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

NECDC 2006
LANL, October 23-27, 2006
SAND2006-6435C
Verification and Validation (V&V) Definitions

**Verification**: Are the equations solved correctly? (Math)

**Validation**: Are the equations correct? (Physics)

**ASC**:
- **Verification**: The process of determining that a model implementation accurately represents the developer’s conceptual description of the model and the solution to the model.
- **Validation**: 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.
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.
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.

1. DP Application
2. Planning
3. Code Verification
4. Experiment Design, Execution & Analysis
5. Metrics
6. Calculation Verification
7. Assessment
8. Prediction & Credibility

Verification

Validation Experiments

Validation Metrics

Credibility

Permanence

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.
- Two NECDC papers discuss this in detail for predictive electrical modeling:
  1. G. Gray et al, “Designing Dedicated Experiments to Support Validation and Calibration Activities for the Qualification of Weapon Electronics”
Planning

It’s all about the APPLICATION of the code(s):
• Application Requirements
• Phenomenology Identification and Ranking Table
• Priorities
• Identify focused verification requirements (supplementing code development focus)
• Identify hierarchical 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 is present.)

Numerical Error
“Calculation” Verification

Software Implementation
“Code” Verification

Algorithms

Mathematics

To believe any numerical error statement requires “code verification.”
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:
  - Convergence studies (highly empirical – can I take these to the bank?)
  - A posteriori error estimation (not in our favorite equations)
  - 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!

Verification is hard!

Convergence studies:

- Work at the state of the art for the NECDC community is illustrated by the LANL 2005 Level II verification milestone.
- Explored convergence error estimation both in asymptotic and non-asymptotic regimes.
- Heavily documented.
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.

😊
Verification is hard!
Probabilistic error models:

- Glimm and colleagues: treatment of numerical error as an uncertainty (incomplete knowledge).
- Probability used to quantify error models.
- 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

\[ \text{Diff} = \text{“Nature} - \text{Calculation”} \]

- 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), \textit{Forecast Verification}; Wilks (1995), \textit{Statistical Methods in the Atmospheric Sciences}
Validation is hard!
Have we detected a trend?

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

See 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!
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 $\text{Diff}$ 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 especially important in prediction.

Estimated numerical error 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 (but not Today!)
• Form mirrors function; see Pilch et al (2006)

How much V&V is enough?
• “Sufficiency” → “Predictive Capability Maturity Model”

Judgment replaced?
• Do codes replace designers?
• Are codes certified?
Decisions are hard!
Remember: ASC is “predictive” computational science.

What does “predictive” mean?

DO PREDICTIVE M&S!

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?

“Risk”-Informed Decision Making

Foundation = V&V

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

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

What does “predictive” mean?
Decisions are hard!
What do we mean by “predictive”?  

“Predictability” versus “Predictive Science” versus “Predictive Capability”

- **Predictability** – A technical concept, conventionally arising in the consideration of complex systems. I.e. as in “predict the stability of the solar system” or “predict the evolution of a chaotic system.”
- **Predictive Science** – might just as well be a philosophical hope in the progress of the human condition. How do you measure it?
- **Predictive Capability** – in particular a computational capability 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 always relative to the intended application.
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.
Predictive Capability Maturity Model – LANL Style!

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation Geometry Fidelity</td>
<td>Cartoon-ish</td>
<td>Better</td>
<td>Better again</td>
<td>Solution time / resolution balance at will</td>
</tr>
<tr>
<td>Physics Model &amp; Alg Fidelity</td>
<td>Ad-hoc</td>
<td>Calibrated / low fidelity</td>
<td>Physics-based / higher fidelity</td>
<td>Validated / algs span application</td>
</tr>
<tr>
<td>Model Integration in Code</td>
<td>Stand-alone</td>
<td>Limited testing</td>
<td>Achieved; validation</td>
<td>Robustness demonstrated</td>
</tr>
<tr>
<td>Solution Verification</td>
<td>Judgment only</td>
<td>Numerical sensitivities</td>
<td>Convergence, numerical errors quantified</td>
<td>Verification test suites created, automated</td>
</tr>
<tr>
<td>NTS Data Validation/QMU</td>
<td>Not applied</td>
<td>Variable calibration, less validation</td>
<td>Global calibration, validation</td>
<td>No ad-hoc calibrations, systematic val</td>
</tr>
<tr>
<td>Impact on codes, knob removal</td>
<td>Ignored</td>
<td>Makes little difference</td>
<td>Clear improvement</td>
<td>Designer adoption</td>
</tr>
</tbody>
</table>
Are we in the business of replacing designers 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 are certified right now! – through designer use (and WILLINGNESS to use).

