their field or non-field nature); (2) *analytical representations of the model* that can accurately reproduce a subset of the acquired data via system identification through successive dynamic model adaptation with the help of optimization techniques (encoding intrinsically the validity of the derived model); and (3), *derived models* for simulating the predictive response of the originally modeled system or any other system that shares the identified continuous behavior with the original system.

Figure 2 shows the data driven M&S process as a modification of that shown in Figure 1. The modules located within the dashed-line region constitute an alternative conceptual model that is defined by the optimization structure comprising these modules. This approach effectively embeds data associated with observed behavior of the physical system, into the conceptual model, thus guaranteeing the intrinsic validity of the conceptual system. For continuous systems, this model usually refers to the mathematical formulation representing the conservation laws and associated constitutive equations. Since no comparative module exists between the physical and simulated data (see Figure 2) this methodology eliminates entirely the weak termination problem. Termination is strong because the adaptively computed model is trained on the behavioral data and therefore must terminate if it satisfies the optimization criteria, including any acceptable error tolerance between actual and simulated data. Accordingly, termination occurs prior to using the model for prediction of behavior.

3 DATA MAY CONTAIN COUPLING BETWEEN EXPERIMENTAL FRAME AND PHYSICAL SYSTEM

The process of forming a system model based on measurements associated with the behavior of a physical system with the aid of an experimental frame falls in the area of inverse modeling and is depicted schematically by Figure 3(a). Repeating the loop of he validation comparison many times is a direct consequence of the dynamic character of a DDDAS. This perspective reflects the most commonly

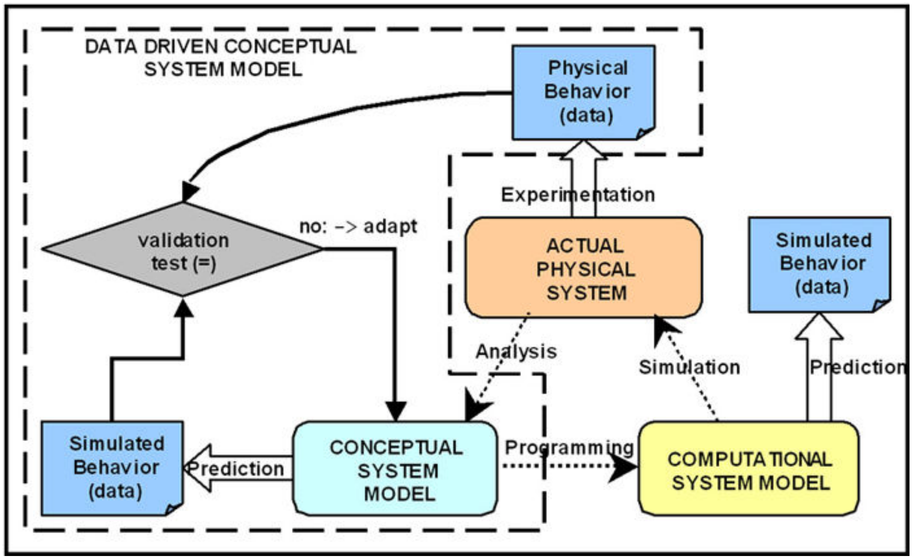

Figure 2: Evolved QV&V for Data-Driven M&S Process

accepted representation of the interaction between the experimental frame, the physical system and the corresponding model as they have been introduced by the the theoretical foundations of M&S (Bernard P. Zeigler 2000).

When the optimizer's tests have been terminated then the established model can be used within the simulator to interact with the experimental frame and adjust certain preferred features of the data acquisition and control process in the context of the continuous interaction of measurements with simulation as predicated by the DDDAS context as shown in Figure 3(b). It is conceivable that any given DDDAS can alternate from the model formation scheme on the left to the data acquisition and control scheme on the right as many times as the validation acceptance criteria are violated during the life of the entire simulation.

