Data Credibility in Scientific Machine Learning

Kyle Neal, Erin Acquesta, Blake Lance, Bill Rider, Matt Barone, Uma Balakrishnan

ASME VVUQ Symposium
Texas A&M University, College Station, TX
May 25-26, 2022
Defining credibility
- the quality of being trusted and believed in [Oxford Dictionary]

Credibility considerations within engineering.

Typical scenario

• Question: what is the max stress? What is the remaining useful life?
• Answer: 1) perform analysis that leverages experiments and CompSim, 2) make a prediction
• Can the prediction be trusted? What is our confidence?
  • Credibility: assert whether conclusions/results are credible enough to support a specific decision

Credibility is garnered through collection and documentation of credibility evidence
What is credibility evidence?

Without a formal definition, it can really be quite subjective…

Expert judgement, I have been doing this for 50 years!

The deliverable is due today, so it better be credible!

I ran the highest fidelity simulation on the best and biggest computer out there!

We built conservatism and plenty of margin into all of our calculations!

We used the same process we have always used, we have never been wrong before!

We’re considering credibility evidence in context of data for SciML

Credit: Aubrey Eckert, Josh Mullins
Outline

- Why credibility is needed in SciML
  - Introduce credibility concepts from CompSim

- Propose a Dataset Credibility Framework
  - Datasheets for Datasets
  - EPICC
  - PIRT

- Turbulence Case Study
Why does Credibility Matter for Machine Learning?

Disney's deepfakes are getting closer to a big-screen debut

Google's Fabricius uses machine learning to decode hieroglyphs

CMU and Facebook AI Research use machine learning to teach robots to navigate by recognizing objects

Deep learning enables early detection and classification of live bacteria using holography

Tencent and Chinese scientists use deep learning to predict fatal COVID-19 cases

Machine learning helps robot swarms coordinate

The Use of Artificial Intelligence in Healthcare Accelerated During the Pandemic. It's Here to Stay.

AI is raising the standards of population health, ultimately making it easier for doctors to make more informed decisions as they come up with ...
Motivating Credibility for Scientific Machine Learning (SciML)

Scientific Machine Learning (SciML)
Machine learned models are used in lieu of, complementary to, or as surrogates for science and engineering computational simulation models.

Applications:
- Classification - Component Health Assessments
- Regression - Data-Driven Turbulence RANS Closures

Challenges:
- Tracking Data Credibility and Provenance
- Code & Solution Verification
- Validation for Use Cases
- Identification, Aggregation, and Quantification of All Sources of Uncertainty

Credit: Aubrey Eckert, Josh Mullins
Credibility Identification through Maturity Models

Capability Maturity Model Integration, Carnegie Mellon University

“Originally created for the U.S. Department of Defense to assess the quality and capability of their software contractors, ISACA's CMMI models have expanded beyond software engineering to help organizations around the world, in any industry, understand their current level of capability and performance and offer a guide to optimize business results.”

(2006) Software Engineering Institute** at CMU

- CMMI: Capability Maturity Model Integration
- MDDAP: Medical Device Discovery Appraisal Program
- CMMI Cybermaturity Platform
- DMM: Data Management Maturity
- People Capability Model

PCMM: Predictive Capability Maturity Model at Sandia [Oberkampf et al. 2007]

Adapted from CMMI to provide a maturity model that can be applied to our CompSim models that are intended to be used in high-consequence decision-making environments.

The computational simulation (CompSim) **credibility process** assembles and documents **evidence** to ascertain and communicate the **believability** of **predictions** that are produced from computational simulations.

Credit: Aubrey Eckert, Josh Mullins

https://www.cmmiinstitute.com
https://www.sei.cmu.edu
Suggested Translation to Credibility for SciML

**Noting for SciML:**
- We prioritize data representation over geometric fidelity.
- Highlight domain-aware (e.g., physics-informed) training to align with terminology used by SciML model developers.

