# Sparse resolutions to inconsistent datasets using L1-minimization

Arun Hegde
Wenyu Li
Jim Oreluk
Andrew Packard
Michael Frenklach

This project is supported by the U.S. Department of Energy, National Nuclear Security Administration, under Award Number DE-NA0002375

# Overview

- Overview of Bound-to-Bound Data Collaboration (B2BDC)
  - models + data = *dataset (model-data system)*
- Dataset Consistency
  - scalar consistency measure
  - New vector consistency measure
- Dataset examples:
  - GRI-Mech 3.0
  - DLR-SynG
- B2BDC protocol for model validation
  - suggested use of B2BDC tools for model validation

# Bound-to-Bound Data Collaboration

UQ as constrained optimization: **parameters** constrained by **models** and **data**

## Models

"True" QOI models

$$f_e : \mathbb{R}^n \to \mathbb{R}$$

Surrogate QOI models

$$M_e(x) \approx f_e(x)$$

Fitting error

$$|M_e(x) - f_e(x)| \le \epsilon_e$$

$$e = 1, \ldots, N$$

## Data

- Prior knowledge on uncertain parameters

$$x \in \mathcal{H} \subseteq \mathbb{R}^n$$

- QOI measurements (w/ uncertainty)

$$L_e \le y_e \le U_e$$

$$e = 1, \ldots, N$$

## Dataset

$$x \in \mathcal{H} \subseteq \mathbb{R}^n$$

$$L_e - \epsilon_e \le M_e(x) \le U_e + \epsilon_e$$

$$\text{for } e = 1, \ldots, N$$

$$\mathcal{F} \subseteq \mathcal{H}$$

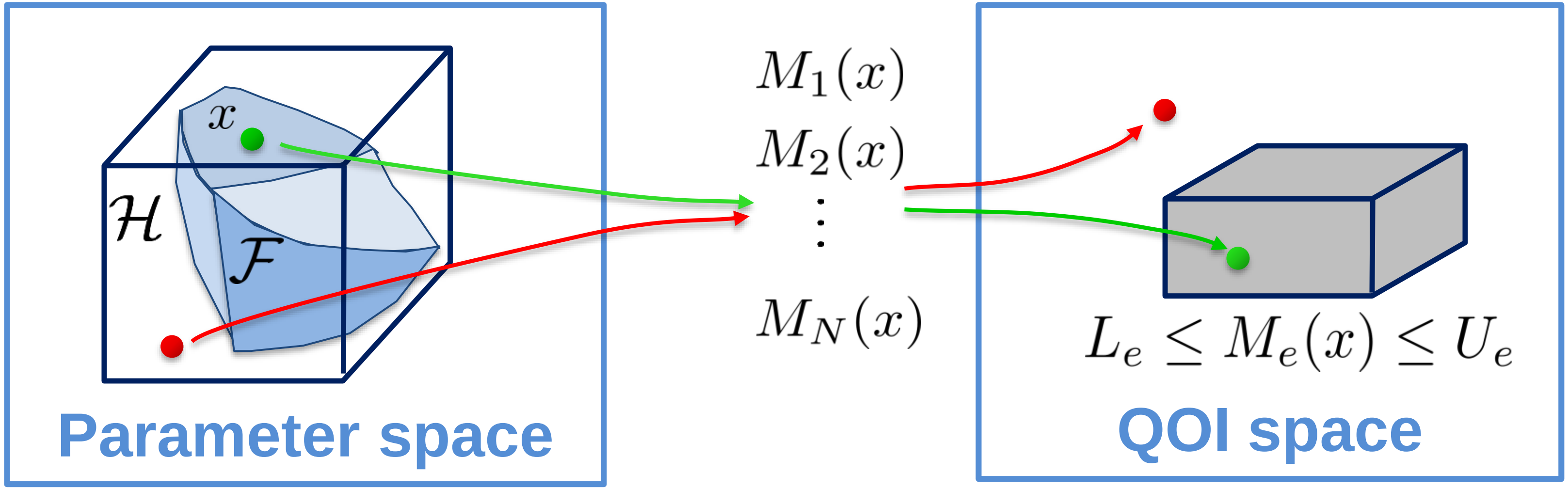
**Feasible set** – parameters for which the models and data agree.

Prediction establishes the range of a model $M_Q(x)$ subject to model-data constraints

$$\left( \begin{array}{c|c} \begin{aligned} \min \quad & M_Q(x) \\ \text{s.t.} \quad & x \in \mathcal{F} \end{aligned} & \begin{aligned} \max \quad & M_Q(x) \\ \text{s.t.} \quad & x \in \mathcal{F} \end{aligned} \end{array} \right)$$

# Consistency of a Dataset

- A dataset is **consistent** if it is feasible
  - Parameters exist for which model predictions match experimental observations




- Consistency analysis is quantifying **model validation**.

# Quantifying Consistency

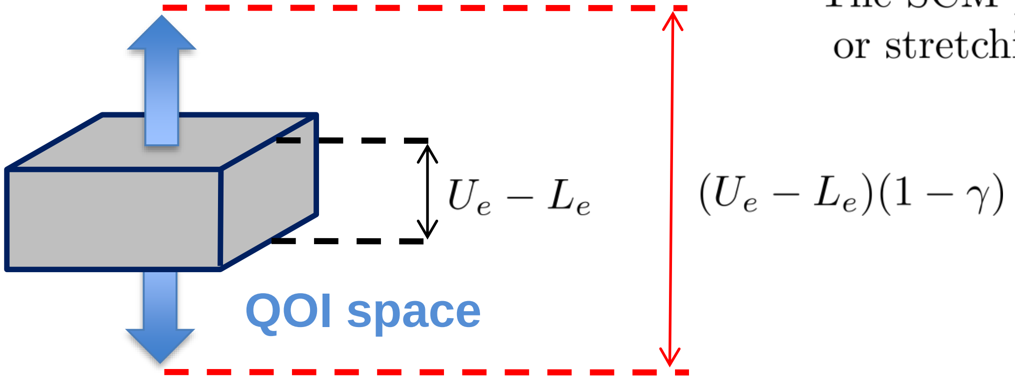
**Q:** Does there exist a parameter vector $x \in \mathcal{H}$ for which the models and data agree, within uncertainty?