However, the experimental frame itself is comprised from the actuation, sensing and control subsystems. Furthermore, all measurements are achieved through the sensing layer of the experimental frame and all excitation of the system is also achieved through the actuation layer subsystem of the experimental frame. Thus, these subsystems appear more as subsystems of the measured and actuated physical system from the perspective of the data used to construct the model. Therefore the acquired data used to identify the model of the physical system in the optimizer, are carrying the interaction of the sensing and actuating layer with the physical system. Thus, the identified system is not really the physical system of interest, but rather a coupled system corresponding to the composition of the sensing and actuating subcomponents of the experimental frame with the actual physical system as shown in Figure 4(a) for the case of the model identification and in Figure 4(b) for the case of the simulator that dynamically adjusts the remaining control layer of the experimental frame. The uncertainty introduced from error propagation of statistically definable noise or deterministically definable deviations through the sensing and actuation subsystems is contained in the data and inescapably will be attributed to the physical model while this is not true in reality. This refactored view of the identified system suggests that in order to isolate the behavior of the actual physical system one has to be able to decompose the sub-behaviors of the sensing and actuating layers of the experimental frame. Thus, identifying the behavior of these subsystems ahead of time as well as during the actual system identification (because it may depend on it) is a direction not to be forgotten in DDDAS development and application practices by the respective researchers and practitioners. Of course, this depends always on the strength of the interaction and its effects on the data. In other words, for some cases it may be ignored but for some other ones it has to be considered.

4 EXISTENCE OF DATA AND MODEL SPACES SUGGEST SEEKING OPTIMAL CHOICES

The two most important concepts of a DDDAS are the data used to make the model or/and adjust the simulation and the model associated with the underlying physical system. In fact, in the inverse problem and system identification bibliographies there is a scarce reference to the data and model spaces (Tarantola 2004). The isomorphic aspects between these disciplines and DDDASs as described elsewhere (Michopoulos and Lambrakos 2005) directly suggest by implication that data space and model space play a