**Data Representation/Credibility**
- Does the data provide a representative population for training/testing/validating?

**Domain-Aware**
- What physical phenomena needs to be preserved in the model? Is it captured?

**Code Verification**
- Software Quality Assurance

**Solution Verification**
- What numerical errors impact the predictions?

**Validation**
- Do the model predictions agree with the ground truth data that was not used during training?

**Uncertainty Quantification**
- What sources of uncertainties are irreducible (aleatory)
- What sources of uncertainties are reducible (epistemic)

---

**Noting for SciML:**
- We prioritize data representation over geometric fidelity.
- Highlight domain-aware (e.g., physics-informed) training to align with terminology used by SciML model developers.
Dataset Credibility Framework
Credibility is Predicated on the Quality of the Training Data
Terminology for Assessing Data Quality

- **Data provenance** – the origin of data and the process by which it arrived at the database

- **Data lineage** – the line of descent. How the data evolves (including transformations/processing) over time

- **Data pedigree** – includes the idea of quality in addition to origin and line of descent

- **Data integrity** – the maintenance of, and the assurance of, data accuracy and consistency over its entire life-cycle

- **Data credibility** –
  - a measure of how much trust you can put in the data you have
  - a level of certitude that the data content corresponds to a real object or has been obtained using a proper acquisition method (i.e., data quality)

---

https://www.igi-global.com/dictionary/data-quality-assessment/6708
Dataset Credibility Framework

- Integrated framework
- Prioritize
  - Documentation
  - Identifying errors/gaps
  - Planning
- Ideally included before data is collected/generated
  - But can be used post facto

---

Dataset Credibility Framework (v.1.0)

The purpose of this framework is to support the credibility assessment of datasets used in scientific machine learning (SciML). It is comprised of the following 3 elements:

<table>
<thead>
<tr>
<th>Datasheet</th>
<th>Through recording the metadata of a dataset, this document provides 1) a quick summary of the dataset, 2) a formal and central archival process, and 3) the ability for information about the dataset to be updated over time.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment Planning for Integrating Credibility and CompSim (EPICC)</td>
<td>This framework is used for analyzing (and planning if possible) physical experiments that generate a dataset. EPICC is adapted from computational simulation (CompSim) credibility assessment.</td>
</tr>
<tr>
<td>The Phenomena Identification and Ranking Table (PIRT)</td>
<td>This process identifies and prioritizes physical phenomenon in an engineering analysis. It is used here to inform the credibility of computational simulations (CompSim) that are used to generate training datasets.</td>
</tr>
</tbody>
</table>

Questions or Comments: Kyle Neal (kNeal@sandia.gov), Erin A. Acuesta (eacuesta@sandia.gov)

Sandia National Laboratories
Datasheets for Datasets

Standardized method, proposed by researcher collaborations from Google, Georgia Tech, Cornell, Microsoft Research, University of Maryland, and AI Now Institute, to report on credibility of the training & testing data used for machine learning applications.

<table>
<thead>
<tr>
<th>Motivation</th>
<th>Clearly articulate the reasons for creating the dataset and to promote transparency about funding interests.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composition</td>
<td>Provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks.</td>
</tr>
<tr>
<td>Collection Process</td>
<td>Provide information that allow others to reconstruct the dataset without access to it.</td>
</tr>
<tr>
<td>Preprocessing/Cleaning/Labeling</td>
<td>Provide dataset consumers with the information they need to determine whether the “raw” data has been processed in ways that are compatible with their chosen tasks.</td>
</tr>
<tr>
<td>Uses</td>
<td>Dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.</td>
</tr>
<tr>
<td>Distribution</td>
<td>Prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.</td>
</tr>
<tr>
<td>Maintenance</td>
<td>Plan for dataset maintenance and communicate this plan with dataset consumers.</td>
</tr>
</tbody>
</table>

EPICC: Experiment Planning for Integrating Credibility and CompSim

Why:
- Teams are becoming more integrated, schedules are being compressed
- We need efficient and effective communication
- CompSim often uses experimental results
- Validation, calibration, material characterization
- A comprehensive treatment of CompSim credibility would involve experiment credibility

What: Framework to...
- guide planning of comprehensive experiments for CompSim with integrated teams of experimentalists and analysts
- facilitate communication with integrated teams
- evaluate and communicate credibility evidence of experiments

Credit: Blake Lance & Sarah Kieweg
Physics and Material Model Fidelity – “Are the important physics models adequate?"

