V&V Principles and Challenges

Timothy Trucano
Optimization and Uncertainty Estimation Department
William Oberkampf and Martin Pilch
Validation and Uncertainty Processes

Sandia National Laboratories
Albuquerque, NM 87185

Phone: 844-8812, FAX: 844-0918
Email: tgtruca@sandia.gov

Trilinos Users Group Meeting 2006

SNL, November 9, 2006

SAND2006-7046C

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

SAND2006-7046C

Verification and Validation (V&V) Definitions

Verification: Are the equations solved correctly? (Math)

Validation: Are the equations correct? (Physics)

ASC:

- <u>Verification</u>: The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model.
- <u>Validation</u>: The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.

V&V targets <u>applications</u> of codes, not codes.

Verification and Validation (V&V) Definitions

Example: linear solvers

$$Ax = b$$

"Solve equations correctly" and "correct equations" still make sense.

- But also keep in mind (IEEE):

  Verification = requirements implemented correctly (not quite Math)

  Validation = correct requirements (not quite Physics)

All the general issues project to Trilinos:

- Hence Trilinos is part of the "problem"
- But Trilinos may also be part of the "solution"

SAND2006-7046C

WHY do we have a V&V program?

“We built a good code – it’s not our fault that nobody used it.”

The ASC Program's Grasp is large!

"The purpose of computing is not insight."

Instead, the NNSA Advanced Simulation and Computing Program states that the purpose of computing is:

"high-performance, full-system, high-fidelity-physics predictive codes to support weapon assessments, renewal process analyses, accident analyses, and certification."

(DOE/DP-99-000010592)

Outline

History

Process

Verification

Validation

Stockpile – Toward QMU

Final Cautions

Perceived Mission Statement: Taking the fun out of computational science.

A Little History

- **Autumn 1997 validation workshop kicks off development of an ASCI “Validation Program” (“code developers do verification”).**
- **ASCI “Validation and Verification Program” launches FY1999 (“maybe we do need to worry about verification”).**
- **FY2007 begins 9th year of “Verification and Validation Program” (“verification is very important and difficult”)**
- **The program has been highlighted by milestones designed to optimize SSP impact.**
- **The ASC Predictive Science Academic Alliance Program (PSAAP) White Paper on V&V does a reasonable job of defining the goals and values of the program.**

1998


DRAFT

DOE Defense Programs
Stockpile Stewardship Program

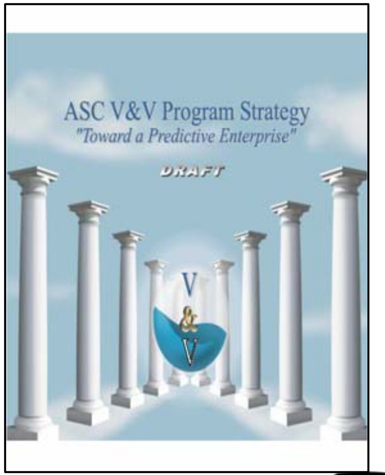
Strategic Computing & Simulation
Validation & Verification Program

Program Plan

April 1998
Version 2.5

U.S. Department of Energy Defense Programs
Lawrence Livermore National Laboratory
Los Alamos National Laboratory
Sandia National Laboratories

2006



Sandia National Laboratories

Do you trust the calculation?

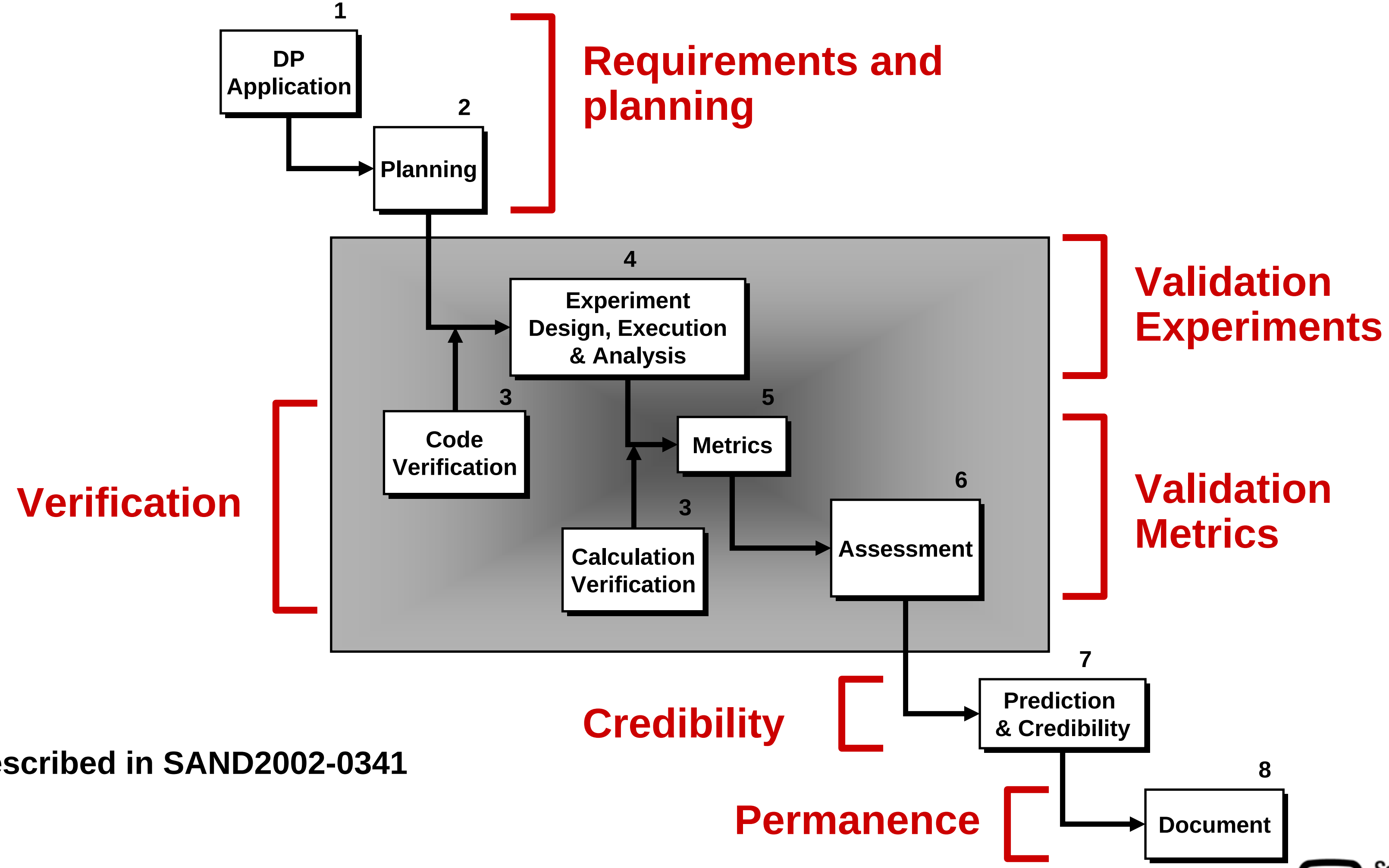
Can you trust the calculation?