**A:** Compute the *scalar consistency measure* (**SCM**)

## Scalar Consistency Measure (SCM)*

$$
\begin{aligned}
\max \quad & \gamma \\
\text{s.t.} \quad & L_e + \frac{(U_e - L_e)}{2}\gamma \leq M_e(x) \leq U_e - \frac{(U_e - L_e)}{2}\gamma \\
& x \in \mathcal{H}, \gamma \in \mathbb{R} \\
& \text{for } e = 1, \ldots, N
\end{aligned}
$$

The SCM produces a symmetric tightening ($\gamma > 0$) or stretching ($\gamma < 0$) of all experimental bounds.





* Feeley, R.; Seiler, P.; Packard, A.; Frenklach, M. *J. Phys. Chem. A.* **2004**, *108*, 9573-9583.

# Quantifying Consistency

**Q:** Does there exist a parameter vector $x \in \mathcal{H}$ for which the models and data agree, within uncertainty?

**A:** Compute the *scalar consistency measure* (**SCM**)

- If consistent, go to prediction. 
- If inconsistent, ??? 

## Scalar Consistency Measure (SCM)*

$$
\begin{aligned}
\max \quad & \gamma \\
\text{s.t.} \quad & L_e + \frac{(U_e - L_e)}{2}\gamma \le M_e(x) \le U_e - \frac{(U_e - L_e)}{2}\gamma \\
& x \in \mathcal{H}, \gamma \in \mathbb{R} \\
& \text{for } e = 1, \ldots, N
\end{aligned}
$$

* Feeley, R.; Seiler, P.; Packard, A.; Frenklach, M. *J. Phys. Chem. A.* **2004**, *108*, 9573-9583.

# Quantifying Consistency

**Q:** Does there exist a parameter vector $x \in \mathcal{H}$ for which the models and data agree, within uncertainty?

**A:** Compute the *scalar consistency measure* (**SCM**)

## Scalar Consistency Measure (SCM)*

$$
\begin{aligned}
\max \quad & \gamma \\
\text{s.t.} \quad & L_e + \frac{(U_e - L_e)}{2}\gamma \leq M_e(x) \leq U_e - \frac{(U_e - L_e)}{2}\gamma \\
& x \in \mathcal{H}, \gamma \in \mathbb{R} \\
& \text{for } e = 1, \ldots, N
\end{aligned}
$$

- If consistent, go to prediction. 
- If inconsistent, ??? 

**Next step:** identify which parts of this model-data system may be at fault.

* Feeley, R.; Seiler, P.; Packard, A.; Frenklach, M. *J. Phys. Chem. A.* **2004**, *108*, 9573-9583.

# Quantifying Consistency

**Q:** Does there exist a parameter vector $x \in \mathcal{H}$ for which the models and data agree, within uncertainty?

**A:** Compute the *scalar consistency measure* (**SCM**)

## Scalar Consistency Measure (SCM)*

$$
\begin{aligned}
&\max \quad \gamma \\
&\text{s.t.} \quad L_e + \frac{(U_e - L_e)}{2}\gamma \leq M_e(x) \leq U_e - \frac{(U_e - L_e)}{2}\gamma \\
&\qquad\quad x \in \mathcal{H}, \gamma \in \mathbb{R} \\
&\qquad\quad \text{for } e = 1, \ldots, N
\end{aligned}
$$

- If consistent, go to prediction. 
- If inconsistent, ??? 

**Next step:** identify which parts of this model-data system may be at fault.

Sensitivities*

- **Local:** $\lambda_j \approx \dfrac{\partial(\text{SCM})}{\partial(\text{bound } j)}$
- **Global:** $\Delta(\text{SCM}) \leq \lambda^T \Delta(\text{bounds})$

* Feeley, R.; Seiler, P.; Packard, A.; Frenklach, M. *J. Phys. Chem. A.* **2004**, *108*, 9573-9583.

# Quantifying Consistency

**Q:** Does there exist a parameter vector $x \in \mathcal{H}$ for which the models and data agree, within uncertainty?

**A:** Compute the *scalar consistency measure* (**SCM**)

## Scalar Consistency Measure (SCM)*

$$\begin{aligned} \max \quad & \gamma \\ \text{s.t.} \quad & L_e + \frac{(U_e - L_e)}{2}\gamma \le M_e(x) \le U_e - \frac{(U_e - L_e)}{2}\gamma \\ & x \in \mathcal{H}, \gamma \in \mathbb{R} \\ & \text{for } e = 1, \ldots, N \end{aligned}$$

- If consistent, go to prediction. 
- If inconsistent, ??? 

**Next step:** identify which parts of this model-data system may be at fault.

**New criteria can be used for the identification:**

- How many experimental bounds do we need to change to become consistent?
  - search for a **sparse** resolution to the inconsistency
  - sparse solutions are **interpretable**

* Feeley, R.; Seiler, P.; Packard, A.; Frenklach, M. *J. Phys. Chem. A.* **2004**, *108*, 9573-9583.

# Vector Consistency

**Q:** Does there exist a parameter vector $x \in \mathcal{H}$ for which the models and data agree, within uncertainty?

**A:** Compute the *scalar consistency measure* (**SCM**)

## Scalar Consistency Measure (SCM)

$$\begin{aligned} \max \quad & \gamma \\ \text{s.t.} \quad & L_e + \frac{(U_e - L_e)}{2}\gamma \leq M_e(x) \leq U_e - \frac{(U_e - L_e)}{2}\gamma \\ & x \in \mathcal{H}, \gamma \in \mathbb{R} \\ & \text{for } e = 1, \ldots, N \end{aligned}$$

If inconsistent, compute the *vector consistency measure* (**VCM**)

- alternative consistency measure
- offers detailed analysis of inconsistency by allowing independent relaxations

## Vector Consistency Measure (VCM)

$$\begin{aligned} \min \quad & \|\Delta^L\|_1 + \|\Delta^U\|_1 \\ \text{s.t.} \quad & L_e - \Delta_e^L \leq M_e(x) \leq U_e + \Delta_e^U \\ & \Delta_e^L, \Delta_e^U \in \mathbb{R}, x \in \mathcal{H} \\ & \text{for } e = 1, \ldots, N \end{aligned}$$

# Vector Consistency

**Q:** Does there exist a parameter vector $x \in \mathcal{H}$ for which the models and data agree, within uncertainty?

**A:** Compute the *scalar consistency measure* (**SCM**)

## Scalar Consistency Measure (SCM)

$$\begin{aligned} \max \quad & \gamma \\ \text{s.t.} \quad & L_e + \frac{(U_e - L_e)}{2}\gamma \le M_e(x) \le U_e - \frac{(U_e - L_e)}{2}\gamma \\ & x \in \mathcal{H}, \gamma \in \mathbb{R} \\ & \text{for } e = 1, \ldots, N \end{aligned}$$

