Consistency of a Reaction Dataset

Ryan Feeley, Pete Seiler, Andrew Packard,* and Michael Frenklach*

Department of Mechanical Engineering, University of California, Berkeley, California 94720-1740

Received: June 8, 2004; In Final Form: August 12, 2004

The numerical approach of data collaboration is extended to address the mutual consistency of experimental observations. The analysis rests on the concept of a dataset, which represents an organization of pertinent experimental observations, their uncertainties, and mechanistic knowledge of the subject of interest. The numerical foundation of data collaboration lies in constrained optimization, utilizing solution mapping tools and robust control algorithms. A rigorous measure of dataset consistency is developed, and Lagrange multipliers are used to identify factors that influence consistency. The new analysis is demonstrated on a real-world example, taken from the field of combustion. In performing the consistency test, the new procedure identifies two major outliers of the dataset, which were corrected upon re-examination of the raw experimental data. The results of the analysis suggest a sequential procedure with step-by-step identification of outliers and inspection of the causes. Altogether, the new numerical approach offers an important tool for assessing experimental observations and model building.

1. Introduction

The understanding of a variety of natural phenomena and industrial processes is reliant on knowledge of the chemical reaction mechanisms and kinetics. Endeavors in such cases begin with identification of the underlying reaction pathways and fundamental mechanisms. When sufficient data accumulate, the interest often shifts to practical applications, motivating the development of mechanistic models.

The “textbook” approach to the development of mechanistic reaction models consists of conjecturing the reaction mechanism, expressing it in a suitable mathematical form, and comparing the predictions of the constructed model to available experimental observations. Typically, such comparisons result in mixed outcomes: some show a reasonably close agreement and some do not. In the latter case, the apparent inconsistency obtained between the model and the experiment is argued to imply either that the model is inadequate or that the experiment (or, rather, its interpretation) is incorrect.

In some areas, such as heterogeneous catalysis and biochemical systems, the fundamental reaction mechanisms are largely unknown and establishing them form the challenge of the current research. Yet, in other fields, such as atmospheric chemistry and the combustion of small hydrocarbons such as methane, there is a broad consensus in regard to the reaction pathways underlying the mechanisms. Thus, any inadequacy of the kinetic models essentially rests in their parameter values. In the following discussion, we assume the latter situation.

If the kinetic parameters of such a “known” mechanism were known exactly, then a direct comparison of model prediction with a given experiment, within its uncertainties, would decisively indicate whether that experiment is consistent or inconsistent with the model. In reality, however, the model parameters themselves have uncertainties that must be included in the analysis.

In principle, the parameter identification of chemical kinetic models can be posed as a classical statistical inference:[1−3] given a mathematical model and a set of experimental observations for the model responses, determine the best-fit parameter values, usually those that produce the smallest deviations of the model predictions from the measurements. The validity of the model and the identification of outliers is then determined using analysis of variance. The difficulty involved in the application of standard statistical methods lies in the fact that chemical kinetics models are stated in the form of differential equations that do not possess a closed-form solution. Further complications result from the highly “ill-structured” character of the best-fit objective function, with long and narrow valleys and multiple local minima, resulting in an ill-conditioned optimization that lacks a unique solution.[2,4]

The best-fit optimization problem for general, nonlinear dynamic models has been addressed with a series of numerical methods: “direct” gradient search,[5,6] gradient search based on sensitivities,[7] solution mapping,[4,8−10] genetic algorithms,[11,12] and Monte Carlo techniques.[7,13] In some cases, it was coupled with statistical inference and estimation of confidence regions.[7,8,13] Recent developments also include formulation of the problem in the form of error propagation: given uncertainty ranges for model parameters, estimate the intervals of variations for model predictions.[13,14]

All of the aforementioned methods essentially view the problem as a two-step process: estimation of model parameters from fitting a selected set of experimental data, followed by exercise of the obtained model, either as validation against an additional set of experiments or making predictions outside the experimentally accessible conditions.

Recently, we have pursued a different approach, which we call *data collaboration*.[15,16] In this approach, we focus not on parametrization of the parameter uncertainty region, which the aforementioned methods engage in and rely upon, but rather on transferring the uncertainties of the “raw” (experimental) data into the model *directly*. Doing so allows one to harvest substantially more of the information content of the data[16] and determine more-realistic bounds on model predictions.[15] Our approach is anchored in the concept of a *dataset* that unites all

* Authors to whom correspondence should be addressed. Telephone: 510-643-1676. Fax: 510-642-6163. E-mail: pack@me.berkeley.edu, myf@me.berkeley.edu.

10.1021/jp047524w CCC: $27.50 © 2004 American Chemical Society
Published on Web 10/12/2004

the pertinent experimental data and the mechanistic knowledge for a given system, and numerical analysis based on combination of solution mapping and optimization techniques used and advanced by robust control theory.[17] This numerical methodology avoids the unnecessary overconstraining of model parameters that plagues many other techniques due to inherent correlations among parameters, and allows one to explore more closely the true *feasible region* of the parameter space in a computationally efficient manner.

The present work expands further on these ideas. Within the framework of a dataset, we develop a numerical measure of dataset consistency, which provides a combined way to examine system uncertainties that originate from either rate parameters and/or experimental observations. The analysis of dataset consistency is assisted by Lagrange multipliers, which gauge the sensitivity of the consistency measure to the dataset uncertainties. We begin with a brief description of the concepts and methodology of data collaboration, followed by mathematical formulation of the problem; we then present new mathematical developments on dataset consistency and conclude with a realistic demonstration of the method.

2. Data Collaboration

2.1. Dataset. Let E denote a physical experiment (for example, a flow reactor or a laminar premixed flame) and Y_e a property of interest that is measured in this experiment (e.g., the intensity of scattered light or a peak concentration). The value of Y_e is designated y_e, and the experimentally measured value is d_e. We assume the experimental uncertainty is not necessarily symmetric about d_e, and, thus, the deviation of the measured value from y_e has lower and upper bounds, namely $l_e \leq y_e - d_e \leq u_e$.

