Credibility Assessment Frameworks for Empirical/Data Driven Models

- Personal Views -

Joshua Kaizer
Nuclear Reactor Regulation
Division of Safety Systems
Nuclear Performance and Code Review

May 15, 2019
Objective

To communicate:

• a brief summary of credibility assessment framework (CAF)

• application of a CAF with a specific empirical model

• proposal for a CAF for all types of empirical / data-driven models
credibility
\,\textipa{\,kred\-ə-\'bi-lə-tē}\noun

The determination that an object can be trusted for its intended purpose.
Credibility is the focus, but never seems to be focused on

- Seems to appear just “off the main stage” of the VVUQ community
- ASME V&V, VVUQ, V&V-10, V&V-20
- Verification and Validation in Scientific Computing
- Verification and Validation in Computational Science and Engineering
- Fundamentals of Verification and Validation
- Model Validation – Perspectives in Hydrological Sciences
- NASA-STD-7009 – Standard for Models and Simulations
- Assessing the Reliability of Complex Models

- Determining credibility of is the main goal of VVUQ
Driving over a bridge

It is safe to drive over the bridge.

G

G1
The bridge can withstand the weight of my car.

<table>
<thead>
<tr>
<th>Level</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Don’t think about it.</td>
</tr>
<tr>
<td>1</td>
<td>Someone has checked it.</td>
</tr>
<tr>
<td>2</td>
<td>I drove over it yesterday.</td>
</tr>
<tr>
<td>3</td>
<td>Someone will drive over it right before I will.</td>
</tr>
<tr>
<td>4</td>
<td>Engineering analysis of the bridge demonstrates a significant margin.</td>
</tr>
<tr>
<td>5</td>
<td>Current measurements of the bridge along with analysis of the bridge shows that it is structurally sound.</td>
</tr>
<tr>
<td>6</td>
<td>A recent bridge inspection showed that the bridge is structurally sound.</td>
</tr>
</tbody>
</table>

G2
There won’t be a natural disaster while I am driving over the bridge.

<table>
<thead>
<tr>
<th>Level</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Don’t think about it.</td>
</tr>
<tr>
<td>1</td>
<td>In my experience, natural disasters are rare.</td>
</tr>
<tr>
<td>2</td>
<td>There hasn’t been a natural disaster in this area for at least 20 years.</td>
</tr>
<tr>
<td>3</td>
<td>Bridge has been built to withstand any seismic activity. Weather forecast shows no adverse conditions expected.</td>
</tr>
</tbody>
</table>
Examples of credibility frameworks

<table>
<thead>
<tr>
<th>Framework</th>
<th>Criteria</th>
<th>Tiers</th>
<th>Evidence Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSAU (EMDAP)</td>
<td>20</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>PCMM</td>
<td>6</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>NASA-STD-7009</td>
<td>8</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>ASME V&amp;V 40</td>
<td>22</td>
<td>5</td>
<td>3-4</td>
</tr>
<tr>
<td>CBT CAF</td>
<td>33</td>
<td>4</td>
<td>[2,3,4,5,12]</td>
</tr>
</tbody>
</table>
Credibility Assessment Levels (NASA)

**Verification:**
Were the models implemented correctly, and what was the numerical error/uncertainty?

<table>
<thead>
<tr>
<th>Level</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Insufficient Evidence</td>
</tr>
<tr>
<td>1</td>
<td>Conceptual and mathematical models verified.</td>
</tr>
<tr>
<td>2</td>
<td>Unit and regression testing of key features.</td>
</tr>
<tr>
<td>3</td>
<td>Formal numerical error estimation.</td>
</tr>
<tr>
<td>4</td>
<td>Numerical errors small for all important features.</td>
</tr>
</tbody>
</table>

**Validation:**
Does the M&S results compare favorably to the referent data, and how close is the referent to the real-world system?

<table>
<thead>
<tr>
<th>Level</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Insufficient Evidence</td>
</tr>
<tr>
<td>1</td>
<td>Conceptual and mathematical models agree with simple referents.</td>
</tr>
<tr>
<td>2</td>
<td>Results agree with experimental data or other M&amp;S on unit problems.</td>
</tr>
<tr>
<td>3</td>
<td>Results agree with experimental data for problems of interest.</td>
</tr>
<tr>
<td>4</td>
<td>Results agree with real world data.</td>
</tr>
</tbody>
</table>
Application of a credibility assessment framework
Framework for critical boiling transition has been applied by the U.S. NRC