If inconsistent, compute the *vector consistency measure* (**VCM**)

- alternative consistency measure
- offers detailed analysis of inconsistency by allowing independent relaxations

## Vector Consistency Measure (VCM)

$$\begin{aligned} \min \quad & \boxed{\|\Delta^L\|_1 + \|\Delta^U\|_1} \\ \text{s.t.} \quad & L_e - \Delta_e^L \le M_e(x) \le U_e + \Delta_e^U \\ & \Delta_e^L, \Delta_e^U \in \mathbb{R}, x \in \mathcal{H} \\ & \text{for } e = 1, \ldots, N \end{aligned}$$

heuristic for sparsity

# Examples*

* Hegde, A.; Li, W.; Oreluk, J.; Packard, A.;  Frenklach, M. **2017**. *arXiv:1701.04695*.

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

### Scalar Consistency

- **Procedure:** Iteratively apply SCM, using sensitivities (Lagrange multipliers) to identify problems.

### Vector Consistency

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

### Scalar Consistency

- **Procedure:** Iteratively apply SCM, using sensitivities (Lagrange multipliers) to identify problems.
- SCM < 0. Analyze ranked sensitivities

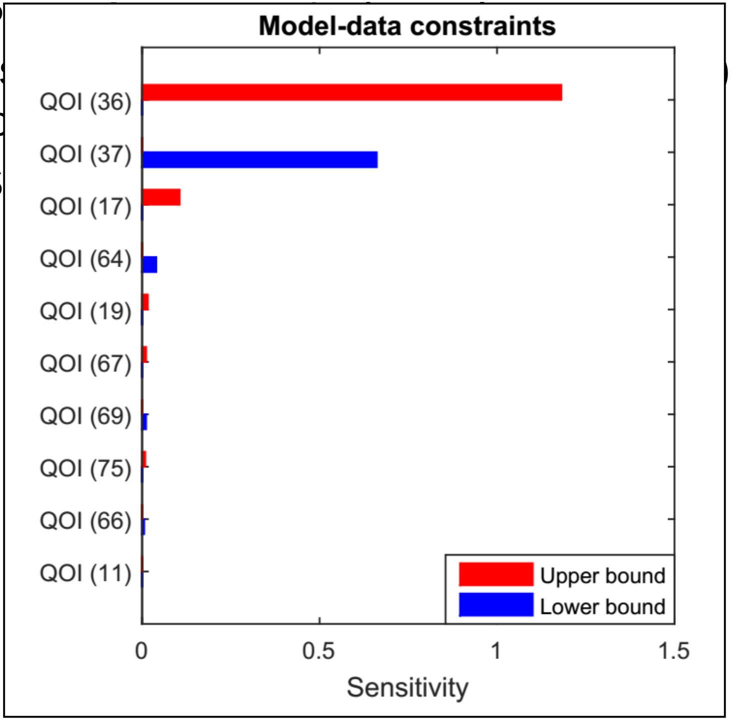
### Vector Consistency

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

## Scalar Consistency

- **P** us to
- S



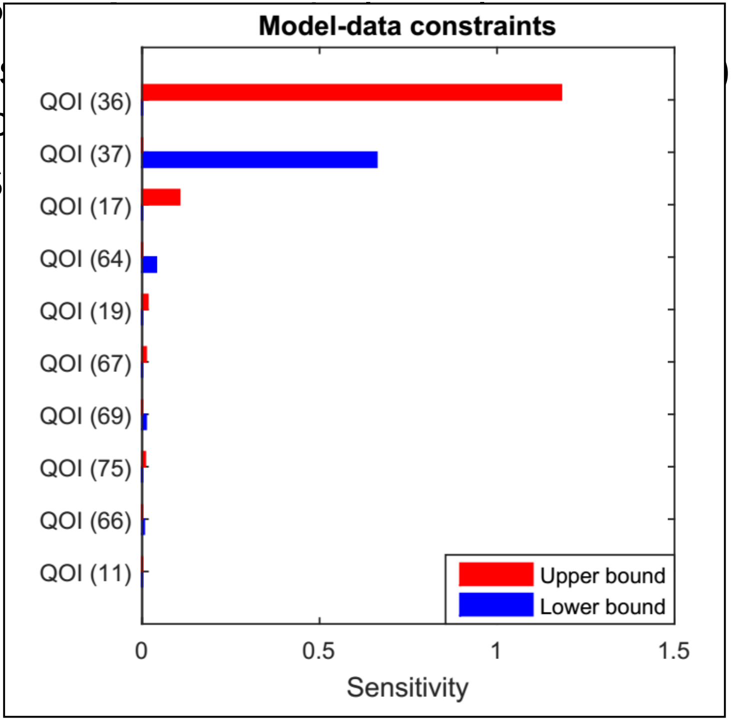

## Vector Consistency

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

## Scalar Consistency

- P
  us
  to
- S




## Vector Consistency

Remove the top most sensitive QOI
Remove the top two most sensitive QOIs
Remove the top *n* most sensitive QOIs
Remove the second most sensitive QOI
(counter intuitive)

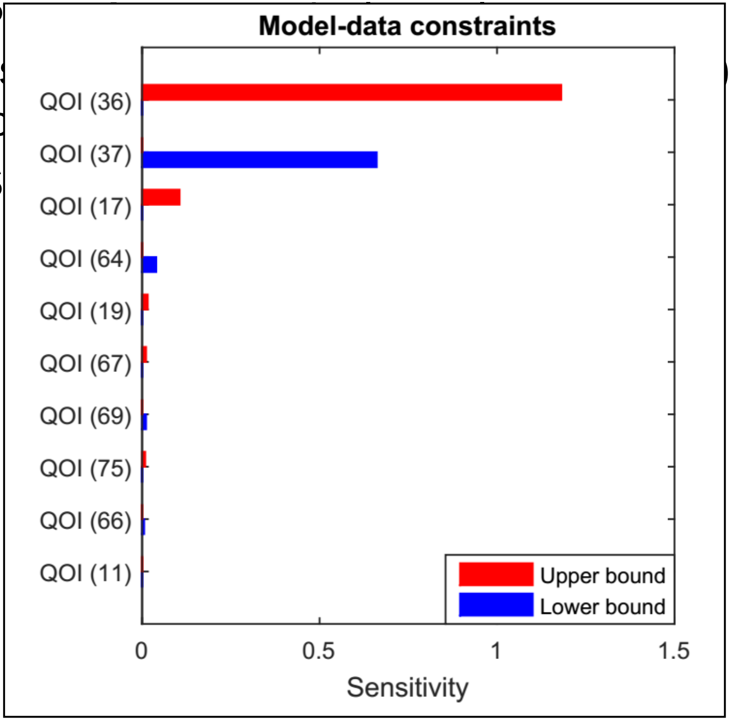
.
.
.