We associate with experiment E a *dataset unit*, (d_e, u_e, l_e, M_e), which consists of the measured value, the reported uncertainty in the measurement (upper bound and lower bound), and a mathematical model, respectively. A *dataset* is a collection of dataset units $\{(d_e, u_e, l_e, M_e)\}$. In the following, we denote a dataset unit as $\mathcal{U}_e$, the dataset as $\mathcal{D}$, and the set of indices e as $\mathcal{E}$; i.e., $\mathcal{U}_e = (d_e, u_e, l_e, M_e)$ and $\mathcal{D} = \{\mathcal{U}_e\}_{e\in\mathcal{E}}$. The model M_e is defined as the functional relation between the model parameters and the prediction for Y_e. Discussion of the model definition follows.

The creation and organization of a dataset is guided by the system in question, for instance, the formation of nitrogen oxides in the combustion of natural gas, the concentration levels of ozone in the atmosphere, or transmembrane signaling in bacterial chemotaxis. A single experiment cannot provide complete information on such a system, but rather probes its particular aspect. A collection of such individual "bits" of pertinent information (i.e., dataset units) forms a dataset. The more extensive and diverse the collection, the more complete is the understanding of the system.[16] The unifying principle, the one that determines the "pertinence" of a given experiment to a given dataset, is a presumption that there exists a single chemical kinetics model, common to all dataset units, that is expected to predict Y_e when exercised at the conditions of experiment E. In other words, it is presumed that a broad consensus exists (at least tentatively) regarding the necessary reaction steps of the system and, hence, the mathematical structure of the unifying kinetic model is known, and that this mathematical model is sufficient, in principle (with the "right" choice of parameter values), to predict all experimental observations included in the dataset.

For a known chemical reaction system, the mathematical form of the kinetic model is a set of ordinary differential equations (ODEs) that describe the time evolution of all chemical species. The ODE formulation is based on reaction-rate laws (such as Mass Action[18] or Michaelis−Menten[19]) that contain either physical or empirical parameters. Their values could be entirely unknown, but usually they are known within some bounding intervals that have been established in prior studies or estimated theoretically. Experience shows[9,20] that, for an individual experiment E, only a small subset of model parameters has a measurable influence on the property Y_e. We denote such a subset as $\mathcal{X}_e$, and we refer to the parameters contained within $\mathcal{X}_e$ as *active variables* for experiment E. For instance, the ignition time of a methane−air mixture is primarily determined by a dozen or so kinetic parameters, and the influence of the rest of the parameters (above 600 in the case of the GRI-Mech model[20]) is largely within the noise and, for all practical purposes, can be safely neglected[9,20] (e.g., by fixing them at their respective "literature" values). This phenomenon is termed *effect sparsity*.[21,22]

We will designate an individual active variable as X_j and use $x_j \in \mathrm{R}$ to refer to a specific value of X_j. Individual dataset units may (and usually do) have different sets of active variables. Thus, X_j might be an active variable for one experiment but not another. We denote the list of active variables for experiment E as $\mathcal{X}_e$. The union over all $e \in \mathcal{E}$ form the dataset active variables, $\mathcal{X} = \cup_{e\in\mathcal{E}}\mathcal{X}_e$. We will denote the total number of dataset active variables as n, and vector $\boldsymbol{x} \in \mathrm{R}^n$ represents dataset active variable values. Associated with a vector $\boldsymbol{x}$, $\boldsymbol{x}_e$ are the values extracted from $\boldsymbol{x}$ that correspond to the active variable set $\mathcal{X}_e$.

In the context of tuning model parameters (e.g., rate constants) through optimization, the initial conditions of ODE integration (pressure, temperature, etc.) for a dataset unit $\mathcal{U}_e$ are fixed to those of experiment E. The only changes occurring from run to run are those in the values of optimization variables (such as pre-exponential factors of rate coefficients, activation energies, ratios of rate coefficients, and enthalpies of formation). Thus, the model M_e of the dataset unit $\mathcal{U}_e$ represents the relationship between values of the active variables of experiment E and model predictions for Y_e. In other words, $M_e(\boldsymbol{x}_e)$ replaces y_e, yielding $l_e \leq M_e(\boldsymbol{x}_e) - d_e \leq u_e$, which ties together the data, the model, and the uncertainty.

2.2. Initial Hypercube and Feasible Region. We further assume that prior information on the possible values of the dataset active variables is available. For instance, the value of activation energy computed quantum-mechanically will have an uncertainty that is associated with that calculation, or there could be several experimental studies, each reporting a different value for the same rate constant. This prior information can be expressed as the confinement of possible values of the active variables to an n-dimensional "hypercube", $\mathcal{H} = \{\boldsymbol{x} \in \mathrm{R}^n\colon \alpha_j \leq x_j \leq \beta_j\}$, where α_j and β_j are the lower and upper bounds on x_j for $j = 1, 2, ..., n$. Each edge of the hypercube $\mathcal{H}$ represents the presumed interval of *physically allowed* values of the corresponding active variable, either the estimated uncertainty or a range that contains the differing values.

Some parameter values drawn from $\mathcal{H}$ may result in model predictions that lie outside the experimentally determined ranges. In other words, not every $\boldsymbol{x} \in \mathcal{H}$ predicts all experimental observations of the dataset within their specified uncertainties. The collection of parameter values that are both contained in the hypercube and satisfy $l_e \leq M_e(\boldsymbol{x}_e) - d_e \leq u_e$ for every $e \in \mathcal{E}$ form the *feasible region*, $\mathcal{F}_{\mathcal{D}}$. A point $\boldsymbol{x}$ that is not contained in $\mathcal{F}_{\mathcal{D}}$ has been eliminated from consideration as a possible value for the dataset active variables by either the prior information, through the α and β bounds of $\mathscr{H}$, or by the experimental observations of the dataset, through intervals $(d_e + l_e, d_e + u_e)$. It is in this manner that experimental observations increase our knowledge of the kinetic parameters: an experiment may eliminate portions of the hypercube $\mathscr{H}$ from consideration, thereby decreasing the uncertainty in the values of the kinetic parameters. Further discussion and illustration of the feasible region and its character can be found in ref 16.

