Session B6
V&V (especially validation) for M&S which employ Significant Aggregation

Session B6 leaders:
Co-Chairs: **Andreas Tolk** (VMASC/ODU) and **Robin Miller** (UK MoD)
Session Recorder: **Bob Senko** (DMSO)

B6 Materials in Foundations '02 proceedings:

Paper

Implications of Metamodeling, Multi-Resolution Modeling (MRM), and Exploratory Analysis for Validation (46 pp)
Paul Davis (RAND)
Jim Bieglow (RAND)

Slides (may contain back-up materials and notes)

Implications of Metamodeling, Multi-Resolution Modeling (MRM), and Exploratory Analysis for Validation (72 slides) [in both pdf and ppt formats]
Jim Bieglow (RAND)
Paul Davis (RAND)

Participants in this session are listed at the end of the Discussion Synopsis.

Discussion Synopsis (to provide perspective on papers & briefings identified above).

The following figure gives an overview of the discussion following the presentation within the B6 session. The various contributions fit into this general schema.

V&V in the Context of MRM

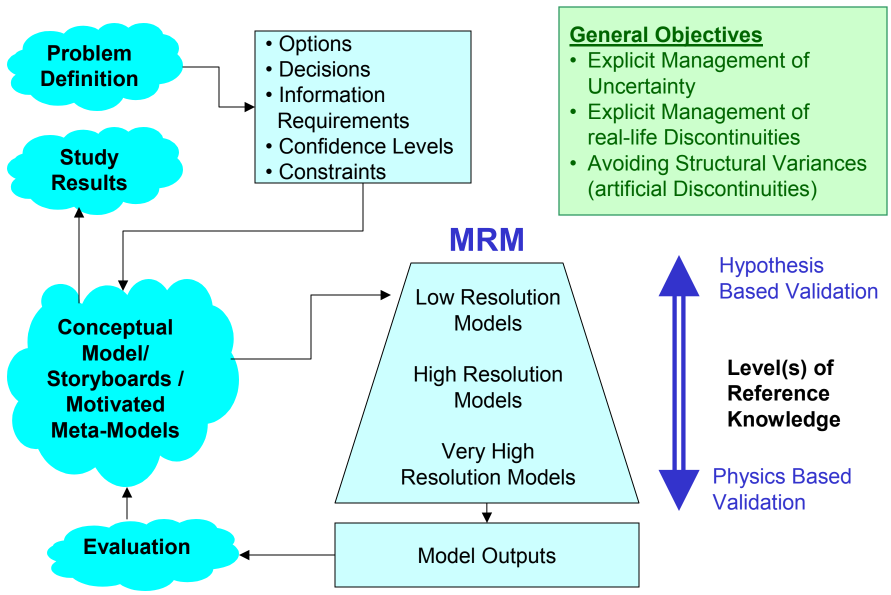


Figure 1: Overview V&V in MRM

Generally, the purpose of a model – or a set of models – is to aid the decision making process (procurement, operational support, etc.). It does this by providing predictions in form of possible consequences of different decisions. Establishing validity is answering the question, "Are we building/using the right model(s)?" positively, i.e., the model is fit for the purpose. In the context of Multi-Resolution Modeling (MRM), in addition the consistency of the set of models to be used has to be ensured, i.e., MRM is building families of mutually consistent models at different levels of resolution.

To be able to do so, the purpose of a study, i.e. the use of a model in a context, is essential. To be able to establish validity, which sometimes even may be the proof of validity of the study ("Are we asking the right questions?"), the implicit knowledge of the stakeholders must become explicit. This results in a first, ideal conceptual model. This conceptual model comprises the relevant factors, the relations, and parameters for the factors (including confidence intervals if applicable). This model may comprise various sub-models of different resolution. It is the bridge between the stakeholder and the analyst and ensures that the problem is understood on both sides.