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

## Scalar Consistency

- **P**
  us
  to
- S




## Vector Consistency

~~Remove the top most sensitive QOI~~

**Remove the top two most sensitive QOIs**

~~Remove the top *n* most sensitive QOIs~~

~~Remove the second most sensitive QOI (counter intuitive)~~

.

.

.

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

## Scalar Consistency

- **Procedure:** Iteratively apply SCM, using sensitivities (Lagrange multipliers) to identify problems.
- SCM < 0. Analyze ranked sensitivities
- SCM > 0. Two QOIs removed, dataset consistent.

## Vector Consistency

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

## Scalar Consistency

- **Procedure:** Iteratively apply SCM, using sensitivities (Lagrange multipliers) to identify problems.
- SCM < 0. Analyze ranked sensitivities
- SCM > 0. Two QOIs removed, dataset consistent.

## Vector Consistency

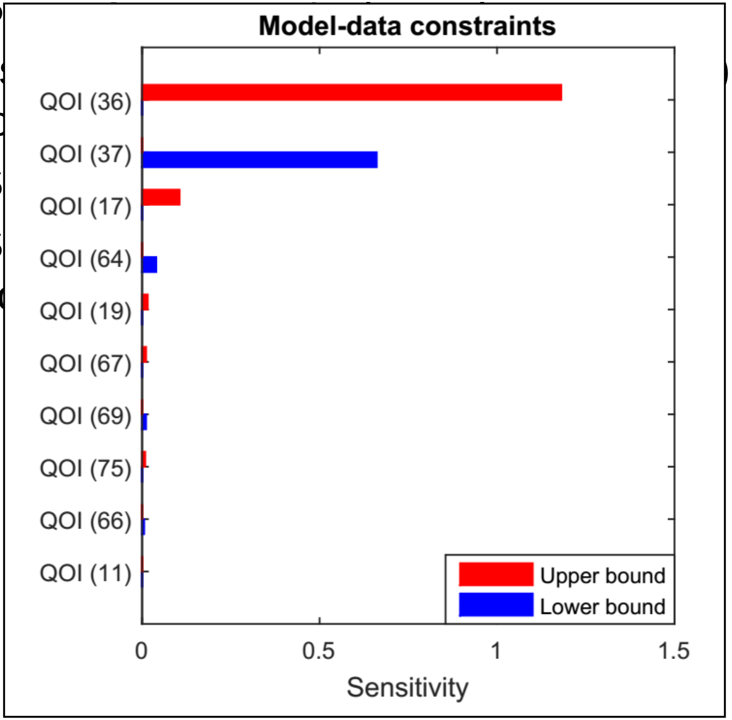
- Compute VCM.

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

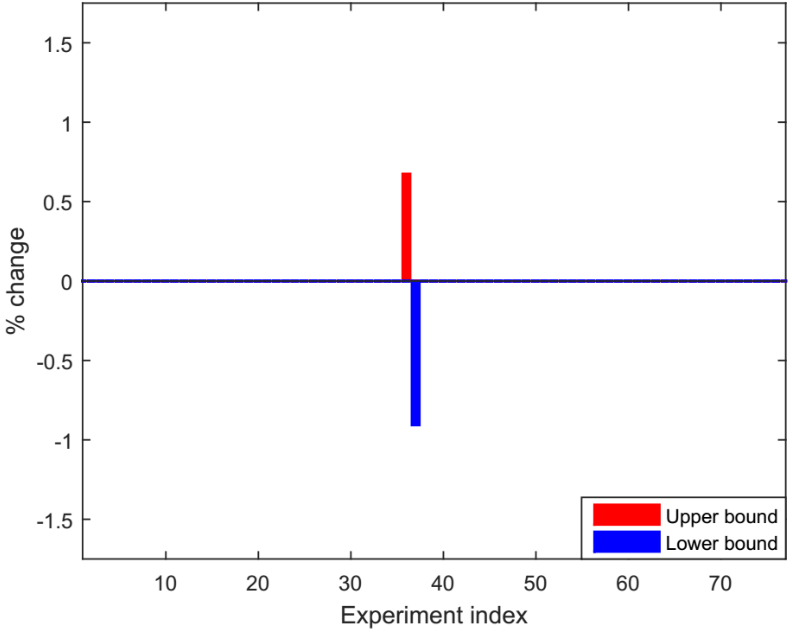
## Scalar Consistency

- P
  us
  to
- S
- S
  co




## Vector Consistency

- Compute VCM.



# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

### Scalar Consistency

- **Procedure:** Iteratively apply SCM, using sensitivities (Lagrange multipliers) to identify problems.
- SCM < 0. Analyze ranked sensitivities
- SCM > 0. Two QOIs removed, dataset consistent.

### Vector Consistency

- Compute VCM.
- Two QOIs relaxed (same as in SCM), dataset consistent.

# Comparison of Methods: GRI-Mech 3.0

GRI-Mech 3.0 dataset (77 QOIs, 102 uncertain parameters) for natural gas combustion.

### Scalar Consistency

- **Procedure:** Iteratively apply SCM, using sensitivities (Lagrange multipliers)

Rapid and interpretable resolution of inconsistency

### Vector Consistency

- Compute VCM.
- Two QOIs relaxed (same as in SCM),

Rapid and interpretable resolution of inconsistency

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

**Scalar Consistency**

**Vector Consistency**

* Slavinskaya, N.; et al. *Energy & Fuels.* **2017**, vol. 31, pp 2274–2297

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

**Scalar Consistency**

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).