2.3. Methodology. Our approach casts a given problem as a constrained optimization over the feasible region, drawn on the entire knowledge content of a dataset. It combines solution mapping (SM), which is used to generate each M_e, and robust control (RC) techniques, which are used to solve the constrained optimizations. The mathematical details can be found in ref 17. Briefly, optimization of a general-form objective function subject to general-form constraints is known to be a "hard" numerical problem (see ref 16 and references therein). However, it turns out that, if the constraints (and the objective function) can be expressed as polynomials, one can employ recent RC techniques to develop computationally efficient algorithms of optimization. This is the essence of our approach.

We develop quadratic approximations for each dataset-unit model M_e, using the SM methodology:[4,8,9] identification of active parameters $\mathscr{X}_e$ via sensitivity analysis and development of a quadratic response surface via computer experiments arranged according to a factorial design.[3,22] The new developments for the problem of dataset consistency, along with the necessary details, are given in the next section.

3. Dataset Consistency

3.1. Problem Formulation. Given a dataset, we are now interested in determining whether the data it contains are mutually consistent. This interest is motivated by practical questions such as establishing whether a given reaction model is in agreement with the available experimental observations, recognizing data outliers, or identifying the source of disagreement between the model and experiment. The framework outlined below develops quantitative measures to address such questions in rigorous terms.

We begin by introducing the following definition: a dataset $\mathscr{D}$ (together with its corresponding prior information) is said to be *inconsistent* if there is no single point $\boldsymbol{x}$ in the hypercube $\mathscr{H}$ that satisfies $l_e \le M_e(\boldsymbol{x}_e) - d_e \le u_e$ for all e in $\mathscr{E}$. Otherwise, the dataset is *consistent*. In other words, the dataset is inconsistent if the feasible region is empty. The mathematical development that follows is aimed, in essence, at determining whether the feasible region for the constraints implied by a given dataset is empty or not.

The dataset constraints are represented by four sets of inequalities:

from the prior information on $\mathscr{H}$:

$$\begin{cases} -x_j \le -\alpha_j \quad (\text{for } j = 1, 2, ..., n) \\ x_j \le \beta_j \end{cases}$$

and from the data set units:

$$\begin{cases} -M_e(\boldsymbol{x}_e) + d_e \le -l_e \quad (\text{for each } e \text{ in } \mathscr{E}) \\ M_e(\boldsymbol{x}_e) - d_e \le u_e \end{cases} \tag{1}$$

By the definition established in the aforementioned discussion, the dataset is inconsistent if no single $\boldsymbol{x}$ satisfies all of these constraints. To condense the notation, we form vectors $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n)^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_n)^T$, $\boldsymbol{l} = (l_{e_1}, l_{e_2}, ..., l_{e_m})^T$, and $\boldsymbol{u} = (u_{e_1}, u_{e_2}, ..., u_{e_m})^T$, where m denotes the number of dataset units in the dataset (i.e., the size of $\mathscr{E}$). In accord with this notation, we refer to the four collections of inequalities in eq 1 as the $\boldsymbol{\alpha}$-, $\boldsymbol{\beta}$-, $\boldsymbol{l}$-, and $\boldsymbol{u}$-constraints.

TABLE 1: Example Dataset $\mathscr{D} = \{\mathscr{U}_1, \mathscr{U}_2, \mathscr{U}_3\}$

	$\mathscr{X}_e$	$M_e(\boldsymbol{x}_e)$	d_e	l_e	u_e
$e = 1$	$\{X_2\}$	x_2	0.75	−0.25	0.25
$e = 2$	$\{X_1, X_2\}$	$x_2 - x_1 + 1$	0.7	−0.25	0.25
$e = 3$	$\{X_1, X_2\}$	$x_1 + x_2 - 0.1$	0.6	−0.25	0.25

The values of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{l}$, and $\boldsymbol{u}$ affect the consistency of a dataset. Indeed, a consistent dataset may become inconsistent with a decrease in the intervals of active variables (affecting $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) and/or in the levels of experimental error (affecting $\boldsymbol{l}$ and $\boldsymbol{u}$). The present status, even in better established fields, is such that the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ values are not well-established and those of $\boldsymbol{l}$ and $\boldsymbol{u}$ are seldom documented. Given this situation, we can consider values of $\boldsymbol{l}$ and $\boldsymbol{u}$ to be very tentative, and by varying them, answer questions such as, "at what level of the experimental error does the dataset become inconsistent?" In fact, this very question leads us to the definition of a consistency measure, as described in the next subsections.

3.2. Pairwise Consistency. Before discussing our final results of how we determine if a dataset is consistent, we introduce a simpler, easier-to-visualize test. In this test, we consider pairs of dataset units $(\mathscr{U}_e, \mathscr{U}_f)$ for e, f (not necessarily distinct) in the index set $\mathscr{E}$. For each pair, we compute the minimum level of uncertainty, u_{ef}, that leaves the two-element dataset $\mathscr{D}_{ef} = \{\mathscr{U}_e, \mathscr{U}_f\}$ consistent:

$$\begin{aligned} u_{ef} = {} & \text{the minimum value of } u \text{ such that the constraints} \\ & |M_e(\boldsymbol{x}_e) - d_e| \le u \text{ and } |M_f(\boldsymbol{x}_f) - d_f| \le u \\ & \text{are satisfied by some value } \boldsymbol{x} \text{ in } \mathscr{H} \end{aligned}$$

If this *threshold* uncertainty u_{ef} does not exceed the reported uncertainty bounds of the corresponding two dataset units, i.e., if $u_{ef} < -l_e, u_e, -l_f, u_f$, then the two-element dataset $\mathscr{D}_{ef}$ is consistent, or, in other words, the two dataset units $\mathscr{U}_e$ and $\mathscr{U}_f$ are mutually consistent with each other. Continuing in this manner, we calculate u_{ef} for each pair of dataset units in the dataset, i.e., for all e, f in $\mathscr{E}$. If at least one of the two-element datasets is pairwise inconsistent, the entire dataset is inconsistent. The converse is not true; i.e., even if all two-element datasets are pairwise consistent, it does not necessarily imply that the entire dataset is consistent. We demonstrate such a case with the following numerical example.