Next, three questions have to be answered in the light of validity:

- What is the necessary level of abstraction for the model/sub-model which is needed to cope with the questions
- What is the obtainable level of reference knowledge against which we might validate the model used to contribute to the solution of the problem
- What models/sub-models are available on which levels that can be used in this context

The last question results from practical constraints. The choice of the best fitting model is part of the validation process, which should not only include the selection criteria, but also a list of successful mappings to the conceptual model as well as a list of open issues, i.e., which parts of the conceptual model couldn't be mapped to parts of the model of choice. Also, the following challenges have to be addressed in this phase:

- Deal explicitly with aggregation, disaggregation, and various levels of resolution
- Deal explicitly with uncertainty
- Deal explicitly with discontinuities
    - Make real world discontinuities explicit
    - Avoid artificial discontinuities (see excursus on structural variances)

To be able to deal with real world discontinuities, it is necessary to be aware of the fact that higher echelons (often represented by lower resolution models) will cross discontinuity borders of lower echelon (represented by higher resolution models), e.g., when dealing with break-points within operations. This enables dealing with resulting effects, e.g., by accepting the discontinuity, override the results of the high-resolution output within the results of the low-resolution model.

In many cases, the explicit formulation of the conceptual model enables hypothesis-based validation instead of the actually often used demonstration/physics-based validation. The hypothesis must be based on the purpose of the study and possible options and decisions to be made. Especially when dealing with studies lying in the future, this seems to be the only feasible way. Nonetheless, the use of (physics-based) validated high-resolution models as a reference model still is seen as a valuable option by parts of the community and therefore shouldn't be neglected.

The level of uncertainty also tends to increase with the timely distance of future studies. Therefore, in general, it is recommended to use low-resolution models to cope with these questions, as the amount of assumptions needed to be able to initialize high-resolution models is to big to ensure a rigid validation process for the resulting model/data combination.

The actual level of knowledge to validate against is often not clear in the study. To be able to deal with this issue, not only the resolution, but also the scope of the study relevant. In particular, areas of continuity are of interest, or – in other words – the borders of discontinuity. Often, transitivity of results – in particular success functions measured by level dependent measures of merit – is assumed which not necessarily needs to be the case. To be able to cope with these challenges, a mission/task-oriented aggregation may make more sense than the often used entity based aggregation. In particular, validation should cope with:

- Looking at the lower level for the boundaries of discontinuities;
- Looking for transitive winning functions; identify intransitive factors and ensure they are properly represented in the model:
- Explicitly identifying external factors that can drive the results over the boundary of continuous domains.

To cross the borders of different resolution, aggregation and disaggregation is necessary. When building a new model, refinement of the model – leading to a higher level of resolution by adding detail- should be treated as aggregation, and generalization of the model – combining details to a super group, thus reducing detail – should be treated as disaggregation. Therefore, aggregation and disaggregation themselves are not simple functions, but are instantiating models, hence need to be validated. To be aware of the aggregation/disaggregation and the underlying models is the first step in the validation process. Therefore, respective documentation is vital to the validation process.

In general, conceptual models based on a general description language can facilitate the validation. This is not only true for multi-resolution models. The conceptual model must be explicit about the constraints. This is seen as a facilitator as many applications do not need to cope with the complexity of high-resolution models.

Validation can also be interpreted as a method to give meaning or semantics to a specification or the model. This includes psychological as well as socio-technical aspects of the environment of the study, for example:

- Who is the stakeholder
- Who is interested in the results of the study
- Who is making the final decision
- What are the constraints of the study (e.g., are there any biases)

Having a good story is important, because:

- It helps establish face validity;
- It presents and explains results to client;
- It facilitates to extrapolate results;
- It facilitates to build a better metamodel/conceptual models

Examples for factors for good metamodel are identified as:

- Goodness of fit
- Plausible story line
- Parsimony – small model not a lot of variables
- Identification of critical elements – sets of variables all of which must work to make the system perform adequately

Hence, the proposal of the invited paper to use motivated meta-models in form of stories and storyboards for the model therefore makes sense. Such stories help to instantiate the conceptual model. Moreover, the same meta-models can be used to explain results of the used models to the non-analysts, i.e., to make the results transparent. These stories, however, introduce a new level of resolution and comprise aggregation that again must be subject of validation. With this point of view, every model is embedded into an MRM context and the results of this track are transferable to a much broader community of interest.