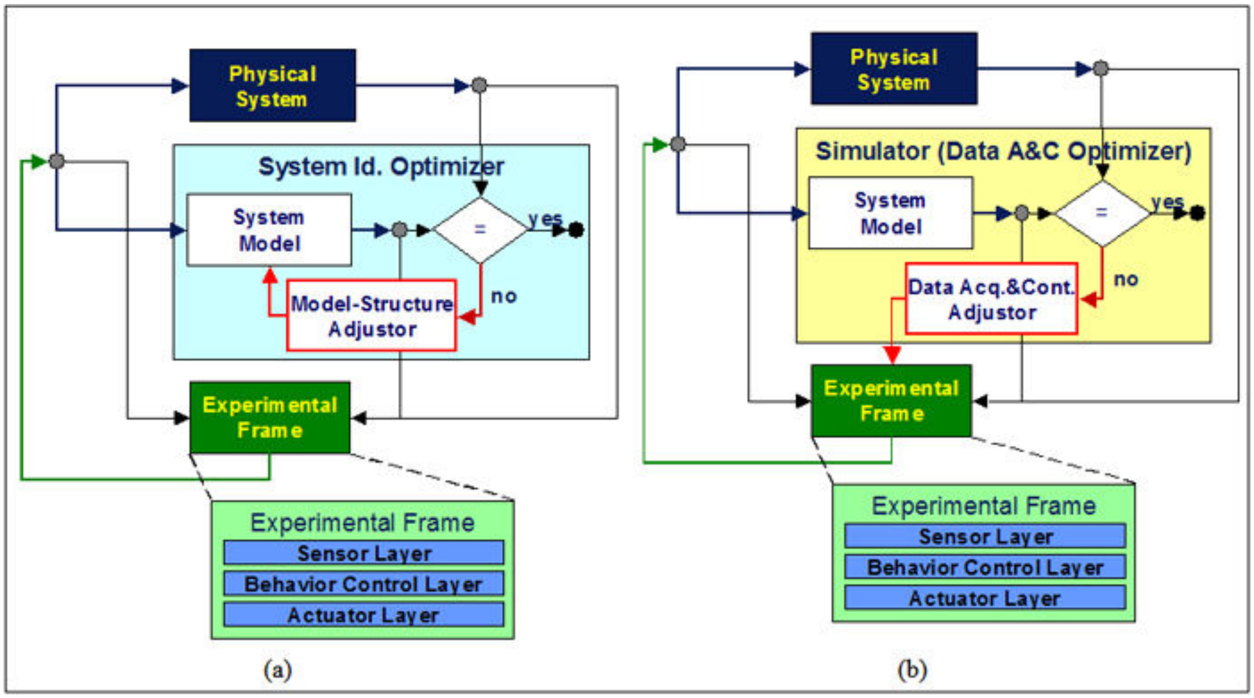

Figure 3: Conventional Coupling among Physical System, Experimental Frame and System Model Optimizer Components (a) and Data Acquisition and Control Simulation Optimizer (b)

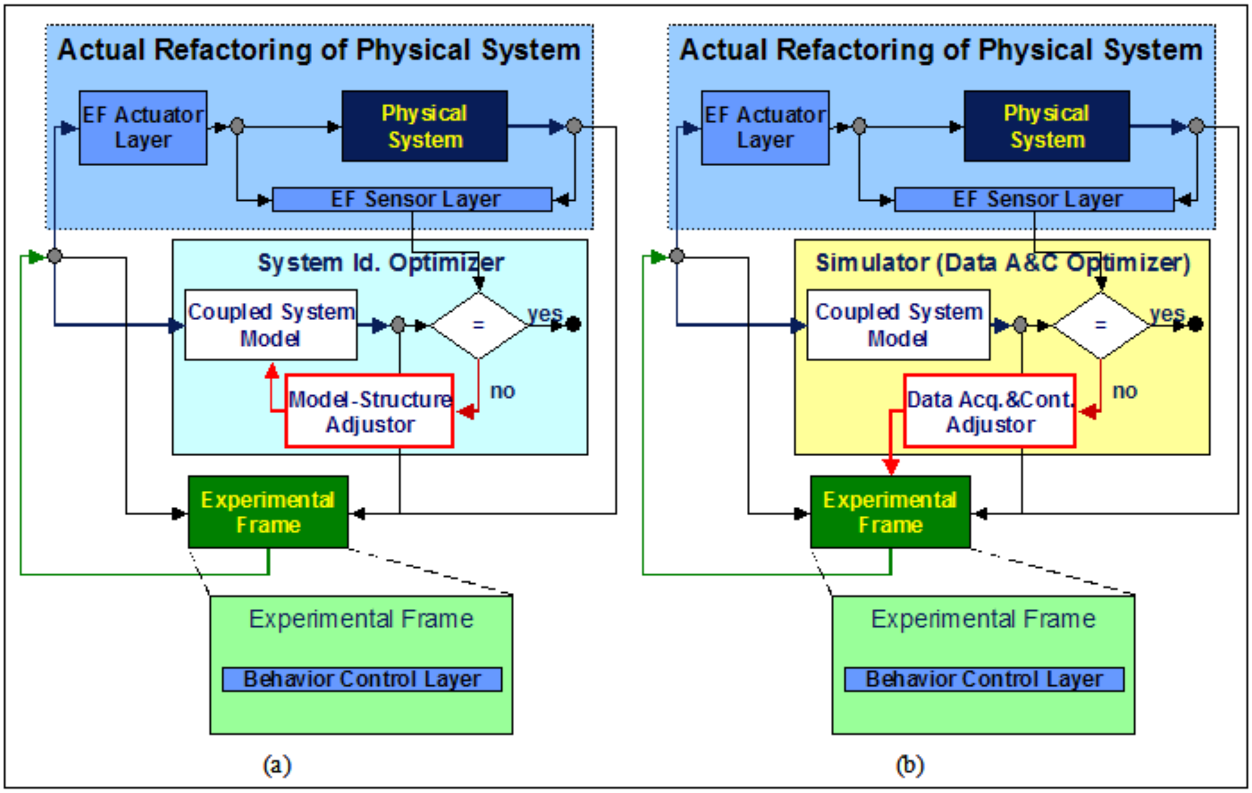

Figure 4: Actual Coupling among Physical System, Experimental Frame and System Model Optimizer Components (a) and Data Acquisition and Control Simulation Optimizer (b)

very important role in the development and application of DDDASs.