• 4 completed reviews
• 2 current reviews
• Framework is currently out for public comment
• Very positive feedback from all stakeholders (internal staff, applicants, oversight)
Three criteria for empirical/data-driven models

1. Model should contain the important physical parameters.

2. No credit is given to the mathematical form of the model
   – Physical intuition may be used, but is not credited
   – Only what and where data validates

3. Model is black box or glass box?
Physical models vs. Not physical models

Physical models

\[ F = m \cdot a, \quad Nu = A \cdot Re^B \cdot Pr^C, \quad f = A \cdot Re^B \]

Not Physical models

\[
q_{CBT} = \frac{(2.022 - 0.0004302 \cdot P) + (0.1722 - 0.0000984 \cdot P) \cdot \exp[(18.177 - 0.004129 \cdot P) \cdot x_c]}{0.44 \cdot \left(1 - x_c\right)^{7.9}}
\]

\[
q''(l_{c,N}) \cdot \left[1 - \exp\left(-0.44 \cdot \frac{(1 - x_c)^{7.9}}{(G/10^6)^{1.72}} \cdot l_{c,EU}\right)\right] \cdot \int_0^{l_{c,N}} q''(z) \cdot \exp\left[-0.44 \cdot \frac{(1 - x_c)^{7.9}}{(G/10^6)^{1.72}} \cdot (l_{c,N} - z)\right] \cdot dz
\]
Proposal for joint standard (V&V 30 and V&V 50)

Credibility Assessment Framework for Empirical/Data Driven Models

AN INTERNATIONAL STANDARD

The American Society of Mechanical Engineers

FOR ASME COMMITTEE USE ONLY
Credibility assessment framework for data-driven models
G - The data-driven model can be trusted for its intended purpose.

G1. The experimental data supporting the model are appropriate.

G2. The model was generated in a logical fashion.

G3. The model has sufficient validation as demonstrated through appropriate quantification of its error.
G1 - The experimental data supporting the model are appropriate.

G1.1. The experimental data have been collected at a credible test facility.

G1.2. The experimental data have been accurately measured.

G1.3. The experiment can reproduced the conditions in the reference system.
G1.1 - The experimental data have been collected at a credible test facility.

G1.1.1. The test facility is well understood.

G1.1.2. The test facility has been verified by comparison to an outside source.
## Evidence for G1.1.1

The test facility is well understood.

<table>
<thead>
<tr>
<th>Level</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A reference that describes the test facility in appropriate detail has been provided. At a minimum, the reference includes loop and test section.</td>
</tr>
<tr>
<td>2</td>
<td>The assessors have visited the test facility. Additionally, a reference that describes the test facility in appropriate detail has been provided. At a minimum, the reference includes loop and test section.</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
</tr>
</tbody>
</table>
Evidence for G1.1.2

<table>
<thead>
<tr>
<th>G1.1.1</th>
<th>The test facility has been verified by comparison to an outside source.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Evidence</td>
</tr>
<tr>
<td>1</td>
<td>The test facility has been verified by comparison of data obtained at the facility to some benchmarks or some known phenomenological behavior.</td>
</tr>
<tr>
<td>2</td>
<td>The test facility has been verified by comparison of data obtained from tests at the facility to data other than the QOI data obtained from a credible facility.</td>
</tr>
<tr>
<td>3</td>
<td>The test facility has been verified by comparison of QOI data obtained at the facility to QOI data obtained from a credible facility.</td>
</tr>
<tr>
<td>4</td>
<td>The test facility has been verified by comparison of QOI data obtained at the facility to QOI data obtained over the same application domain as that of the proposed model at a credible facility.</td>
</tr>
</tbody>
</table>
Proposal Summary

Credibility Assessment Framework for Empirical/Data Driven Models

ASME V&V 30-2020
ASME V&V 50-2020
Summary

• Leverage “Credibility Assessment Framework” format
  – Used by ASME V&V 40
  – Used by NUREG/KM-0013

• Generate a standard for assessing data driven (i.e., empirical) models

- Questions -
Backup Slides
What is critical boiling transition?

• When you boil too much
  – Too many bubbles crowd the surface, or
  – Or you run out of water

• Very complex physics
  – 2-phase flow, turbulence, phase change

• One type of a data-driven model
Interesting “Discoveries”

• Defining the validation domain of the data driven model is mathematically challenging
  – Multiple regions in validation domain, not easily defined by equations (especially in higher dimensions)

• Can’t prove the error is consistent over a domain, best you can do is prove no obviously “bad” regions

• Model should be “applied in the same manner as it was when it was validated”.