Excursus: Structural Variances, MRM, and V&V

Structural Variances are counterintuitive discontinuities in the simulation results that occur due to the structure of the model and not because of some error in the code or the data. A definition as well as references to further sources can be found in [Andreas Tolk, *Non-Monotonicities in HLA-Federations*. Paper 99S-SIW-001, Spring Simulation Interoperability Workshop, Orlando, Florida, 1999].

Structural variances occur – among other possibilities – if the internal decision modes, the internal models of the physical processes, and the external evaluation are not harmonized (see Figure 2). If the internal decision mode decides something that cannot be executed by the internal physical models, aggregation or disaggregation is necessary, which may result in discontinuities. The same is true for the perception of the situation, on which the internal decision algorithms operate, thus what the physic models produce must be harmonize with what the decision algorithm needs. If the internal decision mode is not harmonized with the general objectives measured by the external evaluation process, this will lead to discontinuities as well. And if the external evaluation process needs aggregation or disaggregation of the outcome produced by the physical models, this is another source of structural variances.

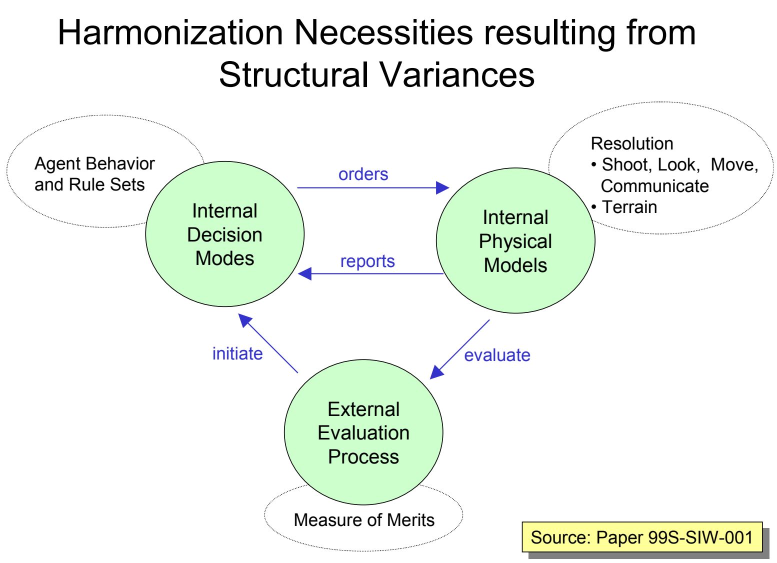


Figure 2: Structural Variances

Structural variances already appear in single, assumed one-level-resolution models. As in the invited paper was shown, even in these models the likelihood of being a multi-resolution model in the whole story of a study is high. The danger of structural variances due to the lack of proper harmonization increases dramatically for federations. Some first examples are given in [Andreas Tolk, *Non-Monotonicities in HLA-Federations*. Paper 99S-SIW-001, Spring Simulation Interoperability Workshop, Orlando, Florida, 1999].

These observations motivate the following extension to the model given in the invited paper, adding the level of evaluation to the possible chain of aggregations and disaggregations.

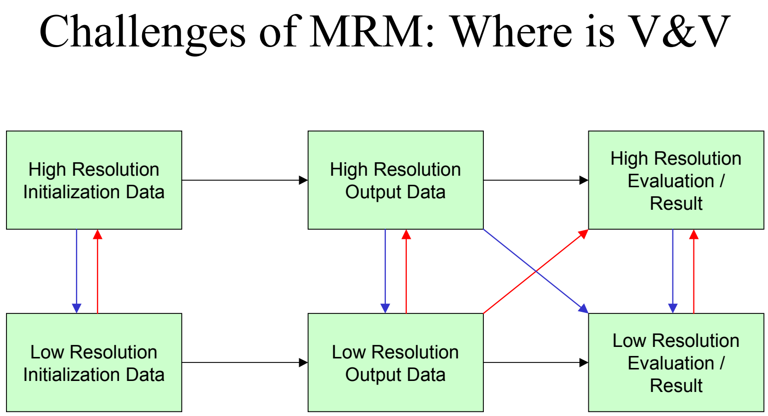


Figure 3: MRM and Structural Variances

V&V should be connected with every aggregation and disaggregation process ensuring that the underlying model. The first steps must be:

- To make the aggregation/disaggregation obvious
- To document the underlying models

SYNOPSIS OF THE B6 INVITED PAPER

Analysts routinely think at multiple levels of resolution. They check complex computations (high-resolution) against back-of-the-envelope calculations (low-resolution). This is sometimes called the "giggle test."

