Uncertainty Quantification Framework
for Modeling Prediction

Michael Frenklach

Collaborators:

– Andy Packard
– W. A. Lester, Jr. (DFT)
– P. Westmoreland (ALS)
– N. Slavinskaya (SynGas)

Supported by: NSF, AFOSR, DOE-NNSA (PSAAP II)
• Introduction: UQ-predictive modeling
• Bound-To-Bound Data Collaboration
• Introductory case: Energetics of water clusters
• Full-blown case: Combustion of natural gas
The key challenge: Prediction
“Model predicts reasonably well the experimental behavior”
“Model matches the experimental data”
“...excellent agreement between model and data.”
“The model predictions match reasonably well the experimental data”

“Model predicts data” ?

“Model falls short in predicting experimental data”
“The prediction matches very well with experimental data”

“Simulation agrees well with the data”
“The model well predicts the data”
“Good agreement was found between the model and the data”
When All Models Are Wrong

More stringent quality criteria are needed for models used at the science/policy interface, and here is a checklist to aid in the responsible development and use of models.
We develop a temporal modeling framework adapted from physics and signal processing ...

The results ... indicate that by using our framework .... we can achieve significant improvements in prediction compare to baseline models ...
• Predictive → UQ-Predictive

• Physics-based models with the focus on data

• Validation is part of the process

• Dimensionality reduction is part of the process

• Practicality → use of surrogate models (Emulators)

• Data/Models
  • Access, sharing, documentation, ...
  • Reproducibility
model
(dif eq, nonlinear)
model
(dif eq, nonlinear)
**Bound-to-Bound Data Collaboration (B2B-DC)**

- an optimization-based framework for combining models and data to ascertain the collective information content

experimental uncertainties

prior knowledge on parameters
INTRODUCTORY CASE:

PREDICT IONIZATION POTENTIAL OF WATER CLUSTERS

**Quantum-Chemistry Approaches**

- **empirical**: force-field – guessed potential, empirically fitted; ...
- **semi-empirical HF**: quantum “core” with some terms replaced by parameters fitted to data (AM1, RM1, PM3, PM6, ZINDO, ...)
- **DFT with fitted parameters**: meta-GGA (Truhlar, M05, M06, M11, ...), double-hybrid DFT (Grimme), ...

\[ E_{XC} = (1 - \alpha_x)E_X^{DFT} + \alpha_x E_X^{HF} + (1 - \alpha_C)E_C^{DFT} + \alpha_C E_C^{MP2} \]

- “static” outcome: the optimized model needs (constant) retuning
- **the optimum is not unique**!
- partial loss of information (two-step process)
B2B-DC

Model: \[ E_{XC} = (1 - \alpha) E_{X}^{\text{GGA}} + \alpha E_{X}^{\text{HF}} + (1 - \beta) E_{C}^{\text{GGA}} + \beta E_{C}^{\text{MP2}} \]

Data

Solve for:

use \( \Delta E \) intervals computed for dimer, trimer, tetramer, and pentamer to predict \( \Delta E \) interval of hexamer
\[ E_{XC} = (1 - \alpha) E_X^{\text{GGA}} + \alpha E_X^{\text{HF}} + (1 - \beta) E_C^{\text{GGA}} + \beta E_C^{\text{MP2}} \]
FULL-BLOWN CASE:

COMBUSTION CHEMISTRY OF NATURAL GAS

• mixture of mostly methane with other light gases

• lowest emissions among fossil fuels; no soot; smallest carbon footprint

• various, expanding sources (biofuels, artificial synthesis,...)

• plenty and cheap; booming US (and world) economy

• technology issues/needs
  • varying compositions – hard to categorize empirically
  • prediction needs: emissions, combustion efficiency, ...
Methane Combustion: \[ \text{CH}_4 + 2 \text{ O}_2 \rightarrow \text{CO}_2 + 2 \text{ H}_2\text{O} \]

Foundation

• A physically-based model

• The network is complex, but the governing equations (rate laws) are known

• Uncertainty exists, but much is known where the uncertainty lies (rate parameters)

• Numerical simulations with parameters fixed to certain values may be performed “reliably”

• There is an accumulating experimental portfolio on the system

• The model is reduced in size for applications
PREDICTION: ignition delay in HCCI engine

- Flow-reactor measurements
- Laboratory flame measurements
- Theoretical rate constants
Experiment, $E$

$L \leq y \leq U$

Dataset unit $U_e = (U_e, L_e, M_e)$

Model, $M(x)$

Feasible set of $x$, $F$

If empty, inconsistent, otherwise, consistent
experiment/theory constrain feasible set

prior knowledge

$M(x_1, x_2)$

bounds on $x_1$
a realistic feasible set:

a set of individual uncertainties does not represent the true compound uncertainty
**Surrogate Models (Emulators)**