Three reasons you may not wish to bet your life on a calculation:

1. **Wrong physics (validation)**
2. **Wrong numerics (verification)**
   - **Wrong math, algorithms, software**
   - **Lousy numerical accuracy**
3. **Wrong use of the results* (decisions)**

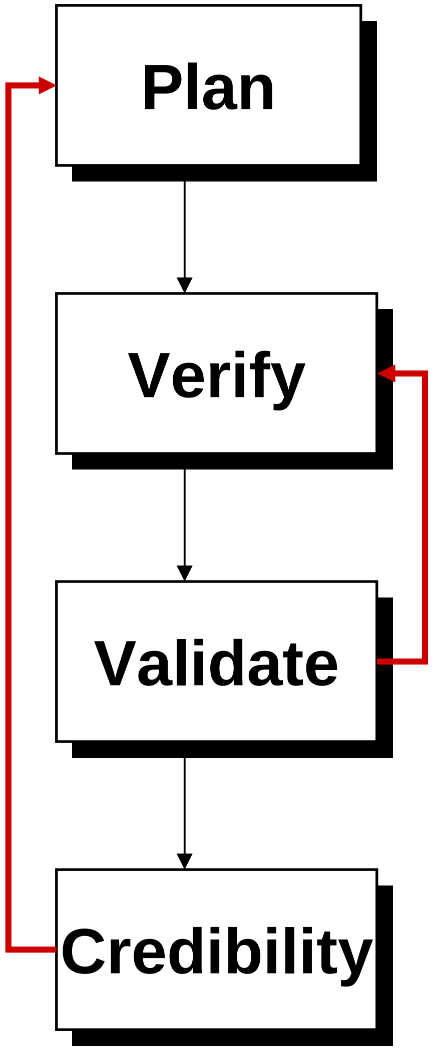
(* Especially scary!)

V&V is a methodology.



Described in SAND2002-0341

Simplified (everybody in ASC does this):



- This methodology has been applied to all SNL ASC V&V milestones.
- Current work includes QASPR and an FY08 HEDP milestone.

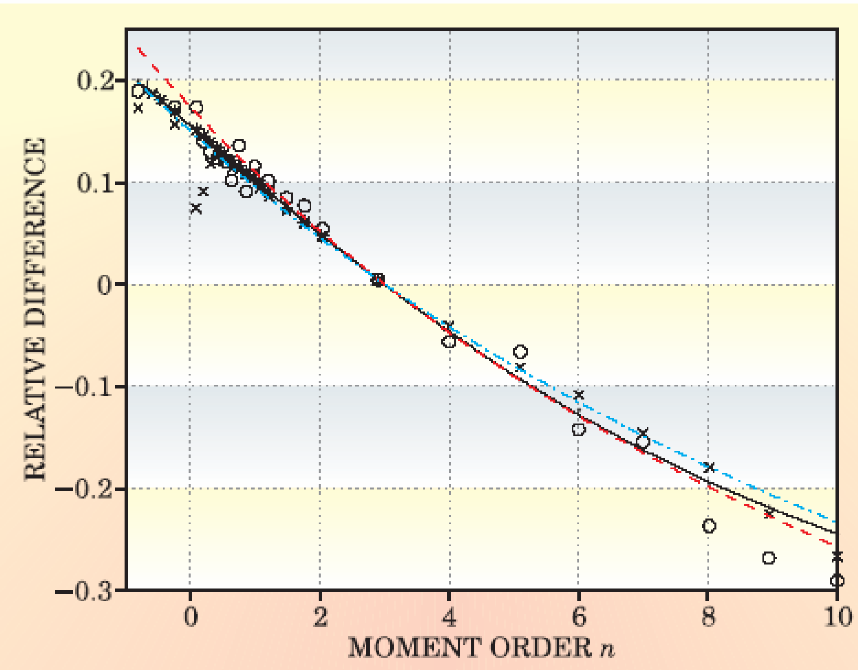
Planning

It's all about the APPLICATION of the code(s):

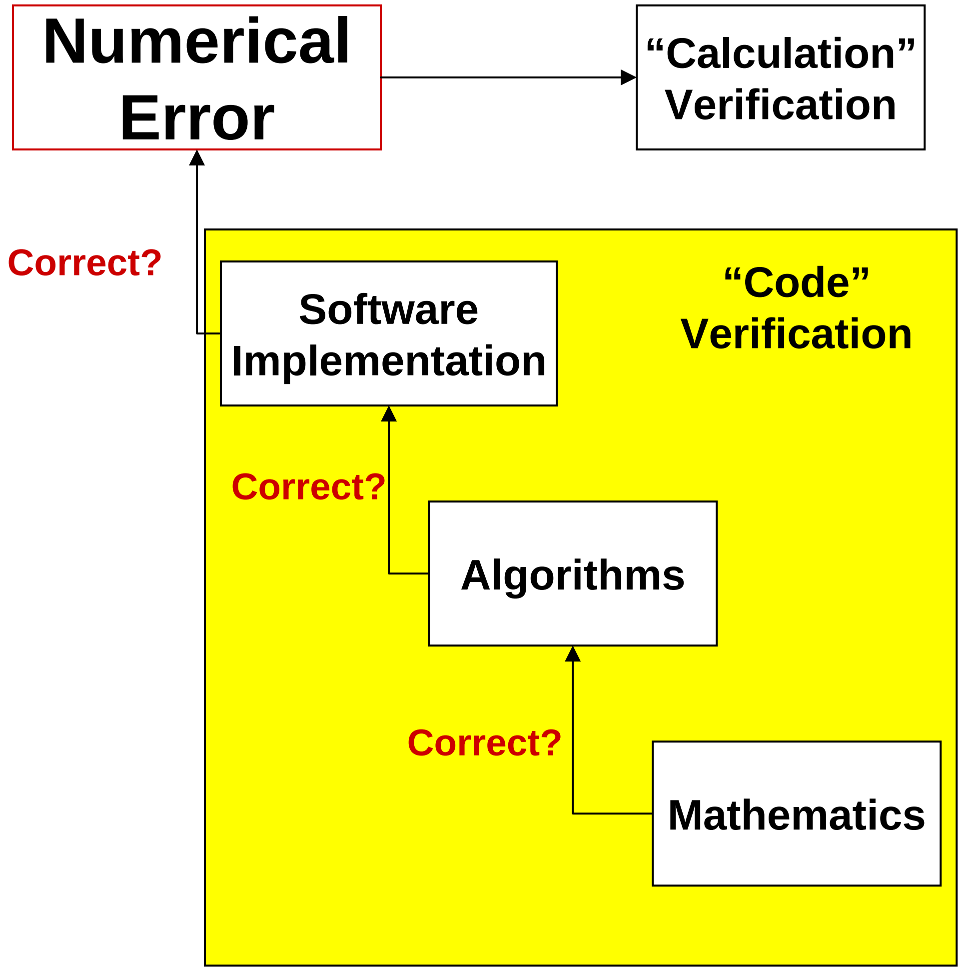
- **Application Requirements**
- **Phenomenology Identification and Ranking Table (PIRT)**
- **Priorities**
- **Identify focused verification requirements (supplementing code development)**
- **Identify <u>hierarchical</u> validation requirements**
- **Understand resultant link to credibility for the application**
- **Described in SAND2000-3101.**

> **"Everybody in the room pretty much knows what to do …"**

Verification
Consider the following comparison with data:



- What is the computational (crosses and stars) error?
- "Good agreement" with experimental data (circles) does not imply numerical accuracy!
- (A DNS resolution study underlies the computational points.)