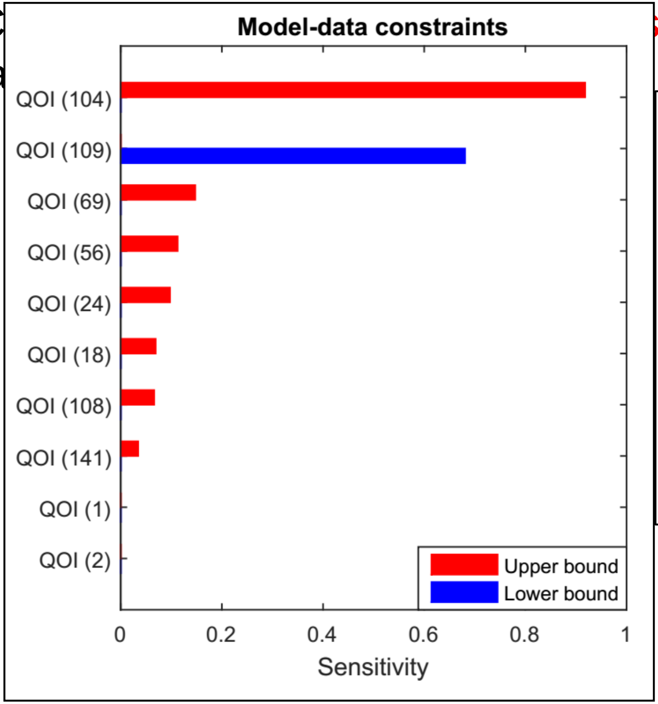
**Vector Consistency**

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

## Scalar Consistency

- SC
(La




## Vector Consistency

Remove the top most sensitive QOI
Remove the top two most sensitive QOIs
Remove the top *n* most sensitive QOIs
Remove the second most sensitive QOI
(counter intuitive)

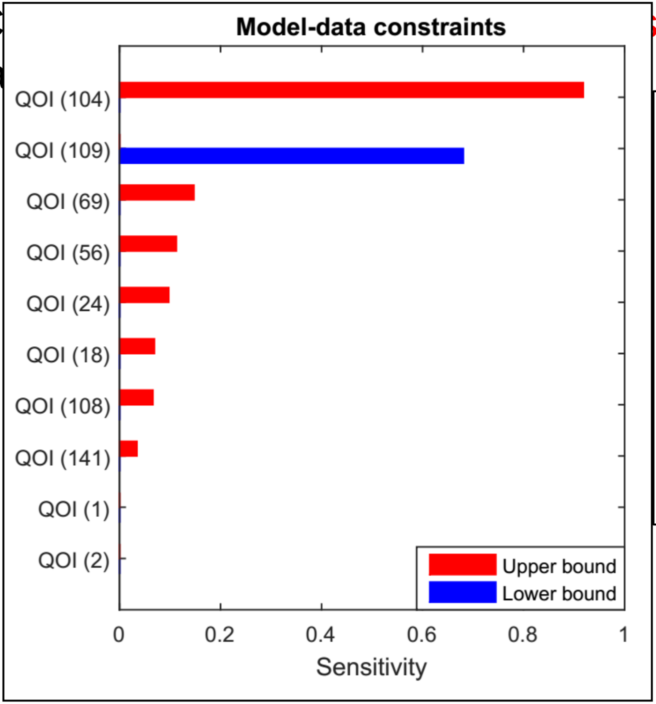
.
.
.

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

## Scalar Consistency

- SC
  (La




## Vector Consistency

**Remove the top most sensitive QOI**

~~Remove the top two most sensitive QOIs~~

~~Remove the top *n* most sensitive QOIs~~

~~Remove the second most sensitive QOI~~

~~(counter intuitive)~~

.

.

.

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

**Scalar Consistency**

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.

**Vector Consistency**

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

### Scalar Consistency

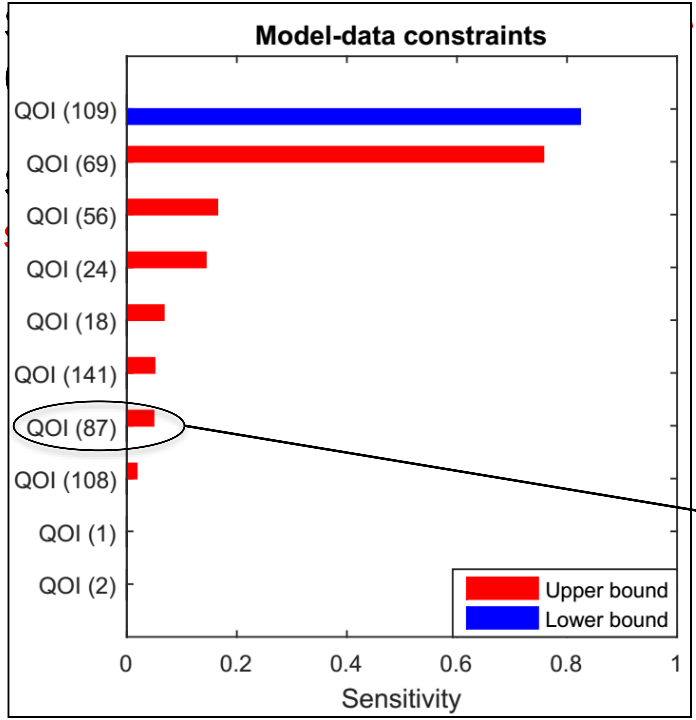
- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.

### Vector Consistency

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

## Scalar Consistency



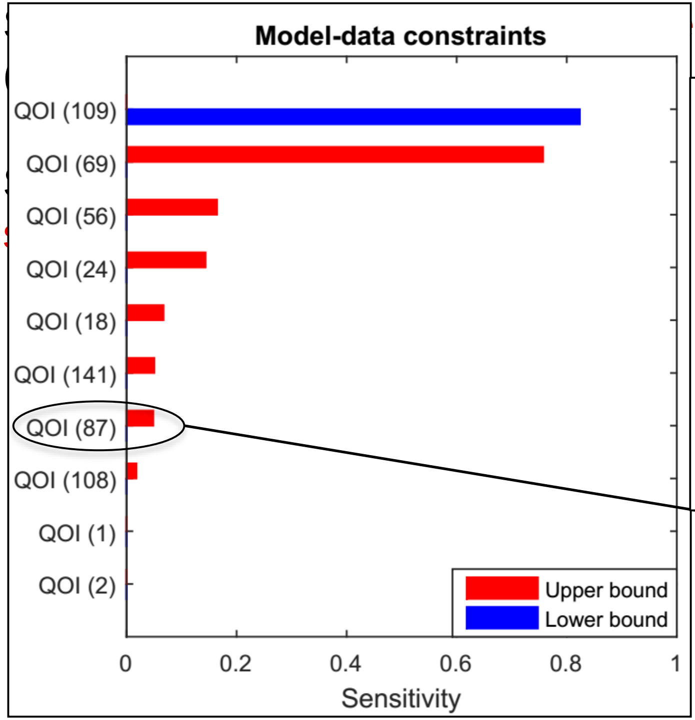

## Vector Consistency

Previously, not identified

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

## Scalar Consistency




## Vector Consistency

**Remove the top most sensitive QOI**

~~Remove the top two most sensitive QOIs~~

~~Remove the top *n* most sensitive QOIs~~

~~Remove the second most sensitive QOI (counter intuitive)~~

.