- build surrogate models for *individual responses* $y$ (rather than for overall objective)
- construct global objective from individual responses (higher fidelity)

\[
\Phi = \sum_{\text{all responses}} w \left( y_{\text{computed}} - y_{\text{observed}} \right)^2 \rightarrow \min_x
\]

\[
y(\{x\}) \approx a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_{1,2} x_1 x_2 + \ldots + a_{1,1} x_1^2 + a_{2,2} x_2^2 + \ldots
\]

*surrogate model*
• build surrogate models for individual responses (rather than for overall objective)
• construct global objective from individual responses (higher fidelity)

\[ \Phi = \sum_{\text{all responses}} w \left( y_{\text{computed}} - y_{\text{observed}} \right)^2 \rightarrow \min_x \]

\[ y(\{x\}) \approx a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_{1,2} x_1 x_2 + \ldots + a_{1,1} x_1^2 + a_{2,2} x_2^2 + \ldots \]

surrogate model
Surrogate Models (Emulators)

- build surrogate models for individual responses (rather than for overall objective)
- construct global objective from individual responses (higher fidelity)

\[ \Phi = \sum_{\text{all responses}} w \left( y_{\text{computed}} - y_{\text{observed}} \right)^2 \rightarrow \min \]

**Dimensionality Reduction**

- Flame speed = \( P_2(x_1, x_2, x_3, x_4, \ldots) \)
- Ignition delay = \( P_2(x_1, x_4, x_5, x_{17}, \ldots) \)
- Species conc. = \( P_2(x_3, x_4, x_5, x_{12}, \ldots) \)

Dimensionality of individual response

Dimensionality of optimization
Uncertainty is constrained by:

- prior knowledge of parameters, \( x \in \mathcal{H} \), the "\( \mathcal{H} \) cube"
- observed data/models, \( M(x) \in \mathcal{D} \), the "\( \mathcal{D} \) cube"

Prediction model: \( f(x) \)

- establish possible range of \( f(x) \), constrained by

\[
\begin{bmatrix}
\min_{x \in \mathcal{H}, M(x) \in \mathcal{D}} f(x) & \max_{x \in \mathcal{H}, M(x) \in \mathcal{D}} f(x)
\end{bmatrix}
\]

If \( f \) and \( M \) are quadratic, then the min and max problems \( \rightarrow \) SDP

and \( p \)'s and \( r \)'s bounds are

- computable
- easily verified as valid
- same for their global sensitivities
A dataset is **consistent** if the Feasible Set is nonempty; i.e., there exists a parameter vector that satisfies:

- all parameters are within prior bounds, $\mathcal{H}$
  
  \[
  x_{1,\text{min}} \leq x_1 \leq x_{1,\text{max}} \\
  x_{2,\text{min}} \leq x_2 \leq x_{2,\text{max}} \\
  \ldots
  \]

- all model predictions are within experimental bounds

  \[
  L_e \leq M_e(x) \leq U_e
  \]

- numerical measure of consistency

  \[
  C_D = \max_{x \in \mathcal{H}} \gamma \\
  L_e(1-\gamma) \leq M_e(x) \leq U_e(1-\gamma), \quad \forall e
  \]
**CONSISTENCY MEASURE**

![Feasible Set](image1)

![Consistent](image2)

DISCRIMINATION AMONG MODELS

Wiesner et al. 1996
27 active variables

Lemon et al. 2003
34 active variables

SENSITIVITY COEFFICIENTS

\[
\frac{\partial \left( \frac{\text{prediction}}{\text{prediction}} \right)}{\partial \left( \frac{\text{prediction}}{\text{uncertainty}} \right)}
\]

\[
\frac{\partial \left( \frac{\text{prediction}}{\text{uncertainty}} \right)}{\partial \left( \frac{\text{prediction}}{\text{uncertainty}} \right)}
\]
\[ \lambda := \frac{\partial \left( \text{prediction} \atop \text{interval} \atop \text{experiment} \atop \text{uncertainty} \right)}{\partial} = \frac{1}{2} \left( \frac{\partial \overline{M}}{\partial U} - \frac{\partial \overline{M}}{\partial L} + \frac{\partial \overline{M}}{\partial U} - \frac{\partial \overline{M}}{\partial L} \right) \]

\[ v := \frac{\partial \left( \text{prediction} \atop \text{interval} \atop \text{parameter} \atop \text{uncertainty} \right)}{\partial} = \frac{1}{2} \left( \frac{\partial \overline{M}}{\partial x_{\max}} - \frac{\partial \overline{M}}{\partial x_{\min}} + \frac{\partial M}{\partial x_{\max}} - \frac{\partial M}{\partial x_{\min}} \right) \]
Sensitivity of uncertainty in predicting $Y_i$ to uncertainty in observing $Y_j$. 
sensitivity
“w.r.t. value” vs “w.r.t. uncertainty”