Spaces in general, are characterized by their respective dimensionality and the associated basis. Determining both of these attributes for both the data and model space such as one achieves the DDDAS simulation objectives can be a central problem when fidelity of simulation depends for example on the tradeoff of accuracy with efficiency as markers of the quality of the simulation.

4.1 The Data Space

Data-space is the space containing all the data that can be generated by observing the systemic behavior via various sensing technologies. It has both a discrete and continuous character as its basis is formed by the parameters (discrete or continuous in time) that constitute the input and output vectors of a given system. Its dimensionality is determined by the total number of these parameters. Ideally, one would prefer that these input-output variables or parameters are independent from each other. In many system analysis approaches like in the case of continuum mechanics, the data space of a system coincides with the state space of the measurable state variables characterizing the behavior of the system. The data-space provides us with the data to be used to identify the system model and/or correct the simulation in real time.

In Figure 5 a section $R \subseteq \mathbb{R}^3$ of a data space is shown spanned by the basis (i_m, i_n, o_i), where i_m, i_n represent two arbitrary input variables of the system and o_i an arbitrary output of the system. Any trajectory in this space such as the one denoted by the black line reflects a behavioral expression of the system. Any data acquired from this space will have to lie one such a trajectory. For systems that the inputs are controllable by the experimental frame the stimulus paths can be defined on the subspace $P \subseteq \mathbb{R}^2$ spanned by the basis (i_m, i_n). One such path corresponding to the black trajectory is the green path on this input subspace.

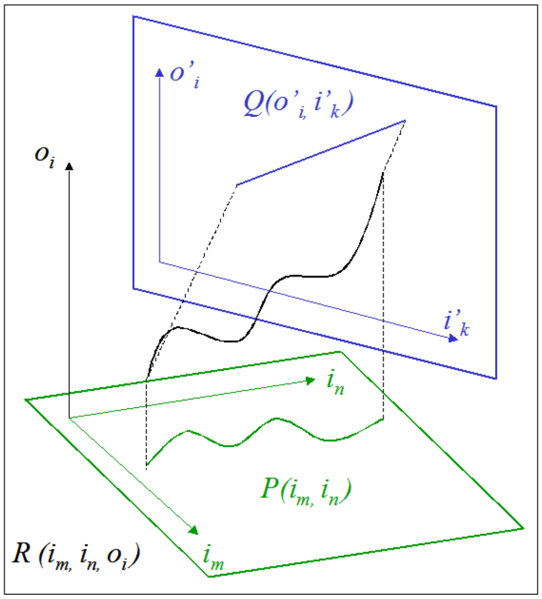

Figure 5: Data Space Section and Potential Subspaces

Important issues that arise within this context are the determination of the data sampling distance along the behavioral trajectory, its variability, the shapes of the proposed excitation paths and their sequencing. More importantly one may ask if indeed the dimensionality and type of basis selected is appropriate, underspecified or overspecified for determining the behavior of the system. A case in point is shown when one considers yet another subspace $Q \subseteq \mathbb{R}^2$ spanned by the basis $(i'k, o'_n)$, where the projection of the actual trajectory (in the higher dimensional superspace $R \subseteq \mathbb{R}^3$) appears as a line without undulations (blue line in Figure 5). This clearly indicates that it may be dangerous to underspecify the observation space because it will yield to an unrealistic system determination. Overspecification of the system also is problematic because it yields to less efficient and more costly system identification or simulation in order to capture redundant and unimportant features of the systemic behavior.