.

.

Previously, not identified

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

**Scalar Consistency**

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.
  - Remove QOI # 109.

**Vector Consistency**

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

**Scalar Consistency**

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.
  - Remove QOI # 109.

Repeat until consistent

**Vector Consistency**

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

### Scalar Consistency

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.
  - Remove QOI # 109.

Repeat until consistent

- This strategy results in the removal of 73 QOIs.

### Vector Consistency

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

**Scalar Consistency**

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.
  - Remove QOI # 109.

Repeat until consistent

- This strategy results in the removal of 73 QOIs.
- Another strategy results in 56 QOIs removed.

**Vector Consistency**

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

**Scalar Consistency**

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.
  - Remove QOI # 109.

Repeat until consistent

- This strategy results in the removal of 73 QOIs.
- Another strategy results in 56 QOIs removed.

**Vector Consistency**

- Compute VCM.

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

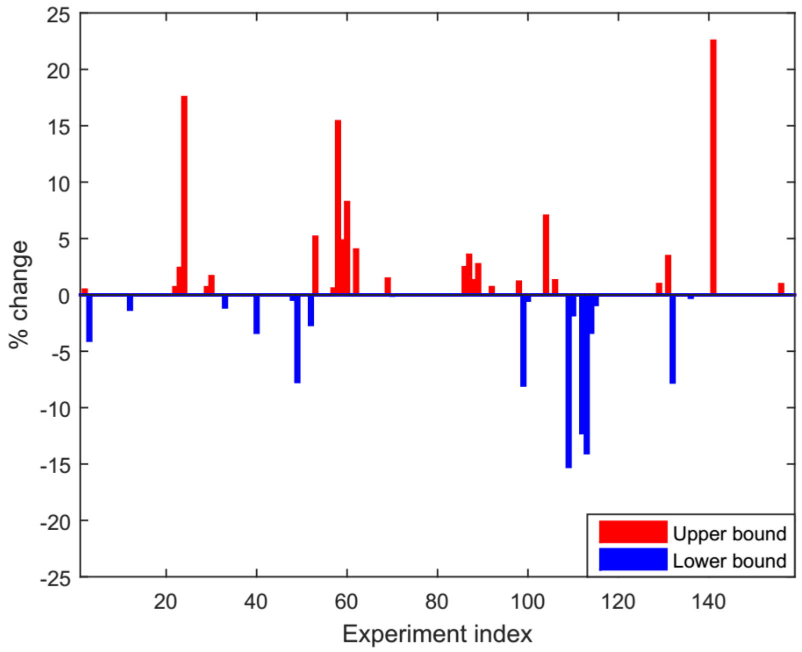
**Scalar Consistency**

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.
  - Remove QOI # 109.

Repeat until consistent

- This strategy results in the removal of 73 QOIs.
- Another strategy results in 56 QOIs removed.

Implementing the 43 relaxations results in a consistent dataset



# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

### Scalar Consistency

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.
  - Remove QOI # 109.

Repeat until consistent

- This strategy results in the removal of 73 QOIs.
- Another strategy results in 56 QOIs removed.

### Vector Consistency

- Compute VCM.
  - Recommends 43 relaxations (18 to lower bounds, 25 to upper bounds)

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

### Scalar Consistency

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset.
- SCM < 0. Analyze ranked sensitivities.
  - Remove QOI # 109.

Repeat until consistent

- This strategy results in the removal of 73 QOIs.
- Another strategy results in 56 QOIs removed.

### Vector Consistency

- Compute VCM.
  - Recommends 43 relaxations (18 to lower bounds, 25 to upper bounds)

Example of what we termed **massive inconsistency**

# Advantages of VCM: DLR-SynG

DLR-SynG dataset* (159 QOIs, 55 uncertain parameters) developed at DLR.

### Scalar Consistency

- SCM < 0. Analyze ranked sensitivities (Lagrange multipliers).
  - Remove QOI #104 from dataset

Indirect and inefficient resolution of inconsistency

73 QOIs.

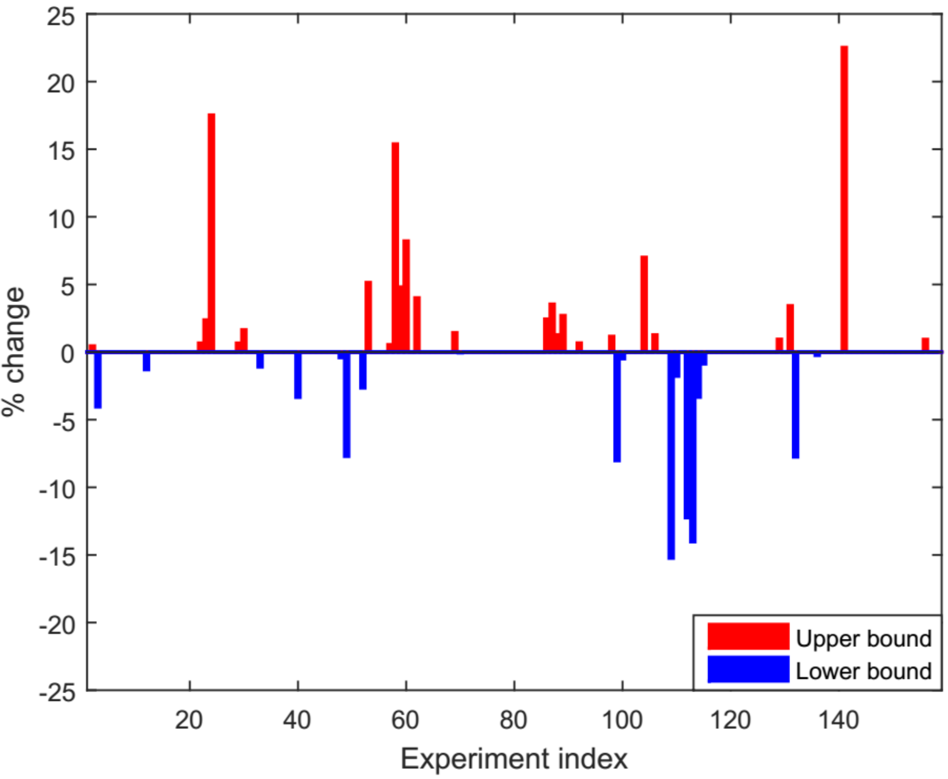
- Another strategy results in 56 QOIs removed.