An analyst will run a large simulation model with gigabytes of inputs and outputs, but then explain why the results came out that way with a simple story. If the analyst explains to himself, it's called "understanding," and it's used to debug the model and data, and to plan the next case to run. If the analyst explains it to other analysts, it's called "face validation."

The analyst takes the results of a dozen or fewer cases from a large model, and prepares two or three briefing slides for the client. The model has shown the analyst a few trees in great detail. The analyst uses a mental model—his story about what the model is doing (or ought to be doing)—to explain and generalize the results, giving the client a view of the forest.

Another use of multi-resolution thinking involves connecting broad issues to more detailed questions. Defence planning must occur at the strategic level, where one deals with questions of force structure, possible future threats, and the implications of future advances in technology.

These broad questions must ultimately be connected to questions of what new weapon systems should we buy. It makes no sense to try to describe the entire landscape of strategic issues at the level of detail of the individual weapon system, so we develop different models (some mental, some explicit but qualitative, some quantitative) to deal with different levels.

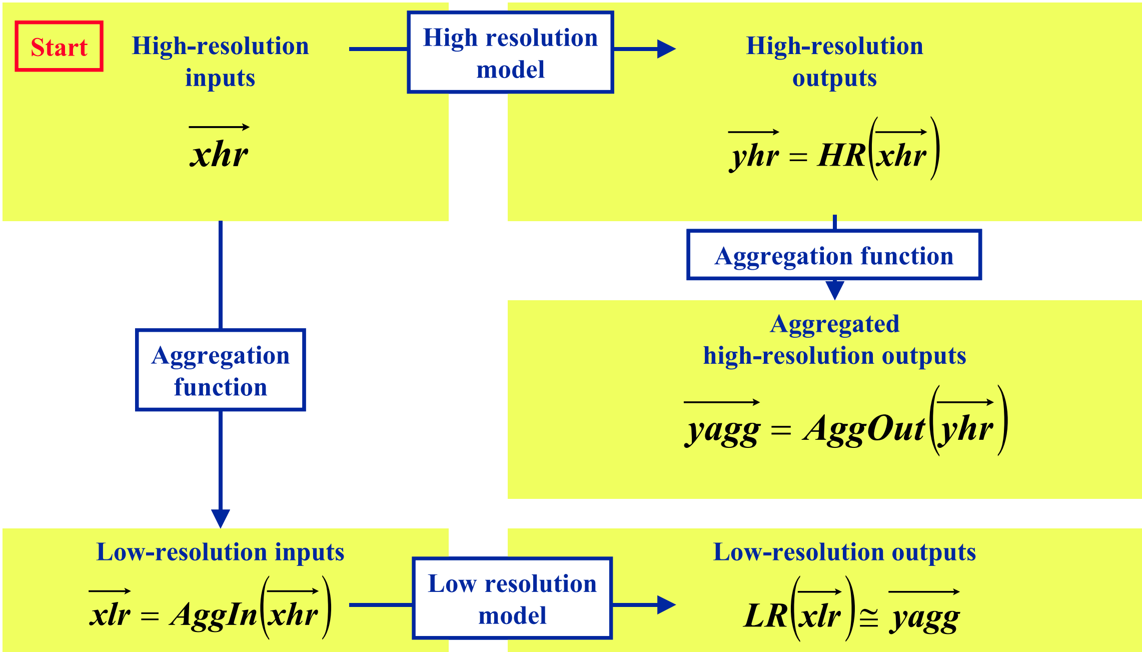
Most of the problems dealt with by the authors have massive uncertainty. The strategic issues just mentioned involve guessing what might happen over the next 20-30 years. These problems call for a very different analysis strategy than do engineering problems for which there is a well-established methodology and experience base. The engineering problems call for predictive models; our problems require exploratory analysis.