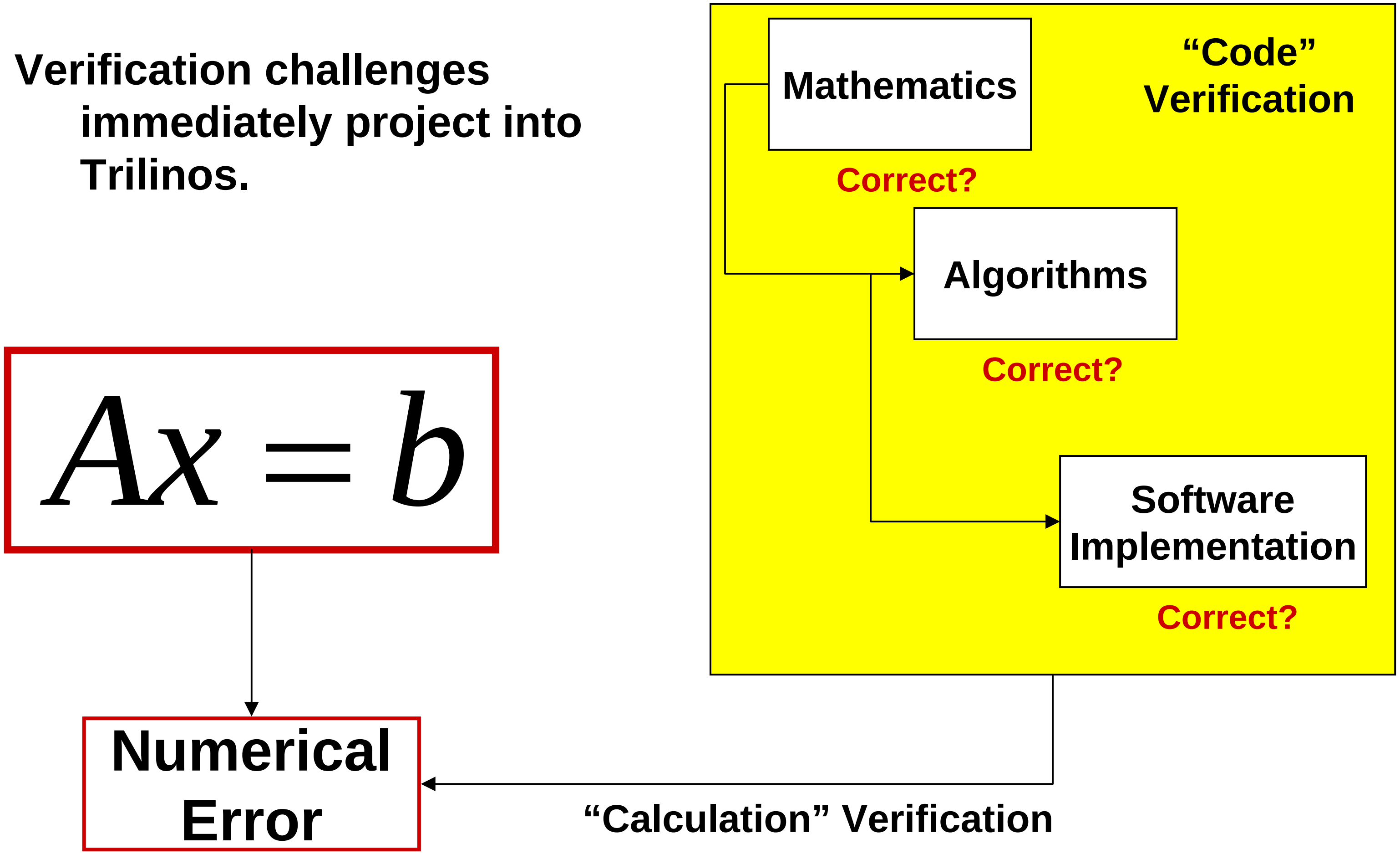


To believe any numerical error statement requires "code verification."

SAND2006-7046C

Verification

Verification challenges immediately project into Trilinos.



Verification is hard!

- This was clearly stated to ASCI, for example in January 1999.
- Credible numerical error statements require a significant code verification foundation:
  - Proofs that math and algorithms are correct
  - Proofs that the software has no bugs
  - Anything less is an approximation and has epistemic (lack-of-knowledge) uncertainty attached to it
- Error statements themselves (solution verification) come from a (presently) limited technology base:
  1. Convergence studies (highly empirical – can I take these to the bank?)
  2. A posteriori error estimation (not in our favorite equations)
  3. Error “models” with intrinsic uncertainty (“the error probably is…”)

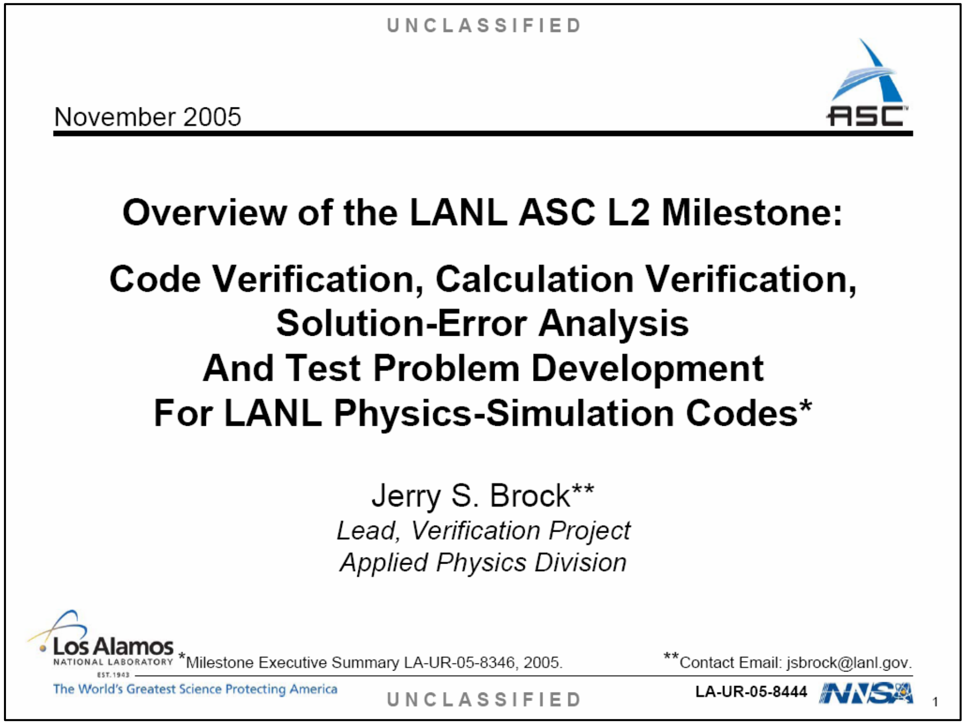
Verification is hard!
Code verification:

Two particulars:

- **Software engineering, which I will spare you since you know much more about this than I do.**
- **Specialized verification testing – "functional testing" (not "unit" testing, not "regression testing," not "structural" testing).**
  1. **How to define tests? See Roy (2005) for current discussion of test design, including Method of Manufactured Solutions**
     - **There is NO AUTHORITATIVE PRESENTATION OF TESTING FOR CS&E IN ANY LITERATURE!**
  2. **Why define formal benchmark sets (so-called Verification Test Suites)? See Oberkampf et al 2002, 2004 for some discussion.**

Verification is hard!
Convergence studies:

- **Work at the state of the art is illustrated by the LANL 2005 Level II verification milestone.**
- **Explored convergence error estimation both in asymptotic and non-asymptotic regimes, as well as other work.**
- **Heavily documented.**


UNCLASSIFIED

November 2005

Overview of the LANL ASC L2 Milestone:

Code Verification, Calculation Verification, Solution-Error Analysis And Test Problem Development For LANL Physics-Simulation Codes*

Jerry S. Brock**
Lead, Verification Project
Applied Physics Division

Los Alamos NATIONAL LABORATORY EST. 1943
The World's Greatest Science Protecting America

*Milestone Executive Summary LA-UR-05-8346, 2005.
**Contact Email: jsbrock@lanl.gov.

LA-UR-05-8444

UNCLASSIFIED

1

Verification is hard!
A posteriori error estimates:

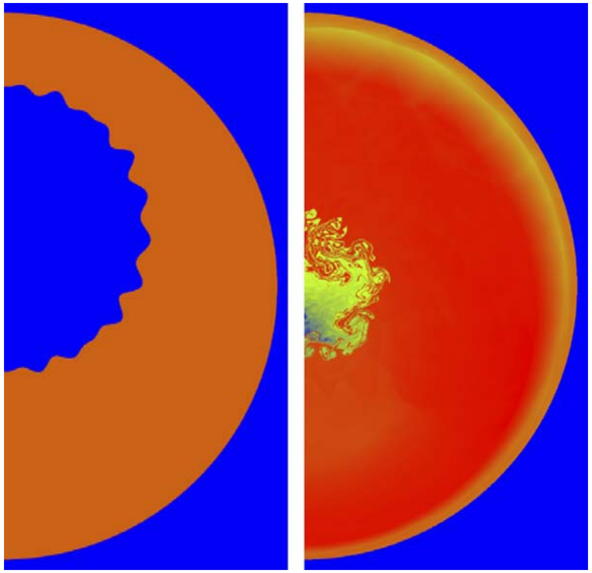
- **Remains a "Holy Grail" effort for multi-material compressible flow hydrodynamics coupled to multiple and multiscale energy transport mechanisms and material descriptions.**
  - **E.g. need error estimates, not error indicators, for non-genuinely hyperbolic systems, including mixed hyperbolic-parabolic systems.**
- **See, for example, Fuentes et al (2006) [Oden's "goal-oriented error estimation program"] for current initial work on transient, nonlinear problems.**
- **Ongoing debate about how long the legs on this program are for NECDC-type problems. ☺**
- **Adaptive mesh refinement (AMR) doesn't necessarily provide error estimates (error indicators versus error estimates).**
  - **Nor does it eliminate the need for verification.**

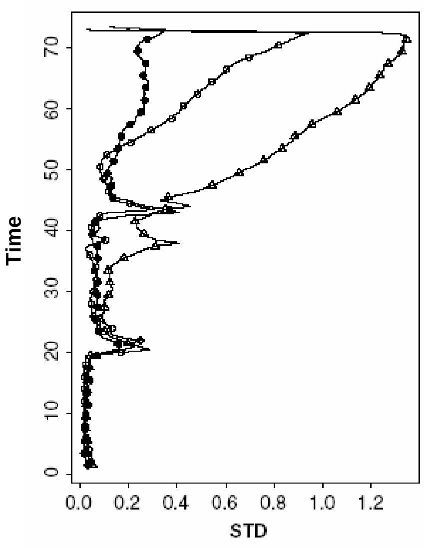
SAND2006-7046C

Verification is hard! Probabilistic error models:

- Glimm and colleagues: treatment of numerical error as an uncertainty (incomplete knowledge).
- Probability used to quantify error models.
- See Yu et al (2006) for latest published example.
- I believe this (type of) work has great importance for the long run. For one thing, it is compatible with QMU.

Imploding shock wave on perturbed material interface.

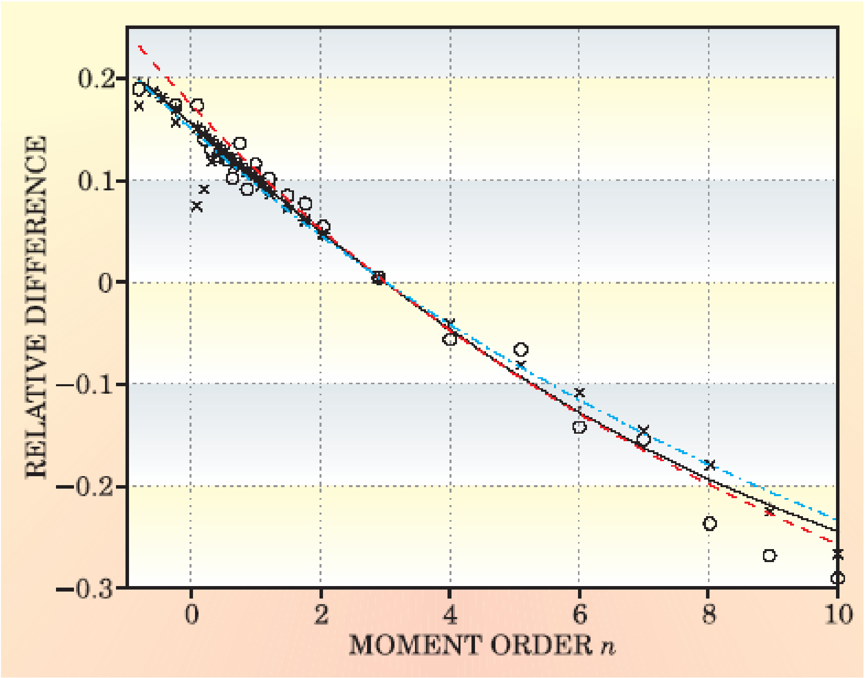




Standard deviation of probabilistically interpreted interface position error as function of time, different meshes.