ASME 2019 V&V
Types of regions in a validation domain

(1) Well covered
(2) Local Hole
(3) Edge
(4) Isolated, Known Behavior
(5) Isolated, Unknown Behavior
(6) Extrapolation
Types of regions in a validation domain

- The error of a data driven model
  - Error is from a normal distribution
  - Error is **not** independent of location in domain.

- One region that has an unusual amount of “bad” error…
A Simple Credibility Assessment Frameworks
“GSN is a graphical argumentation notation…”

(GSN Standard, 2011)

Captures:

- **Goals**
- **Evidence**
- Assumptions
- Strategy
- Context…

It is safe to drive over the bridge.

G1
The bridge can withstand the weight of my car.

G2
There won’t be a natural disaster while I am driving over the bridge.
Currently proposed framework for data driven models
Main Goal (G)

The data-driven model can be trusted for its intended purpose.

G1 The experimental data supporting the model are appropriate.

G2 The model was generated in a logical fashion.

G3 The model has sufficient validation as demonstrated through appropriate quantification of its error.
Sub-Goal G1

G1
The experimental data supporting the model are appropriate.

G1.1
The experimental data have been collected at a credible test facility.

G1.2
The experimental data have been accurately measured.

G1.3
The experiment can reproduced the conditions in the reference system.
Sub-Goal G2

Main Goal (G)

The model was generated in a logical fashion.

- G2.1 The mathematical form of the model is appropriate.
- G2.2 The process for determining the model's coefficients is appropriate.
Sub-Goal G3

G3
The model has sufficient validation as demonstrated through appropriate quantification of its error.

G3.1 The correct validation error has been calculated.
G3.2 The validation error is appropriately distributed throughout the application domain.
G3.3 Any inconsistencies in the validation error have been accounted for appropriately.
G3.4 The model's uncertainty has been appropriately calculated from the validation error.
G3.5 The model has been correctly implemented.
Sub-Goal G1.1

The experimental data have been collected at a credible test facility.

G1.1.1 The test facility is well understood.

G1.1.2 The test facility has been verified by comparison to an outside source.

Sub-Goal G1

G1.1 The experimental data supporting the model are appropriate.

G1.2 The experimental data have been accurately measured.

G1.3 The experiment can reproduce the conditions in the real world system.
Sub-Goal G1.2

The experimental data have been accurately measured.

G1.2.1 The test facility has an appropriate quality assurance program.

G1.2.2 The experiment has been appropriately statistically designed (i.e., the value of a system parameter from any test was completely independent from its value in the test before and after the test).

G1.2.3 The method used to obtain the measured system response quantities (SRQ) results in an accurate measurement.

G1.2.4 The instrumentation uncertainties have been demonstrated to have a minimal impact on the measured SRQs.

G1.2.5 The uncertainty in the measured SRQs is quantified through repeated tests at the same state points.

G1.2.6 Any physical impacts that exist in the test section and not in the reference system should be addressed.
Sub-Goal G1.3

G1.3
The experiment can reproduce the conditions in the reference system.

G1.3.1
The test section used in the experiment should have geometric dimensions equivalent to those of the reference system.

G1.3.2
The components used in the test section should be prototypical of the components used in the reference system.

G1.3.3
The initial conditions of the test section should reflect the expected initial conditions of the reference system.

G1.3.4
The boundary conditions of the test section should reflect the expected initial conditions of the reference system.

G1.3.5
Any differences between the test section and the reference system should have a minimal impact on the measured SRQs.
Sub-Goal G2.1

G2.1
The mathematical form of the model is appropriate.

G2.1.1
The mathematical form of the model contains all the necessary parameters.

G2.1.2
The reasoning for choosing the mathematical form of the model should be discussed and should be logical.

G2 The model was generated in a logical fashion.

G2.1
The mathematical form of the model is appropriate.

G2.2
The process for determining the model's coefficients is appropriate.
Sub-Goal G2.2

The process for determining the model's coefficients was appropriate.

G2.2.1
The training data (i.e., the data used to generate the coefficients of the model) should be identified.

G2.2.2
The method for calculating the model's coefficients should be described.

G2.1
The mathematical form of the model is appropriate.

G2
The model was generated in a logical fashion.
Sub-Goal G3.2

The validation error is appropriately distributed throughout the application domain.