By exploratory analysis we mean exploring the space of inputs to the model, as opposed to predefining a few cases to run (a base case, a couple of excursion cases, and a handful of sensitivity cases). The task of exploring input space grows exponentially with the dimensionality of the space, and becomes infeasible with a model that has more than ten or so input parameters that must be varied. The logic chain, then, says that massive uncertainty breeds the requirement to explore input space, which in turn requires that input space have only a handful of dimensions, which forces us to use low-resolution models.

But when we find an interesting region of input space, we may wish to examine it in greater detail. This implies the need for higher-resolution models of the same phenomena. Voila—multi-resolution modelling!

If the problem has massive uncertainty, and consequently we cannot use models to predict, what should be meant by validation? This is a basic research question. We suggest that to be valid, the model should be structurally sound and the analysis domain should include all cases of interest. But we have done no work on devising practical tests for these notions.

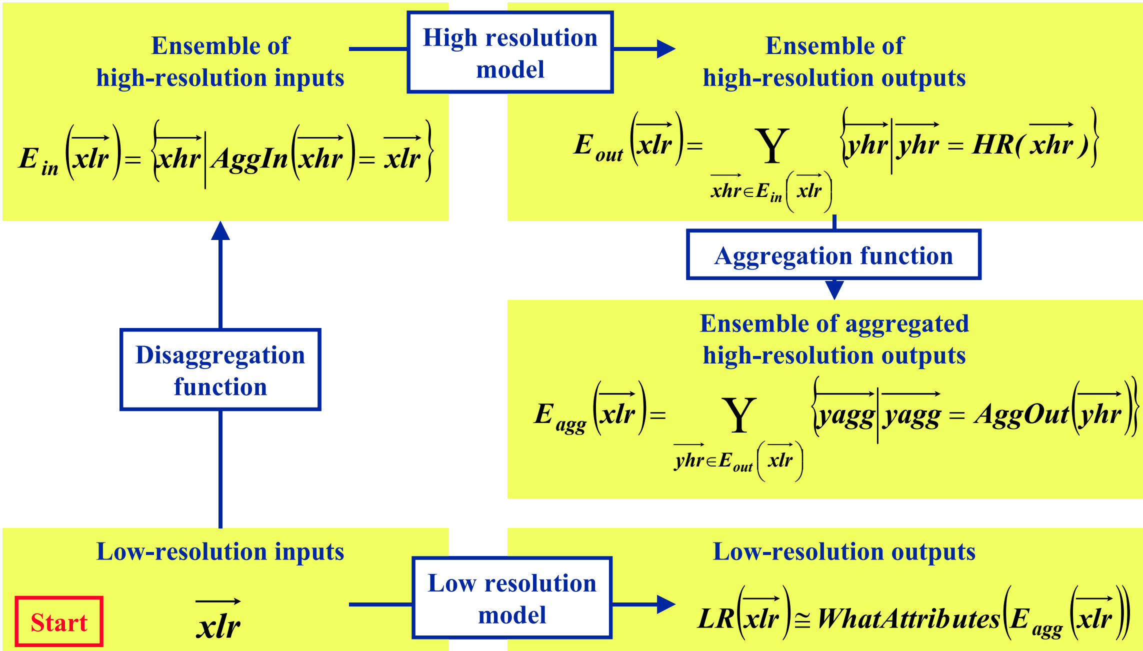
If we do work with models that represent the same phenomena at different levels of resolution, we want to be assured that the models are consistent with one another. The usual definition of consistency is shown here.



Start at the upper left with a particular high-resolution case. We can follow two paths to get to a low-resolution result. We can run the high-resolution model and then aggregate its outputs. Or we can aggregate the high-resolution inputs and feed them to the low-resolution model. The usual definition of consistency requires that the two paths produce approximately the same results.
What the usual definition ignores is that many high-resolution cases are mapped into the same low-resolution case. Resolution is the ability to distinguish one thing from another. So cases we can distinguish at high resolution get "mooshed" together at low resolution.

In this view of consistency, we start with a low-resolution case and follow two paths. The path through the high-resolution model goes up, to the right, and then down. The low-resolution path goes straight right.