Consider a dataset containing three dataset units, the contents of which are listed in Table 1. The prior information is $\alpha_j = 0$, $\beta_j = 1$ for $j = 1, 2$. In other words, $\mathscr{H}$ is the unit square. This example dataset exhibits pairwise consistency. Indeed, $u_{1,2} = u_{1,3} = 0.025$ and $u_{2,3} = 0$, which are well below the allowable uncertainties of ±0.25 in the three dataset units. However, the dataset comprised of all three dataset units is inconsistent. The three shaded regions in Figure 1 depict, for each dataset unit, the set of points that lead to model predictions within the experimental uncertainties. In other words, the three displayed regions are the feasible regions for the singleton datasets: $\mathscr{D}_1 = \{\mathscr{U}_1\}$, $\mathscr{D}_2 = \{\mathscr{U}_2\}$, and $\mathscr{D}_3 = \{\mathscr{U}_3\}$, respectively. It is easily seen that any two of these regions overlap, implying pairwise consistency. Yet, the three regions share no common points and hence have an empty intersection, as illustrated in Figure 2. Consequently the entire dataset $\mathscr{D}_{123} = \{\mathscr{U}_1, \mathscr{U}_2, \mathscr{U}_3\}$ is inconsistent. This example demonstrates that the entire dataset must be considered to determine consistency. A technique that addresses this is described next.

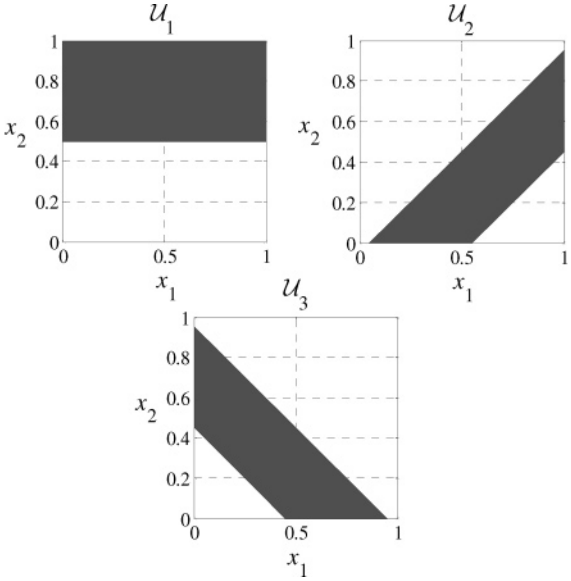

Figure 1. Feasible regions (shaded) for the three dataset units of the example dataset.

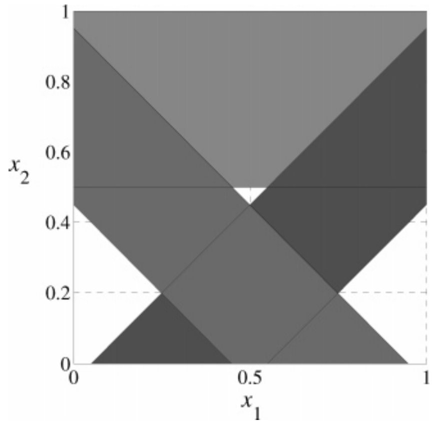

Figure 2. Overlay of the feasible regions for the three dataset units of the example dataset. Any two of the regions overlap, implying pairwise consistency. However, the three regions do not have a point in common; thus, the dataset $\mathscr{D} = \{\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3\}$ is inconsistent.

3.3. Measure of Dataset Consistency. We have defined a dataset to be consistent if there is at least one $\boldsymbol{x} \in \mathscr{H}$ such that, for each e, $M_e(\boldsymbol{x}_e)$ is contained in the interval $[l_e + d_e, u_e + d_e]$. To determine if a dataset is consistent and quantify the degree of consistency, we compute the smallest such intervals that contain $M_e(\boldsymbol{x}_e)$ for some $\boldsymbol{x} \in \mathscr{H}$. This is accomplished by introducing flexibility into the $\boldsymbol{l}$- and $\boldsymbol{u}$-constraints by adding to each l_e and subtracting from each u_e a slack variable γ. Let $C_{\mathscr{D}}$ denote the largest γ for which some value $\boldsymbol{x}$ satisfies both the $\boldsymbol{\alpha}$- and $\boldsymbol{\beta}$-constraints and the *flexible* $\boldsymbol{l}$- and $\boldsymbol{u}$-constraints, i.e.,

$C_{\mathscr{D}} =$ maximum value of γ subject to the constraints:

$$\begin{cases} -x_j \le -\alpha_j & (\text{for } j = 1, 2, ..., n) \\ x_j \le \beta_j & \\ -M_e(\boldsymbol{x}_e) + d_e \le -l_e - \gamma & (\text{for each } e \text{ in } \mathscr{E}) \\ M_e(\boldsymbol{x}_e) - d_e \le u_e - \gamma & \end{cases} \qquad (2)$$

Defined in this way, $C_{\mathscr{D}}$ values greater than zero imply that the $\boldsymbol{\alpha}$-, $\boldsymbol{\beta}$-, $\boldsymbol{l}$-, and $\boldsymbol{u}$-constraints introduced in eq 1 are satisfied, so the dataset is consistent. The magnitude of $C_{\mathscr{D}}$ provides a measure of the relative consistency (or inconsistency) of the dataset, with larger values of $C_{\mathscr{D}}$ indicating enhanced consistency. For this reason, we refer to $C_{\mathscr{D}}$ as the consistency measure of the dataset $\mathscr{D}$.