Validation is hard!

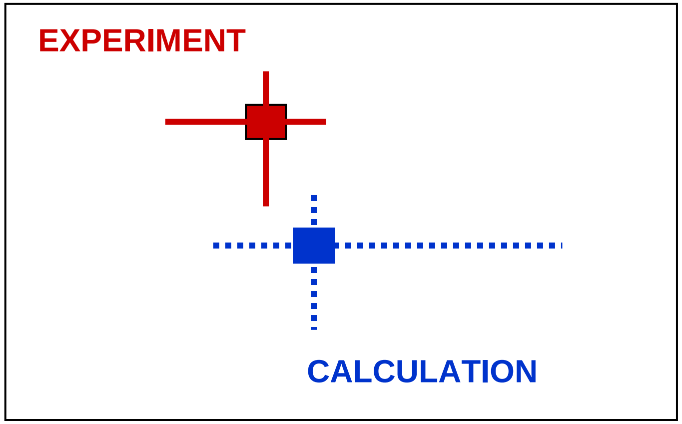
Uncertainty of both the calculation and the experimental data referent is dominant in validation.



- **Assume calculations are converged, say their error bar is the size of the plot symbol.**
- **What does the comparison mean?**
- **THERE ARE NO EXPERIMENTAL ERROR BARS (i.e. experimental uncertainty quantification).**
- **THINK QMU.**

"Experience and instinct are poor substitutes for careful analysis of uncertainty."

Validation lies at the heart of "predictive codes"



In principle, a simple strategy:

"Converge" the calculation.

Put in enough physics to insure "agreement" of calculation and experiment.

This is the 1995 charter of ASCI.

- Experimental uncertainty (variability, bias, diagnostic fidelity) is remarkably hard to quantify.
- Quantitative expt-calc differences are uncertain quantities *
- We aren't converging calculations yet (10 years later).
- What ARE the calculation error bars? Why would anybody believe the reported value? (VERIFICATION IS CRITICAL)
- How much physics do you need?
- How much agreement is good enough?

Validation is hard!
Technically speaking, validation is -

* Characterization of (say) the high-dimensional random field

Diff = "Nature – Calculation"

- Given relatively sparse information
- For the purpose of making a reliability statement about "*Calculation*"
- This interpretation has historical leverage in the atmospheric sciences:
  – For example, see Jolliffe and Stephenson (2003), <u>Forecast Verification</u>; Wilks (1995), <u>Statistical Methods in the Atmospheric Sciences</u>

Validation is hard!
Have we detected a trend?

Where is the numerical accuracy estimation and experimental uncertainty quantification in these kinds of comparisons?

Re: the "Mystery Calculation" Rogue's Gallery

Validation is hard!
Summary of Rogue's Gallery:

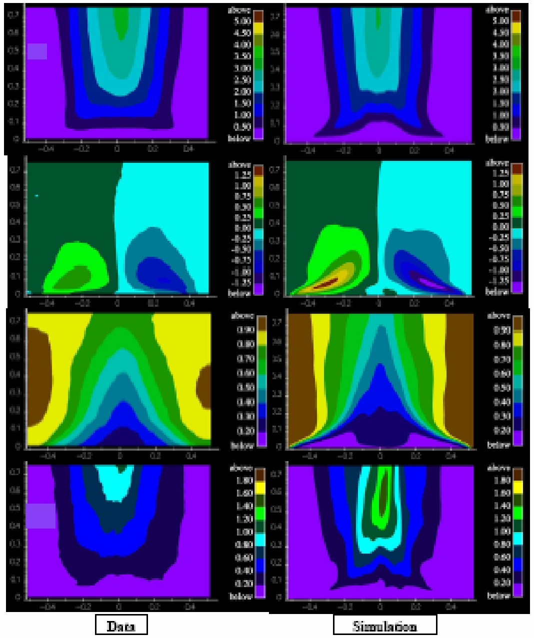
- **Little or no information about V&V**
- **Reported "V&V" has little or no formality**
- **Experimental data have little or no quantified uncertainty**
- **Little or no discussion of computational error**
- **Confusion of robustness with respect to a different grid with "small numerical errors"**
- **Comparison with experiment to claim small numerical errors**
- **Viewgraph norms and spaghetti plots for validation**
- **Information inadequate to repeat calculations**
- **Information inadequate to repeat experiments (or model them with other calculations**
- **Confusion of calibration and validation**
- **Archaic or non-existent editorial policy for computations**
- **…**

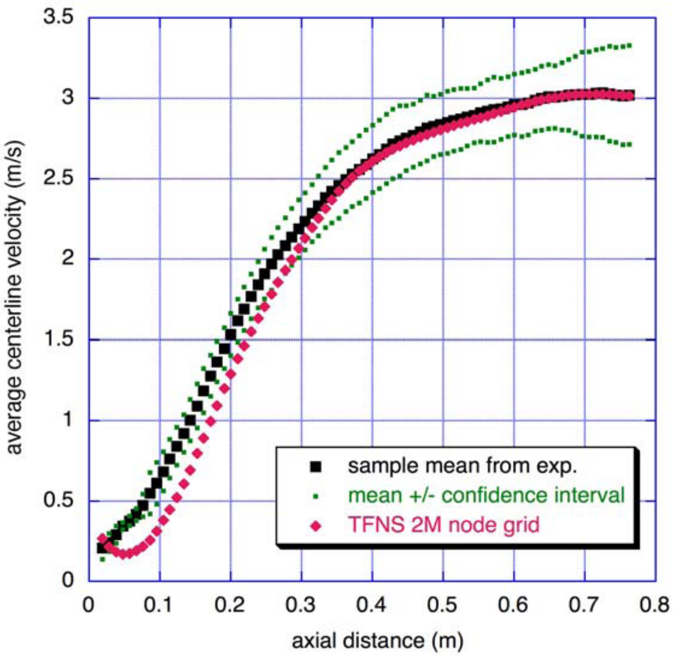
HOW ARE THESE ARTICLES PEER REVIEWED?

This is a challenge for the CS&E profession, not just V&V weenies!

Viewgraph norm to quantitative differences

Tieszen, et al (2005), “Validation of a Simple Turbulence Model Suitable for Closure of Temporally-Filtered Navier-Stokes Equations Using a Helium Plume,” SAND2005-3210.