“Over my dead body …”

Hint: not a designer talking about using an ASC code.
A National “Community of Practice”?

• 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?
Can PSAAP help?

We need to get what we know out in the national community (External Review Panel June 2006). ASC PSAAP has mandatory elements – predictive capability, collaboration with the Labs, … and V&V.

Aside from the opportunity to respond to the above elements, PSAAP provides educational opportunities in V&V/UQ/complex decision making that we should emphasize and take advantage of.
Example:

Mehta

**Diaz Action #4 - Requirements**
*(January 2004)*

- Develop a standard for the development, documentation, and operation of models and simulations
  - Develop processes for tool verification, validation, and certification ...
  - Develop standard for documentation, configuration management, and quality assurance
  - Identify best practices ...
  - Provide a plan for tool management, maintenance, and obsolescence ...
  - Identify any training or certification requirements ...
  - Develop a process for user feedback ...

“NASA and DOD could learn from ASC, NNSA.”
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) and decision makers.

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

“Too hard ... Too slow ... Too expensive ...” ?
Back Up Slides
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?
A few references

Rogue’s Gallery
Mystery Calculation #1

• “…the large-scale features are measured [experimentally] and they are well described [emphasis mine] by all three simulations.”
• Is it validation?
• Is it verification?

Quantitative growth of amplitude in time. No calculation errors presented.
Mystery Calculation #2

- Comparison with experiment:
  - No experimental error bars.
- Grid sensitivity studies reported but not quantified.
- No numerical error quantification.
- Commercial code used.

Figure 5. Pressure coefficient on the concave and convex walls of the duct x, + experiment, —RSM, —— standard \( k-\varepsilon \), —— realizable \( k-\varepsilon \), —— RNG \( k-\varepsilon \).
Mystery Calculation #3

- Comparison with experiment:
  - Experimental error bars.
- No reported convergence studies.
- No numerical error quantification.

“These numerical methodologies have already been validated … [references]. The accuracy of the numerical results has been demonstrated thanks to qualitative and quantitative comparisons between numerical simulations and experimental or analytical references.”
Mystery Calculation #4

- Comparison with experiment:
  - Experimental error bars.
- No reported convergence studies.
- No numerical error quantification (the different curves are different zones in the calculation).
- The actual purpose of the study was diagnostic development.
Decisions are hard!
This wisdom may not be much help…

How do we compare this…

…with THIS?

2-D r-z shell implosion calculation.

…and claim we know what we are doing?

Experimental spectroscopic image of Z-pinched liner stagnation.
Back to decisions – what do we mean by “predictive”?

- M&S typically informs decision making under uncertainty.
- ASC World: “Quantitative 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, Understanding Economic Forecasts, MIT, 2001]
“Rivers of Blood” – note missing elements

• Where is the comparison of observation with prediction?
Lemke example: 1-D MHD driven “flyers”

- ASC V&V milestones have allowed us to put together significant parts of this logic in productive ways (but still not completely!):
  - A 1-D validation study
  - Experimental data has error bars.
  - Numerical error in the calculation is quantified (not represented here)
  - The plot is more than a spaghetti plot.

\[
\Delta x = 21.25 \text{ µm}
\]
Lemke example: calibration = BE

- The DAKOTA optimization toolkit can be unleashed once you know what you are doing.

- A 1-D calibration
  - Experimental boundary data for the calculation is not measured accurately enough to improve the previous calculation

- This calibration is now viewed as a way of reducing the experimental boundary data uncertainty.
Lemke example: Uncertainty quantified

• “Numerical convergence: peak flyer velocity varies by < 0.1% with grid change of 10 to 5 µm.”

- Flyer temperature & density vs. X show significant change with dx
- Magnetic solve convergence tolerance varied by 4 orders of magnitude
- Sensitivity of flyer velocity and state to EOS: six different EOSs used for aluminum flyer
- Sensitivity of flyer velocity and state to EOS
- Sensitivity of flyer velocity and state to electrical conductivity model
- Radiation + thermal conductivity has no significant effect
Lemke example: Prediction – ZR Hardware Change Recommendation

- “Shaped pulse on ZR must be shortened, or must use thicker flyer and longer flight distance: ensures flyer survival; constant velocity; reduces performance.”

Standard pulse; 850 µm Al flyer; 5 mm flight distance

- Shaped pulse; 900 µm Al flyer; 6 mm flight distance