Solving the proposed optimization for the consistency measure is not easy. We require the global maximum of γ, which makes the problem computationally complex. Often, the best one can do is to compute two bounds, $\underline{C}_{\mathscr{D}}$ and $\bar{C}_{\mathscr{D}}$, that satisfy $\underline{C}_{\mathscr{D}} \le C_{\mathscr{D}} \le \bar{C}_{\mathscr{D}}$. As a consequence of bounding the value of the consistency measure, we are left with three scenarios: a value of $\underline{C}_{\mathscr{D}} \ge 0$ implies that $\mathscr{D}$ is consistent; a value of $\bar{C}_{\mathscr{D}} < 0$ implies that $\mathscr{D}$ is inconsistent; and lastly, if $\underline{C}_{\mathscr{D}} < 0 \le \bar{C}_{\mathscr{D}}$, the consistency test is inconclusive. When the consistency test is conclusive, constrained optimization techniques, together with the concept of a dataset, allow us to rigorously determine if the individual experiments that comprise the dataset are mutually consistent with each other, within the proposed kinetic model.

3.4. Threshold Uncertainty. In the development of the dataset consistency measure, we view the uncertainty bounds $\boldsymbol{l}$ and $\boldsymbol{u}$ as fixed quantities (specified experimentally) and ask whether the dataset is consistent at this uncertainty level. Alternatively, we could consider the experimental uncertainties as variable quantities and compute the level at which the dataset switches from being consistent to inconsistent. We refer to this transition point as the *threshold uncertainty*. Mathematically, for each experiment E, the threshold uncertainty is defined as $l_{\text{thresh},e} = l_e + C_{\mathscr{D}}$ and $u_{\text{thresh},e} = u_e - C_{\mathscr{D}}$, with the bounds on $C_{\mathscr{D}}$ translating into bounds on the threshold uncertainties, $\underline{l}_{\text{thresh},e} = l_e + \underline{C}_{\mathscr{D}}$, $\bar{l}_{\text{thresh},e} = l_e + \bar{C}_{\mathscr{D}}$, $\underline{u}_{\text{thresh},e} = u_{\text{e}} - \bar{C}_{\mathscr{D}}$, and $\bar{u}_{\text{thresh,e}} = u_e - \underline{C}_{\mathscr{D}}$.

It is pertinent to mention that the threshold uncertainties, along with dataset consistency, may change with the variation in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. One may repeat the uncertainty analysis with new values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ or perform sensitivity analysis, which is developed in the remainder of this section, to assess the possible effect.

3.5. Lagrange Multipliers. The consistency measure of a dataset is dependent on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{l}$, and $\boldsymbol{u}$. As mentioned in the problem formulation subsection, even for highly researched systems, these values may be and usually are tentative. Therefore, it is useful to estimate the dependence of the introduced consistency measure on the nominal values of these bounds. We gauge this dependence using Lagrange multipliers, which reveal the sensitivity of the solution of an optimization problem (in our case, $C_{\mathscr{D}}$) to the constraints. A technical discussion of these methods lies outside the scope of this paper; the relevant development for the consistency question is given in the Appendix.

The key result used in the present work is the following. Let $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\beta}}$, let $\tilde{\boldsymbol{l}}$ and $\tilde{\boldsymbol{u}}$ denote bounds that possibly differ from the nominal values, and let $C_{\tilde{\mathscr{D}}}$ represent the resulting consistency measure. If we perturb the values of the uncertainty and parameter bounds from their nominal values, the change in the consistency measure is upper-bounded by a linear function of the perturbations. Specifically, the deviation in the consistency measure, $\Delta C_{\mathscr{D}} = C_{\tilde{\mathscr{D}}} - C_{\mathscr{D}}$, satisfies

$$\Delta C_{\mathscr{D}} \le \bar{C}_{\mathscr{D}} - C_{\mathscr{D}} + \sum_{j=1}^{n} (\lambda_j^{(\alpha)} \Delta\alpha_j + \lambda_j^{(\beta)} \Delta\beta_j) + \sum_{e \in \mathscr{E}} (\lambda_e^{(l)} \Delta l_e + \lambda_e^{(u)} \Delta u_e) \qquad (3)$$

where $\Delta\alpha_j = \tilde{\alpha}_j - \alpha_j$ and similar relationships exist for $\Delta\beta_j$, Δl_e, and Δu_e. The scalars $\lambda_j^{(\alpha)}$, $\lambda_j^{(\beta)}$, $\lambda_e^{(l)}$, and $\lambda_e^{(u)}$ are the Lagrange multipliers. Their values are determined by the optimization procedure and satisfy $\lambda_j^{(\alpha)}$, $\lambda_e^{(l)} \le 0$ and $\lambda_j^{(\beta)}$, $\lambda_e^{(u)} \ge 0$. These inequalities are intuitive: for instance, in accord with $\lambda_j^{(\alpha)}$ being negative, a decrease in α_j should increase the consistency measure, because more parameters values are allowed. The inequalities describing $\lambda_j^{(\beta)}$, $\lambda_e^{(l)}$, and $\lambda_e^{(u)}$ follow similar reasoning.

Equation 3 suggests two applications of the Lagrange multipliers, which we describe in the following two subsections. First, the Lagrange multipliers indicate the local sensitivity of the consistency measure to components of the nominal bounds $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{l}$, and $\boldsymbol{u}$. Second, when the dataset is provably inconsistent by our methods, i.e., $\bar{C}_{\mathscr{D}} < 0$, the nonzero Lagrange multipliers identify conflicting bounds. The latter we reference hereafter as the *global outcome*.