- **G3.2.1** The validation data (i.e., the data used to quantify the model's error) should be identified.
- **G3.2.2** The application domain of the model should be mathematically defined.
- **G3.2.3** The expected domain of the model should be understood.
- **G3.2.4** There should be adequate validation error data density throughout the expected and application domains.
- **G3.2.5** Sparse regions (i.e., regions of low data density) in the expected and application domains should be identified and justified.
- **G3.2.6** The model should be restricted to its application domain.
Sub-Goal G3.3

Any inconsistencies in the validation error have been accounted for appropriately.

G3.3.1 The validation error should be investigated to ensure that it does not contain any subgroups that are obviously not from the same population (i.e., non-poolable).

G3.3.2 The expected domain should be investigated to determine if it contains any non-conservative subregions that would impact the predictive capability of the model.

G3.3.3 The model’s predictions trend as expected in each of the various model parameters.
Sub-Goal G3.4

G3.4
The model’s uncertainty has been appropriately calculated from the validation error.

G3.4.1
The validation error statistics should be calculated from an appropriate database.

G3.4.2
The validation error statistics should be calculated using an appropriate method.

G3.4.3
The model’s uncertainty should be appropriately biased.
Sub-Goal G3.5

- **G3.5**
  - The model has been correctly implemented.

  - **G3.5.1**
    - The model has been implemented in the same computer code that was used to generate the validation error.

  - **G3.5.2**
    - The model's prediction of the measured SRQ is being applied using the same evaluation methodology used to predict the validation data set for determining the validation error.

  - **G3.5.3**
    - The model results in an accurate or conservative prediction when it is applied in practice.

Sub-Goal G3

- **G3.1**
  - The correct validation error has been calculated.

- **G3.2**
  - The validated error is propagated through the application system.

- **G3.3**
  - Key uncertainties in the validation data set are accounted for appropriately.

- **G3.4**
  - Key model's uncertainty has been appropriately calculated from the validation error.

- **G3.5**
  - The model has been correctly implemented.
# Evidence Table for Identifying Non-Conservative subregions

<table>
<thead>
<tr>
<th></th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>G3.3.2</strong></td>
<td>The expected domain should be investigated to determine if it contains any non-conservative subregions that would impact the predictive capability of the model.</td>
</tr>
<tr>
<td><strong>Level</strong></td>
<td><strong>Evidence</strong></td>
</tr>
<tr>
<td>1</td>
<td>Plots of each model input parameter versus the validation error (i.e., predicted versus measured or measured versus predicted) are provided. This visual method (e.g., the 1-D method) demonstrates that there are no trends in the validation error with any input parameter.</td>
</tr>
<tr>
<td>2</td>
<td>Plots of each model input parameter versus the validation error (i.e., predicted over measured or measured over predicted) are provided. This visual method (e.g., the 1-D method) demonstrates that there are no trends in the validation error with any input parameter. Additionally, a method similar to the one proposed by Kaizer (2015) is used to demonstrate that there are no obvious non-conservative subregions in the application domain.</td>
</tr>
<tr>
<td>3</td>
<td>A method further refined from the one proposed by Kaizer (2015) is used. Such a method is able to consider all N-dimensions at the same time and does not call for the user to visually identify any suspected non-conservative subregions.</td>
</tr>
<tr>
<td>CRITERIA</td>
<td>LEVELS</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Model Risk - Model influence</td>
<td>3</td>
</tr>
<tr>
<td>Model Risk – Decision consequence</td>
<td>3</td>
</tr>
<tr>
<td>Verification - Software quality assurance</td>
<td>3</td>
</tr>
<tr>
<td>Verification - Numerical code verification</td>
<td>4</td>
</tr>
<tr>
<td>Verification - Discretization error</td>
<td>3</td>
</tr>
<tr>
<td>Verification - Numerical solver error</td>
<td>3</td>
</tr>
<tr>
<td>Verification - Use error</td>
<td>4</td>
</tr>
<tr>
<td>Validation - Model form</td>
<td>3</td>
</tr>
<tr>
<td>Validation - Model inputs (2)</td>
<td>3</td>
</tr>
<tr>
<td>Validation - Test samples (4)</td>
<td>3</td>
</tr>
<tr>
<td>Validation - Test conditions</td>
<td>3</td>
</tr>
<tr>
<td>Validation - Equivalency of input parameters</td>
<td>3</td>
</tr>
<tr>
<td>Validation - Output comparison (4)</td>
<td>2</td>
</tr>
<tr>
<td>Applicability - Relevance of the quantities of interest</td>
<td>3</td>
</tr>
<tr>
<td>Applicability - Relevance of the validation activities to the COU</td>
<td>4</td>
</tr>
</tbody>
</table>