Let's follow the two paths in the consistency diagram. We start with the inputs to a low-resolution case. Moving up, we identify all the high-resolution cases that map into the low-resolution case. We run the high-resolution model for all of them, not just one. This yields a set of high-resolution outputs. Then we aggregate the results to arrive at a set of possible low-resolution results. The second path feeds the low-resolution inputs we started with into the low-resolution model. For the models to be consistent, the results from following the two paths must be "comparable." But what should "comparable" mean?



(This more general notion of consistency should also apply to consistency between a model and the real world—that is, to validity. That is, a model is valid if it's consistent with observations on the real world. The difference is that there is no high-resolution model to transform inputs into outputs. There are just observations. Any partition of observed quantities into inputs (i.e., causes) and outputs (i.e., effects) represents a human imposition of structure onto the real world.)

So just what attributes of the ensemble of aggregated high-resolution outputs should we be trying to estimate? For simplicity, suppose we are estimating a single quantity, such as kills per shot of ATACMS/BAT, or the distance an invading armoured column can reach before it is halted. Then, if the high-resolution model is well-behaved, the ensemble will be an interval. (As an aside, it doesn't have to be an interval. If the high-resolution model is chaotic it can be a very strange-looking set indeed. But that's a subject for another time.) If it's a short interval, we can pick any point in it--say the midpoint—for the low-resolution model to estimate.

If it's a longer interval, we may be able to make do with the MAX or MIN. This would be our choice if we want to use the low-resolution model to build an *a fortiori* argument. That is, we could say that if we chose such-and-such a policy, it would perform no worse than X.

But what if it is a long interval and we can't use the MAX or MIN? Well, in a pinch we could use both. This would give us the full range of possible results. But in many instances this would be so wide as to be nearly useless.

The obvious alternatives come from statistics--estimate a confidence interval or a mean and standard deviation. The hitch in this strategy is that to calculate any of these quantities you need a probability distribution for the high-resolution cases. There often isn't one that's natural to the problem. So the analyst must invent one, typically by choosing a strategy for sampling the high-resolution cases (aka experimental design).

Finally, the paper discusses strategies for building a low-resolution model (called a metamodel) of a high-resolution model (called the object model). The two extreme approaches seem to be statistical and phenomenological. The statistical approach analyses data generated by running the object model many times, disregarding prior knowledge about how the object model works or any theory it may be based on. The phenomenological approach derives the metamodel by formal methods, e.g., approximating and simplifying the mathematical relations in the object model, and disregarding data generated by running the object model. No real analyst would use either pure approach, but where on the spectrum between them is the best approach?

We conducted an experiment that suggests that a heavy dose of theory improves the metamodel. One can think of this as using theory to suggest the form of the metamodel, and then fitting the resulting form to data generated by running the object model. In our experiment, incorporating theory resulted in a metamodel that fit the object model much better than the ones we obtained by relying more heavily on statistical methods. The theory-based metamodel also had fewer independent variables, fewer parameters to calibrate from data, and it more successfully identified critical components—sets of variables that all had to have favourable values for the system to work well.

Another term for the “theory” behind the metamodel is “story.” As suggested at the start, a plausible story is a valuable adjunct to any model. It helps to establish the model’s face validity, and to explain and generalize results to the client. It provides a basis for extrapolating beyond the range of our data, and since our problems have massive uncertainty, extrapolation is always necessary. And our experiment suggests that it is extremely helpful in building a model of a model—a metamodel. This should be no surprise. Analysts never build models from data alone. They always have a cause-and-effect concept they are trying to capture in the model.

B6 Session Participants (11)

First Name	Last Name	Organization
James	Bigelow	RAND
Sam	Johnson	Air Force M&S Policy Division
Joe	Kovalchik	JHU/Applied Physics Lab (APL)
David	MacKay	AFSAA/SAAP
Waits	May	Los Alamos National Laboratory
Robin	Miller	DSTL
Robert	Senko	Consultant
Jack	Sheehan	Defense Modeling & Simulation Office
John	Smale	GAO
Susan	Solick	TRADOC Analysis Center
Andreas	Tolk	VMASC