4.2 The Model Space

Following the inverse-problem approach, a system is represented by a model and associated set of adjustable parameters. The particular choice of a model (or equivalently, model and associated set of parameters) is termed a “parameterization” of the system. The choice of a parameterization to be used to describe a system, however, is in general not unique. In order to address the property of non-uniqueness of system parameterization, inverse problem theory has adopted the concept of ”model space,” where each point of this space represents a “conceivable” parameterization of the system. Given a model space for a specific system, quantitative system identification is further enhanced, or optimized, by isolating those regions of model space corresponding to parameterizations that establish a well-conditioned mapping between model and process parameters, i.e., between model and data space, over a sufficiently wide range of values of the parameters. Specification of a well-conditioned mapping between model and data space is equivalent to (or implies) the specification of a complete set of basis functions. It follows that one may establish a correspondence between an optimal system parameterization and an optimal set of basis functions for parametric representation.

The concept of a model space and optimal parameterization establishes a foundation for providing a relatively more rigorous definition of the term “data-driven model.” Accordingly, the general relationship between data space and model space is such that the vector space character of the data space as determined by the functional representation of the data is what determines what are the optimal regions of model space to be adopted for model representation. This is in contrast to using a particular model parameterization to determine what regions of the data space are to be sampled.

4.3 Important Questions

The existence of an optimal parameterization imposes the need to examine what are the practical considerations for development of numerical procedures for effecting an optimal model representation. Among these considerations are: 1) what are the optimal dimensionalities and sets of basis functions for both the data and model spaces, or equivalently, can a subspace be defined in practice that tends minimize data sampling, while still maximizing the encoding of system response characteristics of interest; 2) can one in practice isolate the optimum path(s) and determine the best path sequencing for acquiring data from the data space and isolating optimal regions of model space; and 3) can one in practice determine those regions of data and model space that tend to satisfy distinct performance criteria.

A reasonably optimal system parameterization having been established, it follows intuitively that not every segment of the data space will be mapped into the associated region of model space. Accordingly, the existence of any reasonably optimal parameterization will imply the existence of a reasonably optimal subspace within a given data space. It is significant to note that the structure of the subspace establishes a criterion for optimal filtering of data, which can be interpreted as a projection operation with respect to the subspace. It follows that regions of data space that cannot be projected onto the subspace should not be sampled. The existence of a reasonably optimal subspace within data space implies that numerical procedures whose algorithmic structure is such that they sample regions of data space that do not project onto the optimal subspace will tend to be ill conditioned.

This line of thinking indicates that another optimization scheme is necessary. It lies above the level of the one used to determine the values of the parameters of a model and it rather determines the optimal parameterization in the model space and the optimal subspace characteristics and sampling data path features in the data space. For this reason it may be called “meta optimization” to denote the next hierarchical level of its application.

5 CONCLUSIONS

Our discussion has examined several significant aspects of validation related to M&S. In particular, that a posteriori validation is unnecessary for data-driven adaptive M&S methods because validation is implicit in these methods. This implicit validation makes data-driven methods very desirable with respect to reliability and suggests that when sufficient data is available, data-driven methods are preferable. In addition, one must be aware of a fundamental property of causality, i.e. that any measurement process influences the system being measured. This awareness is important owing to the fact that the bias introduced by the character of the coupling between measuring system and the physical system itself, may be significant.

Finally, an important point of this paper is that DDDAS implementations have been intimately associated with data and model spaces. These spaces have certain properties and features suggesting that there may be a wide range of methodologies for exploiting their vector space properties. Seeking the optimal methodologies is therefore a desirable goal. The associated questions introduced by this final point suggest that a meta-optimization process should be considered for establishing the suitable restricted neighborhoods within the data space and model space relating to the DDDAS being considered.

ACKNOWLEDGMENTS

The authors acknowledge the support by the National Science Foundation under grants EIA-0205663 and CNS-0540419. Partial support from NRL’s 6.1 core-funding is also greatly acknowledged.