° The process of characterizing modeling completeness and adequacy for intended application.

➢ Tool: PIRT

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>Importance</th>
<th>Adequacy for Intended Use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Math Model</td>
<td>Code</td>
</tr>
<tr>
<td>Phenomena 1</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>Phenomena 2</td>
<td>M</td>
<td>H</td>
</tr>
<tr>
<td>Phenomena 3</td>
<td>L</td>
<td>H</td>
</tr>
</tbody>
</table>

Phenomena Identification and Ranking Table (PIRT) often used at Sandia to support this element.

- Define **key physical phenomena** and rank their importance for a particular quantity of interest
- Importance is relative to **quantity of interest** in the application scenario
- Assess **adequacy** and **gaps** in simulation capabilities and available data
- Adequacy of capabilities is relative to **intended use**
- Gaps are identified when adequacy scoring is below importance ranking

Wilson, Gary E., and Brent E. Boyack. “The role of the PIRT process in experiments, code development and code applications associated with reactor safety analysis.”
# Summary of Data Credibility Tools and Limitations

<table>
<thead>
<tr>
<th>Data Credibility Tools</th>
<th>PIRT</th>
<th>EPICC</th>
<th>Datasheets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machine Learning</strong></td>
<td>≈</td>
<td>≈</td>
<td>√</td>
</tr>
<tr>
<td><strong>Data Sources</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CompSim</td>
<td>✔</td>
<td>✗</td>
<td>≈</td>
</tr>
<tr>
<td>Experiment Data</td>
<td>✗</td>
<td>✔</td>
<td>≈</td>
</tr>
<tr>
<td>Data Mining</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Remote Sensing</td>
<td>✗</td>
<td>✗</td>
<td>≈</td>
</tr>
<tr>
<td>Multi-Fidelity</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
</tr>
</tbody>
</table>

- ✔ - Tool developed for this application
- ≈ - Tool could extend to this application
- ✗ - Tool is not designed for this application
Case Study: Turbulence Dataset
Turbulence Background

Turbulence occurs at high values of Reynold’s Number ($Re$)
- High inertial force
- Low viscous force

Varying time and spatial scales make simulation/modeling a challenge

\[ Re = \frac{\rho V D}{\mu} = \text{Inertial Force} \div \text{Viscous Force} \]

- Reynolds Averaged Navier Stokes (RANS)
- Large Eddy Simulations (LES)
- Direct Numerical Simulation (DNS)

Accuracy

Computational Cost

Modeling
Capture the effect of smaller eddies by increasing viscosity

Simulation
Resolve all eddies

https://www.youtube.com/watch?v=tlefAxicsGY
https://www.youtube.com/watch?v=G32LXeCx7H0
Summary of Recent Work

• Goal: assess the quality of a turbulence datasets for constructing data-driven turbulence closures

• Feature selection
  • Identify optimal features that will divide dataset into distinct clusters
  • Candidate features are the statistics typically provided by turbulence simulation runs

• Clustering to capture different physical regimes
  • “Some clusters are closely identified with the anisotropy state of the turbulence, whereas others can be connected to physical phenomena, such as boundary-layer separation and free shear layers”

• Datasets of different types of flow
  • Plane channel, wavy-walled channel, a bump in a channel, and a square cylinder

Investigate Data Provenance

We lose data provenance if we don’t document early.
Current Work: Applying Dataset Credibility Framework