3.6. Local Sensitivity. In determining $C_{\mathscr{D}}$, the bounds $\underline{C}_{\mathscr{D}}$ and $\bar{C}_{\mathscr{D}}$ are computed. When these bounds are close to each other, $C_{\mathscr{D}} \approx \bar{C}_{\mathscr{D}}$, and eq 3 reduces to

$$\Delta C_{\mathscr{D}} \lesssim \sum_{j=1}^{n}(\lambda_j^{(\alpha)}\Delta\alpha_j + \lambda_j^{(\beta)}\Delta\beta_j) + \sum_{e\in\mathscr{E}}(\lambda_{\mathrm{e}}^{(l)}\Delta l_e + \lambda_e^{(u)}\Delta u_e) \quad (4)$$

It is useful to condense the notation by defining two column vectors, the array of bounds $\boldsymbol{t} = (\alpha_1, ..., \alpha_n, \beta_1, ..., \beta_n, l_{e_1}, ..., l_{e_m}, u_{e_1}, ..., u_{e_m})^{\mathrm{T}}$ and the array of Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_1^{(\alpha)}, ..., \lambda_n^{(\alpha)}, \lambda_1^{(\beta)}, ...)^{\mathrm{T}}$. In the new notation, eq 4 becomes

$$\Delta C_{\mathscr{D}} \lesssim \boldsymbol{\lambda}^{\mathrm{T}}\Delta\boldsymbol{t} \quad (5)$$

where $\Delta\boldsymbol{t} = \tilde{\boldsymbol{t}} - \boldsymbol{t}$. Based on our experience (e.g., with the GRI-Mech dataset discussed in the next section), eq 5 approximately holds with equality when each component of $\Delta\boldsymbol{t}$ is small, i.e.,

$$\Delta C_{\mathscr{D}} \approx \boldsymbol{\lambda}^{\mathrm{T}}\Delta\boldsymbol{t} \quad (6)$$

Equation 6 approximates how small perturbations about the nominal value $\boldsymbol{t}$ affect the consistency measure. Therefore, the components of $\boldsymbol{\lambda}$ indicate the local sensitivity of the consistency measure to the corresponding components of $\boldsymbol{t}$.

The components of $\boldsymbol{t}$ provide estimates of the experimental uncertainties and of intervals that contain the parameter values. If a dataset is inconsistent and its consistency measure is highly sensitive to only a few components of $\boldsymbol{t}$, it is not unreasonable to question if these components of $\boldsymbol{t}$ are realistic bounds on the corresponding measurement uncertainties or parameter bounds from the prior information. Perhaps the dataset would be consistent if more-judicious bounds were used. Consequently, the corresponding experimental results and components of prior information should be examined first before questioning the ODE kinetic model that generated the dataset unit models M_e. On the other hand, if the dataset is inconsistent and none of the elements of $\boldsymbol{t}$ stand out as having a strong influence on the consistency measure, the possibility that the ODE kinetic model is the source of the inconsistency deserves more-careful scrutiny.

3.7. Global Outcome. The components of $\boldsymbol{\lambda}$ provide more information than just an indication of local sensitivity. In contrast to the experience with reaction kinetics, where a zero local sensitivity does not necessarily indicate a reaction is unimportant,[4] in the present case, it has global implications: for *finite* changes to components of $\boldsymbol{t}$ where $\boldsymbol{\lambda}$ is zero, the consistency measure is less than $\bar{C}_{\mathscr{D}}$. Consequently, constraints associated with the zero components of $\boldsymbol{\lambda}$ can be completely ignored without improving consistency.

To develop this global result, we form a new dataset from $\mathscr{D}$ by allowing the components of $\boldsymbol{\alpha}$ and $\boldsymbol{l}$ to which $C_{\mathscr{D}}$ has zero local sensitivity to be arbitrarily negative and letting the components of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ to which $C_{\mathscr{D}}$ has zero local sensitivity be arbitrarily large. Specifically, a new dataset $\tilde{\mathscr{D}}$ is generated from $\mathscr{D}$ by setting

$$\tilde{\alpha}_j = \begin{cases} -\infty & (\text{if } \lambda_j^{(\alpha)} = 0) \\ \alpha_j & (\text{otherwise}) \end{cases}$$
$$\tilde{\beta}_j = \begin{cases} \infty & (\text{if } \lambda_j^{(\beta)} = 0) \quad (\text{for } j = 1, 2, ..., n) \\ \beta_j & (\text{otherwise}) \end{cases}$$
$$\tilde{l}_e = \begin{cases} -\infty & (\text{if } \lambda_e^{(l)} = 0) \\ l_e & (\text{otherwise}) \end{cases}$$
$$\tilde{u}_e = \begin{cases} \infty & (\text{if } \lambda_e^{(u)} = 0) \quad (e \text{ in } \mathscr{E}) \\ u_e & (\text{otherwise}) \end{cases} \quad (7)$$

(Here, we use the convention $0 \cdot \infty = 0$, and note that the maximum in eq 2 must be replaced with a supremum when $\boldsymbol{x}$ is allowed to take unbounded values.) By this construction of $\tilde{\mathscr{D}}$, we have effectively eliminated each constraint in eq 2 to which the consistency measure has zero local sensitivity. $\tilde{\mathscr{D}}$ may then be viewed as a "smaller" dataset than $\mathscr{D}$, obtained after eliminating from the original dataset the information to which the consistency measure is insensitive.