### Vector Consistency

- Compute VCM.
  - Recommends 43 relaxations (18 to lower bounds, 25 to upper bounds)

Direct, one-shot resolution of inconsistency

# Including weights




What if we are unwilling to change certain experimental bounds?

# Including weights

- Domain expert knowledge and opinions enter VCM as weights.

- **Idea:** If a <u>dataset</u> is inconsistent, one should be less willing to relax model-data constraints they trust and more willing to relax constraints that are less reliable. The same goes for parameter bounds.

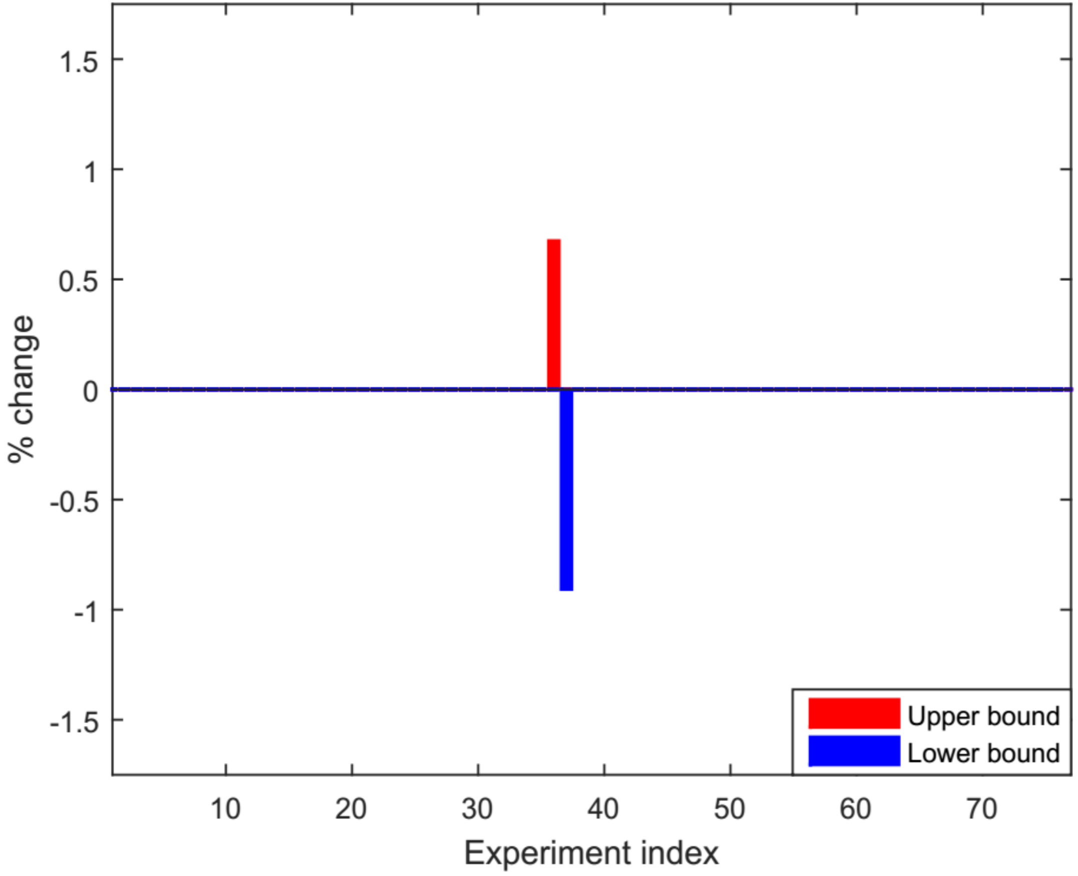
**Weighted VCM**

$$\begin{aligned}\min_{x,\Delta^L,\Delta^U,\delta^l,\delta^u} \quad & \|\Delta^L\|_1 + \|\Delta^U\|_1 + \|\delta^l\|_1 + \|\delta^u\|_1 \\ \text{s.t.} \quad & L_e - W_e^L \Delta_e^L \leq M_e(x) \leq U_e + W_e^U \Delta_e^U \qquad \text{for } e = 1, \ldots, N \\ & l_i - w_i^l \delta_i^l \leq x_i \leq u_i + w_i^u \delta_i^u \qquad \text{for } i = 1, \ldots, n\end{aligned}$$

- Small weight - less willing to change bound.
- Large weight - more willing to change bound.