- Documenting the dataset (via datasheet)
- Although data was generated from simulations, we do not have access to conduct PIRT analysis

**Training Cases:**
- Duct_formatted_CEVM.dat (10201, 50) (Duct flow at Re=3500 (Pinelli et al. 2010))
- AnthonyIC_formatted_CEVM3.dat (179487, 50), AnthonyIC_slice_wSGS2.dat (5292, 50) (Perpendicular jet-in-cross-flow (Ruiz et al. 2015))
- channel_sept2016_formatted.dat (59, 50), channel_sept2016_formatted_corrected.dat (59, 50) (One Reynolds number; Channel flow at Re=590 (Moser et al. 1999))
- JIC_interp_formatted_downsample1000_CEVM2.dat (39106, 50) (Inclined jet-in-cross-flow (Ling et al. 2016))
- Square_Cyl_formatted_CEVM2.dat (5777, 50) (Square cylinder (Ray et al. 2014))
- ConvDiv_formatted_CEVM3.dat (69189, 50) (Converging-diverging channel (Marquillie et al. 2011))

**Validation Cases:**
- CubeFlow_formatted_downsample30_CEVM.dat (18147, 50) (Wall-mounted cube at Re=5000 (Rossi et al. 2010))

**Testing Cases:**
- Duct_NN_ke_Re2000_formatted.dat (2744, 50) (Duct flow at Re=2000)
- Wavy_Wall_slice_CEVM.dat (24929, 50) (Flow over a wavy wall at Re=6850)
- SIJC_formatted_CEVM2.dat (7056, 50) (UNKNOWN CASE AND SOURCE UNKNOWN)

### Channel Features (2D)

<table>
<thead>
<tr>
<th>Columns</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yplus (non-dimensional quantity of y)</td>
</tr>
<tr>
<td>2</td>
<td>Barycentric coordinate, x</td>
</tr>
<tr>
<td>3</td>
<td>Barycentric coordinate, y</td>
</tr>
<tr>
<td>4</td>
<td>Barycentric coordinate, c1</td>
</tr>
<tr>
<td>5</td>
<td>Barycentric coordinate, c2</td>
</tr>
<tr>
<td>6</td>
<td>Barycentric coordinate, c3</td>
</tr>
<tr>
<td>7</td>
<td>S_tau (theta; angle between tensors and vectors)</td>
</tr>
<tr>
<td>8</td>
<td>S_tau (phi)</td>
</tr>
<tr>
<td>9</td>
<td>S_tau (zeta)</td>
</tr>
<tr>
<td>10</td>
<td>lambda_1 (= -lambda_2) (scalar invariants of the strain rate tensor, S, and rotation rate tensor, W)</td>
</tr>
<tr>
<td>11</td>
<td>lambda_5 (lambda_3 = lambda_4 = 0)</td>
</tr>
<tr>
<td>12</td>
<td>eta_1 (additional invariants that involve anisotropy tensor)</td>
</tr>
<tr>
<td>13</td>
<td>eta_2</td>
</tr>
<tr>
<td>14</td>
<td>eta_3</td>
</tr>
<tr>
<td>15</td>
<td>eta_4</td>
</tr>
<tr>
<td>16</td>
<td>eta_5 (all invariants contain information on the local stress-strain relationship in a turbulent flow)</td>
</tr>
<tr>
<td>17</td>
<td>TKE (ratio of turbulent production to dissipation rate)</td>
</tr>
<tr>
<td>18</td>
<td>U (mean velocity)</td>
</tr>
<tr>
<td>19</td>
<td>Case index (1-5, denotes Reynolds number)</td>
</tr>
</tbody>
</table>
Conclusions

• Credibility for SciML discussed
• Dataset Credibility Framework proposed
  • Combine Datasheets, EPICC, & PIRT
• Preliminary demonstration on Turbulence case study

Gaps and Future Considerations:
• Multi-fidelity datasets
• CompSim generated data, haven’t deployed PIRT yet