The following inequality, which we call the *global result*, is proven in the Appendix:

$$C_{\mathscr{D}} \le C_{\tilde{\mathscr{D}}} \le \bar{C}_{\mathscr{D}} \quad (8)$$

In terms of the threshold uncertainties, eq 8 translates to

$$\tilde{l}_{\mathrm{thresh},e} \le \bar{l}_{\mathrm{thresh},e} \qquad (\text{for all } e \text{ in } \mathscr{E}) \quad (9a)$$

$$\underline{u}_{\mathrm{thresh},e} \le \tilde{u}_{\mathrm{thresh},e} \qquad (\text{for all } e \text{ in } \mathscr{E}) \quad (9b)$$

where $\tilde{l}_{\mathrm{thresh},e}$ and $\tilde{u}_{\mathrm{thresh},e}$ are the threshold uncertainties for the dataset $\tilde{\mathscr{D}}$. The relationships in eqs 8 and 9 indicate there is a limit to how much the consistency measure (or the threshold uncertainties) can improve if all constraints to which the original consistency measure $C_{\mathscr{D}}$ is insensitive are removed. Although these two equations hold in generality, they are of foremost interest when the dataset is provably inconsistent ($\bar{C}_{\mathscr{D}} < 0$). We demonstrate the utility of eq 9 by focusing on the upper bounds $\boldsymbol{u}$; the interpretation for the lower bounds $\boldsymbol{l}$ is analogous.

Suppose the dataset is provably inconsistent. Then, for each e, $u_e < \underline{u}_{\mathrm{thresh},e}$. Equation 9 indicates that, upon eliminating all constraints to which $C_{\mathscr{D}}$ is insensitive, the threshold uncertainties $\tilde{u}_{\mathrm{thresh},e}$ are still larger than the acceptable uncertainties u_e. This means that a problem must exist in the constraints that correspond to the nonzero Lagrange multipliers, because effectively eliminating the remaining constraints will not bring the threshold uncertainties down to an acceptable level. The key result can be summarized as follows. The removal of all constraints corresponding to the zero components of $\boldsymbol{\lambda}$ from a provably inconsistent dataset leaves $C_{\mathscr{D}}$ still negative. The source of the inconsistency thus lies with the remaining constraints.

4. Test Case

4.1. GRI-Mech Dataset. We demonstrate the outlined methodology on a real-world example, the GRI-Mech 3.0 dataset,[20] which is taken from the field of combustion chemistry. It is a collaborative data repository for the development of detailed kinetic models for pollutant formation in the combustion of natural gas and was used in our previous studies.[15,16]

The GRI-Mech 3.0 dataset is composed of 77 dataset units, each of which represents an individual observation or a "representation" for a group of them; e.g., an average of several observations or a temperature dependence for a series of measurements. The experimental apparatuses include shock tubes, flow reactors, stirred reactors, and laminar premixed flames. The properties of interest (Y_e) of the dataset units include species concentrations, ignition delays, laminar flame velocities, shifts in peak positions, etc.

The overall kinetic model is comprised of 325 reversible "elementary" reactions among 53 chemical species. This model is presumed to be complete enough to simulate, in a physically plausible manner, the experiments of the dataset units. In other words, it is presumed that the reaction mechanism, and, hence, its formulation as an ODE system, is known and that the uncertainty of the model predictions arises solely from uncertainty in the values of the model parameters.

There are more than 650 total parameters: reaction rate constants, their activation energies, species thermodynamic and transport properties, as well as instrumental constants, i.e., absorption coefficients and the like. Not all of these are active or assumed to be active in the present dataset. Only a fraction of them, specifically 102, are considered as the overall set of the dataset active parameters. Their selection is based on the following considerations.

Each dataset unit model represents a parametrization of the overall ODE model for the specific conditions of that particular dataset unit. Namely, the numerical response of the kinetic model for the conditions of the dataset unit is expressed as a simple algebraic function (a quadratic polynomial in our case) of parameters active at these conditions. The latter are determined by a screening sensitivity analysis and consideration of their uncertainty ranges. For instance, the rate constant of the reaction $H + O_2 \rightarrow OH + O$, although making the top of the list on the sensitivity chart, is excluded from the active parameter list, because its value is sufficiently well-known. On the other hand, the rate constant of the reaction $CH_3 + HO_2 \rightarrow OH + CH_3O$, which has a relatively low sensitivity, is an active parameter, because of its large range of uncertainty. The union of all active parameters comprise the 102-member set of the dataset active parameters.

4.2. Consistency Analysis. For simplicity in demonstration of the technique, and in light of insufficient records of experimental uncertainties $\boldsymbol{l}$ and $\boldsymbol{u}$, even for such a well-documented case as GRI-Mech 3.0, an artificial but realistic uniform level of experimental uncertainties $-l_e = u_e =$ constant for all e in $\mathcal{E}$, was used in the present analysis (hence, $-l_{\mathrm{thresh},e} = u_{\mathrm{thresh},e}$ for all e). For $-l_e = u_e = 0.087$, the GRI-Mech 3.0 dataset is consistent. For uncertainty levels of <0.083, the dataset is inconsistent. In the following, we describe how this was determined and how insights provided by these results happened to identify a "typo" in the original data entry. We begin by discussing the pairwise consistency and $C_{\mathcal{D}}$ measure developed previously. We then identify the contributing factors to the dataset inconsistency and conclude with the specific implications of the consistency analysis for the GRI-Mech 3.0 dataset.