# Weights and GRI-Mech 3.0

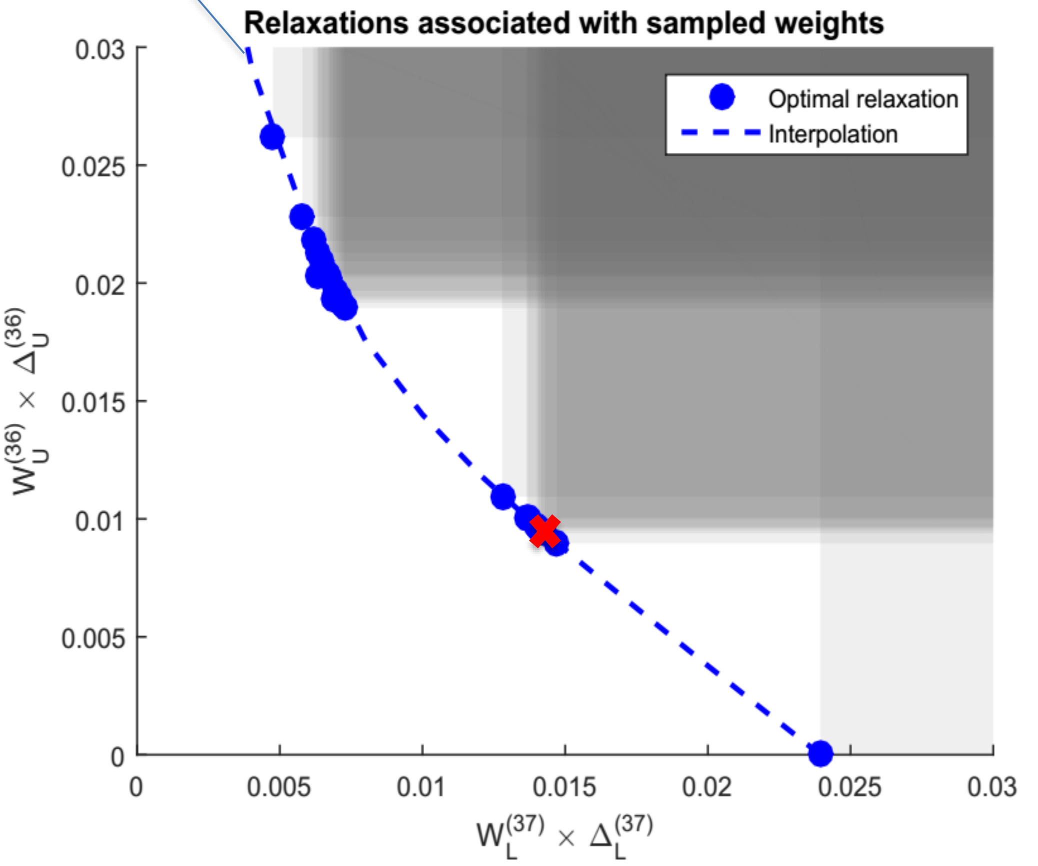
- Single application of VCM identifies two experimental bounds.



# Weights and GRI-Mech 3.0

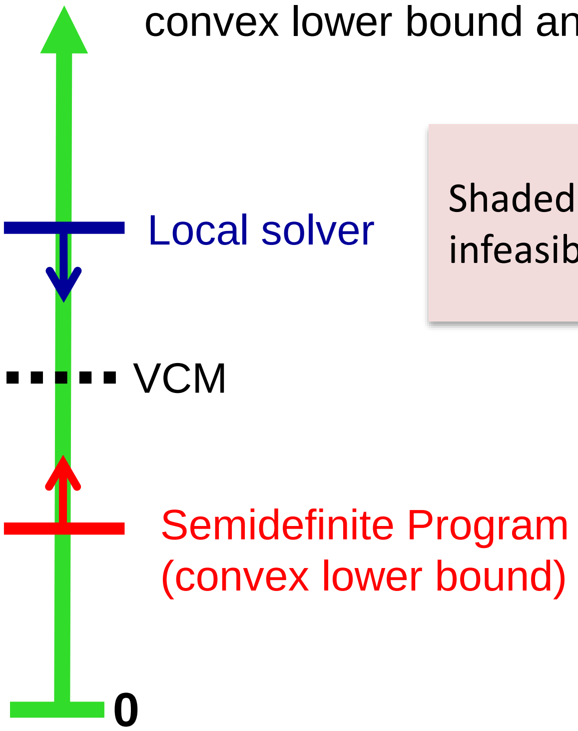
- Single application of VCM identifies two experimental bounds.
- Weights applied to only the previous two bounds.

Tradeoff effect from using different weights



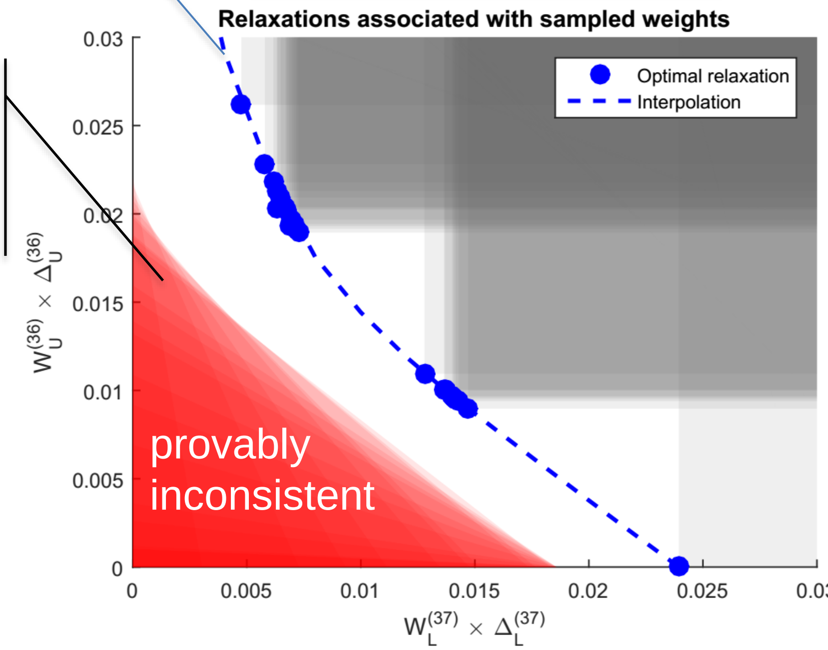
# Computing the VCM

For **quadratic surrogate models**, we can approximate the true solution (NP-hard) using convex lower bound and local optima.




Shaded red region certified infeasible by SDP.

Tradeoff effect from using different weights




Example: GRI-Mech 3.0 dataset – VCM relaxations with varying weights, relaxations allowed to two constraints

# Protocol for model validation

## <u>B2BDC protocol</u>

**Step 1:** Construct *dataset* - QOI selection, model building, data collection, etc.

**Step 2:** Remove self-inconsistent QOIs

**Step 3:** **<u>Scalar consistency (SCM)</u>** analysis

IF inconsistent

perform **<u>vector consistency (VCM)</u>** analysis

**Step 4:** Prediction & further analysis

# Summary

- **Consistency** analysis is **model validation**

- **Vector Consistency** offers an efficient approach to resolving inconsistent datasets
  - particularly efficient for resolving massive inconsistency
  - incorporates expert knowledge through weights
  - examples: GRI-Mech 3.0, DLR-SynG

- Utilizing both the **SCM** and the **VCM** offers a powerful strategy for model validation

# Acknowledgements

This work is supported as a part of the CCMSC at the University of Utah, funded through PSAAP by the National Nuclear Security Administration, under Award Number DE-NA0002375.

National Nuclear Security Administration

CARBON CAPTURE MULTIDISCIPLINARY SIMULATION CENTER

# Questions?