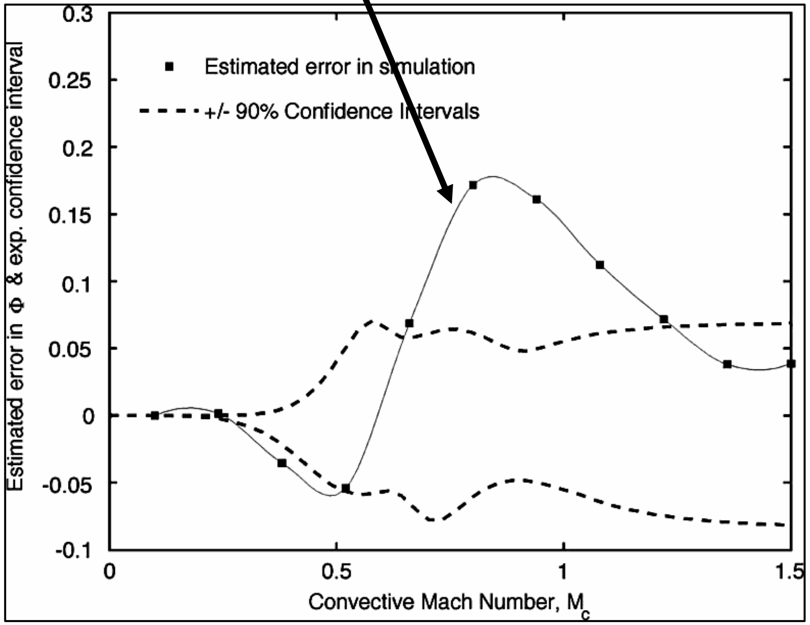




Validation is hard!
Validation Metrics

- Our formal engagement with this topic goes back to 2001.
- See Oberkampf and Barone (2006) for a recent summary of principles.
- "Metrics" are really metrics, but the general topic has to do with rigorous methods for quantifying *Diff* and drawing rigorous conclusions about predictive capability (per the ASC mission).
- Many benefits to thinking rigorously, not least of which is strong clarification of the difference between calibration and validation (Trucano et al, 2006), which is <u>especially</u> important in prediction.

Estimated computational error bars disjoint from experimental uncertainty confidence intervals – target for future validation experiments.



Credibility for WHAT? SSP aka QMU

Will our calculations be used properly?

- More than code developers and users need to believe ASC M&S and understand why it is believable.
- Best Estimate Plus Uncertainty
  - See Pilch et al (2006)
- How much V&V is enough?
  - "Sufficiency" → "Predictive Capability Maturity Model"
- Judgment replaced?
  - Do codes replace humans?
  - Are codes certified?

SAND2006-7046C

Decisions are hard!
Remember: ASC is "predictive" computational science.

DO PREDICTIVE M&S!

What does "predictive" mean?

Large-scale computational simulations supplement or replace physical experiments and tests for stockpile stewardship.

Use the science and experience of high-consequence system design/performance assessment in rigorous decision environments.

"Risk"-Informed Decision Making

Reliability Dimension

1. What can happen?
2. How likely is it?
3. What are the likely consequences if it does happen?

Confidence Dimension

4. What is your confidence in predicting the answers to the three questions?

Foundation = V&V

Decisions are hard!
What do we mean by "predictive"?

"Predictability" versus "Predictive Science" versus "Predictive Capability"

- <u>Predictability</u> – A technical concept, conventionally arising in the consideration of complex systems. I.e. as in "<u>predict</u> the stability of the solar system" or "<u>predict</u> the evolution of a chaotic system."
- <u>Predictive Science</u> – might just as well be a philosophical hope in the progress of the human condition. How do you measure it?
- <u>Predictive Capability</u> – in particular a <u>computational capability</u> with some (rigorous?) basis for credible interpolation or extrapolation of current knowledge, for example existing experimental data.

We (ASC) believe that "predictive capability" can be measured, although such capability is <u>always relative to the intended application</u>.

What do we mean by "predictive"?

- M&S typically informs <u>decision making under uncertainty</u>.
- ASC World: "Quantification of Margins and Uncertainty" (QMU)
  - Technical performance margins for engineered systems
  - Uncertainty in the underlying information and characterization of margins
  - Decisions required that reflect this uncertainty
- Many complex factors enter into using M&S in a complex technical endeavor, like Stockpile Stewardship (or climate warming policy).
- Our bottom line: Produce, communicate, and use M&S in the form of:

Best Estimate Plus Uncertainty

Example of BE+U: "Rivers of Blood"

- **Inflation projections from the Bank of England (February 2005 Inflation Report)**
- **Hendry: "Surprisingly, reporting of forecasts alone was the norm for the Bank, even until relatively recently; and it is still the norm among many forecasters." [Hendry and Ericsson, <u>Understanding Economic Forecasts</u>, MIT, 2001]**

Inflation Report: February 2005

Chart 6.3
Current CPI inflation projection based on market interest rate expectations