laminar flame speed in a stoichiometric atmospheric C$_2$H$_6$-air mixture

sensitivity of methane dataset consistency
to uncertainty in model parameters

prediction on the feasible set

$M_p(x_1, x_2)$

Initial prediction from prior info

Final prediction
prediction interval

is the range of values $M_p$ takes over the set of feasible values of parameters

$$= \overline{M}_p(x) - \underline{M}_p(x)$$

$$\overline{M}_p(x) = \max_x M_p(x)$$

$$\underline{M}_p(x) = \min_x M_p(x)$$

subject to: 

$$x \in \mathcal{H}$$

$$L_e \leq M_e(x_e) - d_e \leq U_e, \quad \forall e$$
Combining kinetic and instrumental models, B2B-DC predicts noisy/weak signals

<table>
<thead>
<tr>
<th>Prediction Feature</th>
<th>Prediction Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>O Peak Value</td>
<td>$[2.7, 4.3] \times 10^{-2}$</td>
</tr>
<tr>
<td>O Peak Location</td>
<td>$[1.9, 2.2]$ cm</td>
</tr>
<tr>
<td>OH Peak Value</td>
<td>$[3.0, 3.6] \times 10^{-2}$</td>
</tr>
<tr>
<td>OH Peak Location</td>
<td>$[1.60, 1.67]$ cm</td>
</tr>
<tr>
<td>C$_2$H$_3$ Peak Value</td>
<td>$[0.09, 1.15] \times 10^{-4}$</td>
</tr>
<tr>
<td>C$_2$H$_3$ Peak Location</td>
<td>$[0.6, 3.9] \times 10^{-2}$ cm</td>
</tr>
</tbody>
</table>

**Advance Light Source Flame Experiments**

35th Combust. Symp. 2014
Information Content of an Experiment

Including Experimental Observations

Posterior Range

Prior Range

Information Gain:

\[ I = 1 - \frac{\text{Posterior Range}}{\text{Prior Range}} \]
Given a budget $T$, determine the best strategy for reducing the uncertainty in model prediction.

\[ \Delta_{\text{opt}} = \arg \min \]  

subject to  

\[ C = \left( \frac{\Delta_{\text{current}}}{\Delta} \right)^a - 1 \]  

range predicted for $C_2H_6$ flame speed (cm/s)

$J. \text{Phys. Chem. A} \ 112:2579 \ (2008)$
Analysis of parameter distributions, predictions, and uncertainty correlations by sampling the feasible set.

Slavinskaia, et al., 2013, 2014
While a $M(x)$ formally depends on all $n$ active variables, in reality it mostly vary in $r \ll n$ linear combinations of the variables.

For creating a surrogate of $M(x)$ we would like to do the design in the $r$-dimensional space.

Factor $f(x) = g(S^T x)$

Compute gradients of $f(x)$ at points of $\mathcal{H}$

$$
\begin{bmatrix}
\nabla f(x^{(1)}) \\
\vdots \\
\nabla f(x^{(N)})
\end{bmatrix} =
\begin{bmatrix}
\nabla g(S^T x^{(1)}) \\
\vdots \\
\nabla g(S^T x^{(N)})
\end{bmatrix} \cdot S^T
$$

Perform SVD of $F$; this gives $S$

Sample $r$-subspace of $\mathcal{H}$ to build surrogate design
Even in case of rigorous Bayesian, use of a prescribed prior (e.g., Gaussian) underestimates the uncertainty in prediction (Phillip Stark, “Constraints versus Priors”, 2012) AND we unlikely to have Gaussian priors!

Approximations, even seemingly “harmless”, may lead to substantial differences in prediction of uncertainty (Russi et al, Chem. Phys. Lett. 2010)

Optimization-based methods, transferring uncertainty in two-steps — from data to parameters and then from parameters to prediction — necessarily overestimate the predicted uncertainty
B2B-DC and rigorous Bayesian produce consistent results

An ongoing collaborative study with
Jerome Sacks, *National Institute of Statistical Sciences*
Rui Paulo, *ISEG Technical University of Lisbon*
Gonzalo Garcia-Donato, *Universidad de Castilla-La Mancha, Spain*

Example: H2/O2
- 21 active variables
- 12 experimental targets
- predicting one “blind”

Example: GRI-Mech 3.0
- 102 active variables
- 76 experimental targets
- predicting one “blind”

B2B-DC prediction for this blind target is [1.89 2.12]
PERSONAL OBSERVATIONS

• current inability of truly predictive modeling
  – conflicting data in/among sources
  – poor documentation of data/models
  – no uncertainty reporting or analysis
  – not much focus on integration of data

• resistance to data sharing
  – no personal incentives
  – no easy-to-use technology

• no recognition of the problem
SUMMARY: B2B-DC

- is mathematically **rigorous**, numerically **efficient**, and **UQ-rich** approach to analysis of practical systems

- is **data-centric**, handles heterogeneous data, and is easily scalable to a large number of data sets

- is **scalable** to a large number of parameters through Solution Mapping features, combined with the Active Space Discovery

- establishes a clear **measure of consistency** among data and models, and identifies the cause of inconsistency if detected

- “measures” **information content** of an experiment
  - assess an **impact** of a given or planned experiment (“what if”)
  - design new experiments/theory that impact the most

- **reduces uncertainties** of known and **predicts correctly uncertainty** of unknown
• What causes/skews model predictiveness?