REFERENCES

AIAA 1998. Standards program, guide for the verification and validation of computational fluid dynamics simulations. Technical Report AIAA report G-077-1998, American Institute of Aeronautics and Astronautics.

ASME-JFE 1993. Journal of fluids engineering editorial policy statement on control or numerical accuracy. *Jnl. of Fluids Engineering* 115 (3): 339–340.

Bernard P. Zeigler, Tag Gon Kim, H. P. 2000. *Theory of modeling and simulation*. 2nd Edition ed. San Diego: Academic Press.

Darema, F. 2004. Dynamic data driven applications systems: A new paradigm for application simulations and measurements. In *International Conference on Computational Science*, ed. M. Bubak, G. D. van Albada, P. M. A. Sloot, and J. Dongarra, Volume 3038, 662–669: Springer-Verlag.

DoD. Verication, validation, and accredization (vv&a) recommended practices guide, defense modeling simulation office, office of the director of defense re- search and engr., available: www.dmso.mil/docslib.

Michopoulos, J. 2004, September 5-10. Mechatronically automated characterization of material constitutive respone. In *Proceedings of the 6th World Congress on*

Computational Mechanics (WCCM-VI), 486–491. Beijing China: Tsinghua University Press and Springer.

Michopoulos, J., C. Farhat, E. N. Houstis, P. Tsompanopoulou, H. Zhang, and T. Gullaud. 2005. Dynamic data driven methodologies for multiphysics system modeling and simulation. See Sunderam, van Albada, Sloot, and Dongarra (2005), 616–623.

Michopoulos, J., and S. Lambrakos. 2005. On the fundamental tautology of validating data-driven models and simulations. See Sunderam, van Albada, Sloot, and Dongarra (2005), 738–745.

Michopoulos, J., P. Tsompanopoulou, E. Houstis, C. Farhat, M. Lesoinne, J. Rice, and A. Joshi. 2005, April. On a data-driven environment for multiphysics applications. *Future Generation Computer Systems* 21:953–968.

Michopoulos, J., P. Tsompanopoulou, E. Houstis, J. Rice, C. Farhat, M. Lesoinne, and F. Lechenault. 2003. Design architecture of a data driven environment for multiphysics applications. In *ASME 2003 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Chicago*, Volume I668CD: ASME Press. Paper-No.:DETC2003/CIE-48268.

Pilch, M., T. Truncano, J. Moya, G. Froelich, A. Hodges, and D. Peercy. 2001, Jan.. Guidelines for sandia asci verification and validation plans - content and format. Technical Report SAN2000-3101, Sandia Reports.

Sunderam, V. S., G. D. van Albada, P. M. A. Sloot, and J. Dongarra. (Eds.) 2005. *Computational science - iccs 2005, 5th international conference, atlanta, ga, usa, may 22-25, 2005, proceedings, part ii*, Volume 3515 of *Lecture Notes in Computer Science*. Springer.

Tarantola, A. 2004. *Inverse problem theory and methods for model parameter estimation*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Van Benthem, J. 1949. *The logic of time*. D. Reidel Publishing Company.

AUTHOR BIOGRAPHIES

JOHN G. MICHOPOULOS is a senior research scientist and engineer at the Naval Research Laboratory where he is leading the Computational Multiphysics Systems Lab within the Center of Computational Material Science. From 2004 to today, he serves as an associate editor of *ASME Transactions Journal of Computational and Information Sciences in Engineering*. He has also held various offices in Navy related committees and his the chair of the *Computational Technologies in Engineering and Science Applications* committee of the *Computers and Information in Engineering* division of the ASME; He is a member of ASME and ACM. His e-mail address is <john.michopoulos@nrl.navy.mil>, and his group's web page is <cms.nrl.navy.mil>.

SAMUEL G. LAMBRAKOS is research physicist in the Center of Computational Material Science at the Naval Research Laboratory for over 20 years. His research areas include inverse methods applied to spectroscopy, particle beam scattering and heat deposition processes. His e-mail address is <samuel.lambrakos@nrl.navy.mil>.