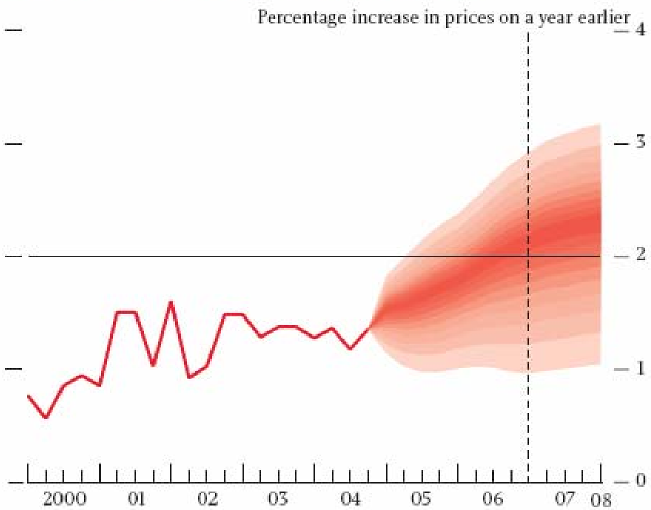
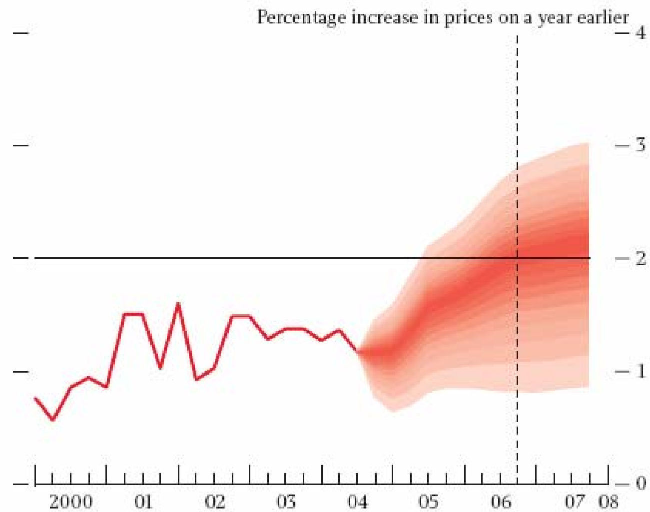
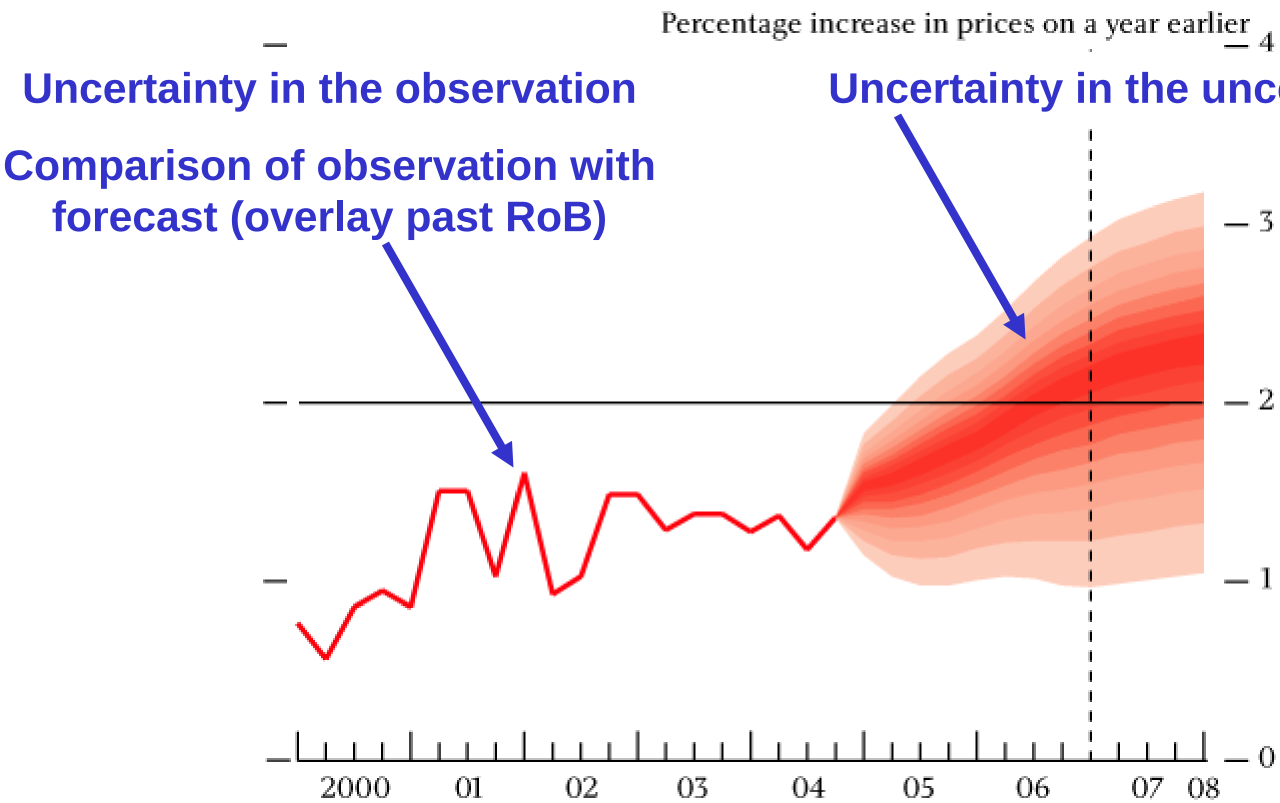


Chart 6.4
CPI inflation projection in November based on market interest rate expectations



“Rivers of Blood” – note missing elements

- Where is the comparison of observation with prediction?



Decisions are hard!
V&V Sufficiency – How much is enough?

Two options:

1. Keep going until you run out of money or until management can't take it anymore. **YUK!** ☹
2. Come up with a constructive basis for assessing sufficiency.

- The latter is inevitably tied to the application and the associated decisions, that is QMU.
- Sufficiency raises challenges of accumulation, communication and preservation of information.

SAND2006-7046C

Predictive Capability Maturity Model – LANL Style!

Predictive Attribute [SNL]	Lowest [Exploratory]	Low [Design]	High [Qualification]	Highest [Certified]
Simulation Geometry Fidelity	Cartoon-ish	Better	Better again	Solution time / resolution balance at will
Physics Model & Alg Fidelity	Ad-hoc	Calibrated / low fidelity	Physics-based / higher fidelity	Validated / algs span application
Model Integration in Code	Stand-alone	Limited testing	Achieved; validation	Robustness demonstrated
Solution Verification	Judgment only	Numerical sensitivities	Convergence, numerical errors quantified	Verification test suites created, automated
NTS Data Validation/QMU	Not applied	Variable calibration, less validation	Global calibration, validation	No ad-hoc calibrations, systematic val
Impact on codes, knob removal	Ignored	Makes little difference	Clear improvement	Designer adoption

Decisions → PCMM → Sufficiency

Decisions

Risk Tolerance

Predictive Attribute [SNL]	Lowest [Exploratory]	Low [Design]	High [Qualification]	Highest [Certified]
Simulation Geometry Fidelity	Cartoon-ish	Better	Better again	Solution time / resolution balance at will
Physics Model & Alg Fidelity	Ad-hoc	Calibrated / low fidelity	Physics-based / higher fidelity	Validated / algs span application
Model Integration in Code	Stand-alone	Limited testing	Achieved; validation	Robustness demonstrated
Solution Verification	Judgment only	Numerical sensitivities	Convergence, numerical errors quantified	Verification test suites created, automated
NTS Data Validation/QMU	Not applied	Variable calibration, less validation	Global calibration, validation	No ad-hoc calibrations, systematic val
Impact on codes, knob removal	Ignored	Makes little difference	Clear improvement	Designer adoption

Are we in the business of replacing humans or their judgment? NO!

- But what are the requirements on the codes for the SSP and the future evolution of the stockpile in a no-nuclear-test environment?
- Designers are "certified" from a variety of perspectives, from explicit training to the tacit knowledge they embody.
- No code will ever replace the explicit and tacit knowledge of a designer.
- But must future codes be certified?
- Codes <u>are</u> certified right now! – through designer use (and WILLINGNESS to use).

"Over my dead body …"

Hint: not a designer talking about using an ASC code.

Should we be constructing "Black Box" software systems?

If "Users" don't understand ASC codes they shouldn't be using them.

A National “Community of Practice”? (Beyond V&V)

- Where do we stand on defining and measuring the credibility of our work for important applications?

> **“Ground-water models cannot be validated.”**

- How do we define, use and empower benchmarks?

> **“My code passes more benchmarks than your code.”**

- Are “standards” needed? Appropriate? Frightening?