• Are there new experiments to be performed, old repeated, theoretical studies to be carried out?

• What impact could a planned experiment have?

• What is the information content of the data?

• What would it take to bring a given model to a desired level of accuracy?
NEED A PARADIGM SHIFT

from *algorithm-centric* view

```text
[diagram showing data → code → data]
```

to *data-centric* view
Valley Character of Objective

\[ \sum_{k=1}^{2} \text{deviations} \]
ADVANCE LIGHT SOURCE FLAME EXPERIMENTS


Integrated signal
Data Analysis

\[ y = f(x, c) \]

Concentration profile

Model

Integrated signal
Using Instrumental Model

Data Analysis

\[ y = f(x, c) \]

Calibration, \( c \)

Assumptions

Integrated signal
Using Instrumental Model

Data Analysis:
\[ y = f(x, c) \]

Calibration, \( c \)

Model

Instrumental Model

Integrated signal

Integrated Signal Profile for \( O_2 \)

Distance from Burner (mm)

Integrated Signal
O, OH, C$_2$H$_3$

- Peak Value
- Peak Location

<table>
<thead>
<tr>
<th>Prediction Feature</th>
<th>Prediction Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>O Peak Value</td>
<td>[2.73, 4.29] $\times 10^{-2}$</td>
</tr>
<tr>
<td>O Peak Location</td>
<td>[1.87, 2.20] cm</td>
</tr>
<tr>
<td>OH Peak Value</td>
<td>[2.97, 3.59] $\times 10^{-2}$</td>
</tr>
<tr>
<td>OH Peak Location</td>
<td>[1.60, 1.67] cm</td>
</tr>
<tr>
<td>C$_2$H$_3$ Peak Value</td>
<td>[0.09, 1.15] $\times 10^{-4}$</td>
</tr>
<tr>
<td>C$_2$H$_3$ Peak Location</td>
<td>[0.60, 3.90] $\times 10^{-2}$ cm</td>
</tr>
</tbody>
</table>
SCIENTIFIC METHOD

- experiments
- theory

hypothesis
collection
model
prediction
SENSITIVITY ANALYSIS:
What conditions will maximize sens to k?
UNCERTAINTY QUANTIFICATION:
What experiment will be most informative?
challenge: prediction

flame code

reaction model

input conditions

 thermo and transport data

validation data

prediction
curse of dimensionality

Volume of sphere
Volume of cube

dimension
$M_p(x)$

$M_p(x)$

$M_p(x)$
Data analysis performed in isolation leads to loss of information
Bayes theorem:

\[
\text{probability (hypothesis | data)} \propto \text{probability (data | hypothesis)} \times \text{probability (hypothesis)}
\]
prior likelihood posterior

\begin{align*}
    \text{prior} \quad \text{likelihood} \quad \text{posterior}
\end{align*}

\begin{align*}
    \text{probability} \quad \text{probability} \quad \text{probability}
    \begin{align*}
        \text{hypothesis|data} \quad \text{data|hypothesis} \quad \text{hypothesis}
    \end{align*}
\end{align*}

\begin{align*}
    \text{prediction} \quad \text{parameters}
    \begin{align*}
        \text{model / analysis}
    \end{align*}
\end{align*}
Quadratic surrogates enable mathematically rigorous, numerically efficient, and UQ-rich approach of B2B-DC to practical systems.

Quadratics work in practice because model parameters are limited
- by physical constraints; e.g., $0 < k < \text{collision limit}$
- by reaction theory / chemical analogy
- by prior experimental / theoretical studies
- and can be linearized; e.g., by the $\log$ transformation

And if they do not work, then
- rational quadratics (“native” with B2B framework)
- a two-level surrogate approach
  - first, use machine learning to build “high-order surrogates”, e.g., Gaussian Process, Kriging, $\varepsilon$-SVM, Polynomial Chaos
  - then, build/use on-demand piece-wise quadratics from the high-order surrogates
B2B-DC can account for emulator errors

\[ L \leq y \leq U \]

\[ y = M(x) + \varepsilon \]

\[ L - \varepsilon \leq M(x) \leq U + \varepsilon \]

\[ L' \leq M(x) \leq U' \]