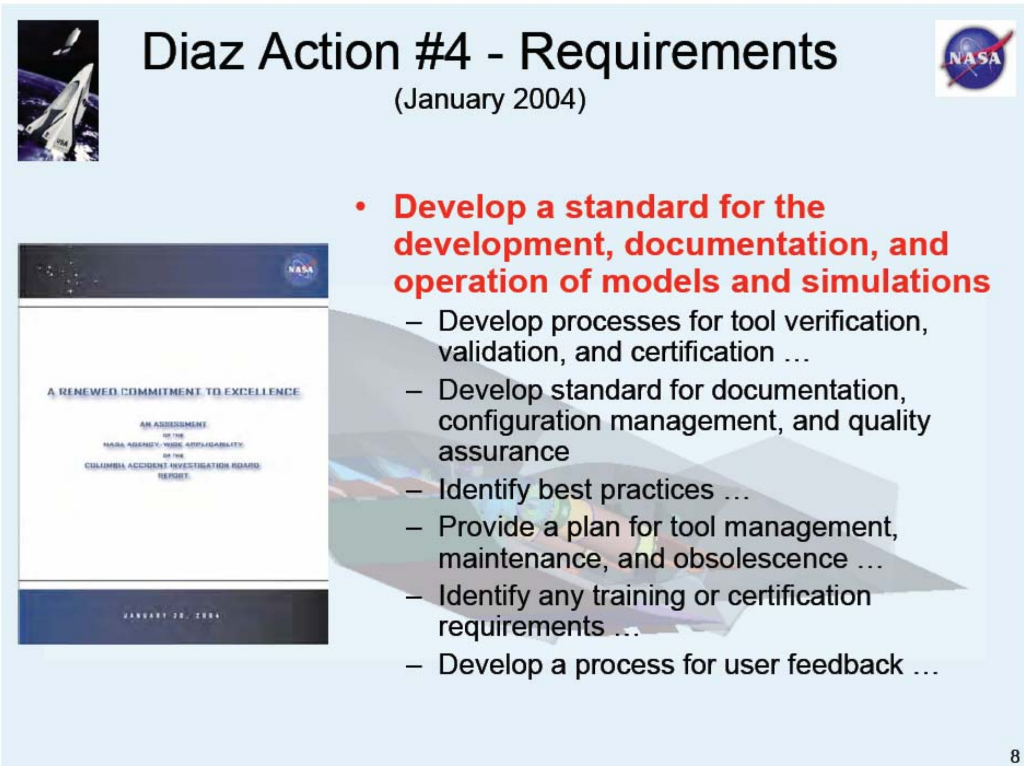
> **“I feel free to ignore those DMSO Guidelines.”**

- Are journals helping or hurting?

> **“Good enough for a journal does not imply good validation.”**

- Will we ever solve enough of the technical problems to make the above questions reasonable?
- How can education help?

Example:



> **"NASA and DOD could learn from ASC, NNSA."**

Mehta

My take-home lessons from the past 11 years of worrying about this are:

1. Comparing calculations with experimental data has no obvious impact on the problem of estimating numerical errors.
2. There is no obvious benefit to be gained by comparing with experimental data that have undefined uncertainty.
3. Code comparisons have no clear relationship to V&V.
4. Confusing calibration and validation is dangerous in prediction.
5. V&V is a risk-management component for high-consequence decision making under uncertainty.
6. Social elements are important.
7. Absence of evidence that something is wrong is not evidence that something is right.

"I've had sixteen fights. I won all but twelve of them..."

In conclusion:

V&V is a collaboration between code developers, experimenters, designers, people with specialized V&V knowledge (to the extent they exist) <u>and</u> decision makers.

Success or failure in V&V directly mirrors the success or failure of this collaboration.

"Too hard … Too slow … Too expensive …" ?

Key technical challenges:

- Numerical error quantification.
- Quantifying epistemic uncertainty.
- Validation metrics for high-complexity data sets (e.g. 4-D data depending on N uncertain parameters, N>>1, with aleatory and epistemic uncertainty separately accounted for).
- Measuring predictive capability and progress in achieving it for the important applications.
- Sufficiency – define and implement "What's good enough."
- What is the best way to collaborate with the user community, in particular beta-users.
- Supporting QMU, that is integrating decision concerns.
- Where is the computing going to come from?

Sandia National Laboratories

A few references

1. ASME (2006), "Guide for Verification and Validation in Computational Solid Mechanics," to be published.
2. D. Fuentes et al. (2006), "Extensions of Goal-Oriented Error Estimation Methods to simulations of highly-nonlinear response of shock-loaded elastomer-reinforced structures," Computational Methods in Applied Mechanics and Engineering, Volume 195, 4659-4680.
3. Klein, et al (2006), "ASC Predictive Science Academic Alliance Program Verification and Validation Whitepaper," UCRL-TR-220342-Rev, to be released.
4. National Science Foundation (2006), "Simulation-Based Engineering Science: Revolutionizing Engineering Science through Simulation," Report of the National Science Foundation Blue Ribbon Panel on Simulation-Based Engineering Science.
5. Oberkampf and Trucano (2002), "Verification and validation in computational fluid dynamics," Progress in Aerospace Sciences, Vol. 38, No. 3, 209-272. (Review)
6. Oberkampf, Trucano, and Hirsch (2004), "Verification, validation, and predictive capability in computational engineering and physics," Applied Mechanics Reviews, Vol. 57, No. 5, 345-384. (Review)
7. W. L. Oberkampf and M. F. Barone (2006), "Measures of Agreement Between Computation and Experiment: Validation Metrics," Journal of Computational Physics, Volume 217, 5-36.
8. Oberkampf and Trucano (2006), "Design of and Comparison with Verification and Validation Benchmarks," SAND2006-5376C (International Workshop on "The Benchmarking of CFD Codes for Application to Nuclear Reactor Safety"
9. Pilch, et al (2000), "Guidelines for Sandia ASCI Verification and Validation Plans – Content and Format: Version 2.0," SAND2000-3101.
10. Pilch et al (2006), "Ideas Underlying Quantification of Margins and Uncertainties (QMU): A White Paper," SAND2006-5001.
11. C. J. Roy (2005), "Review of Code and Solution Verification Procedures for Computational Simulation," Journal of Computational Physics, Volume 205, 131-156
12. Trucano et al (2001), "Description of the Sandia Validation Metrics Project," SAND2001-1339.
13. Trucano et al (2002), "General Concepts for Experimental Validation of ASCI Code Applications," SAND2002-0341.
14. Trucano et al (2003), "On the Role of Code Comparisons in Verification and Validation," SAND2003-2752.
15. Trucano et al (2006), "Calibration, validation, and sensitivity analysis: What's what," Reliability Engineering and System Safety, Volume 91, 1331-1357
16. Y. Yu et al. (2006), "Uncertainty Quantification for Chaotic Computational Fluid Dynamics," Journal of Computational Physics, Volume 217, 200-216.