4.2.1. Pairwise Consistency. We begin by examining the pairwise consistency of the GRI-Mech dataset. The results of the test are displayed in Figure 3. The height of each bar protruding from the (e,f) coordinates of the $u = 0$ plane represents the minimum level of uncertainty for which the two-element dataset $\{\mathcal{U}_e, \mathcal{U}_f\}$ is consistent. For many pairs, this height is negligibly small, indicating that there are values $\boldsymbol{x}$ in $\mathcal{H}$ for which $M_e(\boldsymbol{x}_e)$ is essentially d_e and $M_f(\boldsymbol{x}_f)$ is essentially d_f. The two "walls" along $e,f = 25$ occur because $|M_{25}(\boldsymbol{x}_{25}) - d_{25}| \geq 0.008$ for every value $\boldsymbol{x}$ in $\mathcal{H}$; therefore, u_{ef} must be at least 0.008 for any pair $\{\mathcal{U}_e, \mathcal{U}_f\}$ that contains $\mathcal{U}_{25}$. Another noteworthy feature in Figure 3 is the emergence of a few large peaks. One of them is for the $\{\mathcal{U}_{57}, \mathcal{U}_{58}\}$ pair, showing that an uncertainty level of at least 0.082 (10-fold larger than that for the "wall") is required for this pair of dataset units to be mutually consistent. The extreme magnitude of these few outstanding peaks signals a possible cause for concern regarding the affected experiments, and, as will be shown below, this concern is not without a reason.

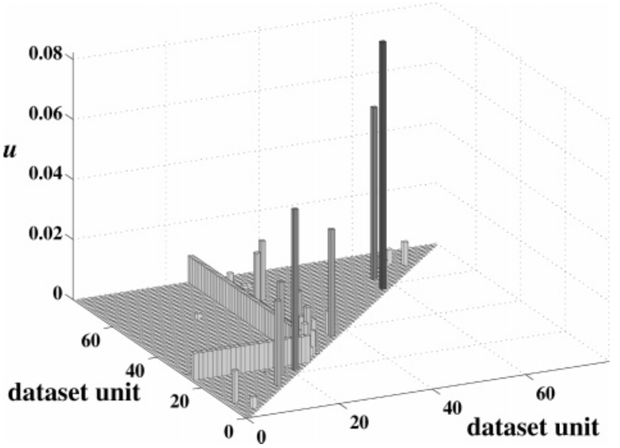

Figure 3. Threshold uncertainty levels u_{ef} calculated for each $\mathcal{D}_{ef} = \{\mathcal{U}_e, \mathcal{U}_f\}$ pair of the GRI-Mech dataset units. The highest peak is $u_{57,58}$, flagging $\mathcal{U}_{57}$ and/or $\mathcal{U}_{58}$ as possible outliers from a pairwise perspective.

4.2.2. Consistency Measure. We now turn to the analysis of consistency of the GRI-Mech dataset as a whole. Consistency of the dataset is determined by the sign of its consistency measure $C_{\mathcal{D}}$. For the given case of the GRI-Mech 3.0 dataset with $-l_e = u_e = 0.08$ for all e in $\mathcal{E}$, the computed bounds are $\underline{C}_{\mathcal{D}} = -0.0065$ and $\bar{C}_{\mathcal{D}} = -0.0033$, indicating that the consistency measure $C_{\mathcal{D}}$ is contained in the interval $[-0.0065, -0.0033]$. Because $C_{\mathcal{D}}$ is negative, the dataset is inconsistent at the 0.08 level of experimental uncertainty. In the present case, the threshold uncertainty lies in the interval $[0.08 - \bar{C}_{\mathcal{D}}, 0.08 - \underline{C}_{\mathcal{D}}] = [0.0833, 0.0865]$. This implies that, for an uncertainty level of <0.0833, the dataset is inconsistent, and for an uncertainty level of >0.0865, it is consistent.

4.2.3. Lagrange Multipliers. Having determined that the GRI-Mech 3.0 dataset becomes inconsistent at $-l_e = u_e = 0.083$, we examine what causes this inconsistency. We ask the following question: Do all or most of the dataset units contribute more or less equally to the inconsistency of the dataset as a whole, or is there, among the dataset units, a single experiment or a small group of them that brings about this outcome? The numerical apparatus presented in the previous section allows us to answer all such questions.

Our analysis is based on the use of $\boldsymbol{\lambda}$, which is the vector of Lagrange multipliers. The components of this vector are shown in Figure 4. The top two panels of Figure 4 display sensitivities of the consistency measure $C_{\mathcal{D}}$, with respect to the components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and the two bottom panels show values with respect to those of $\boldsymbol{l}$ and $\boldsymbol{u}$. Examination of these sensitivity values leads to the following observations.

First, we note that most components of $\boldsymbol{\lambda}$ are zero. This is in accord with the effect sparsity,[3,21] which states that usually only a small number of model parameters have a significant effect on model responses. This phenomenon has been observed in

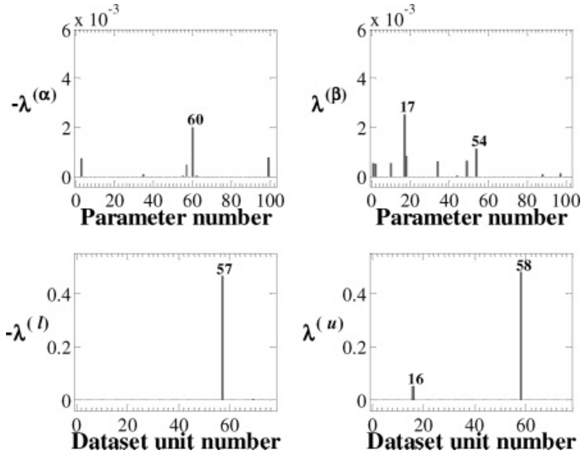

Figure 4. Lagrange multipliers computed for the nominal GRI-Mech dataset; $\lambda_{57}^{(l)}$ and $\lambda_{58}^{(u)}$ have significantly larger magnitudes than the remaining components of $\boldsymbol{\lambda}$, indicating the dataset consistency measure is most sensitive to l_{57} and u_{58}.

numerous sensitivity studies of the role played by individual reactions of (large) chemical reaction networks. The present results show that this also seems to be the case for the error analysis. Furthermore, the global nature of the sensitivity implies that, for the consistency measure to improve (increase), modifications must be made to components of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{l}$, and $\boldsymbol{u}$ that correspond to the nonzero components of $\boldsymbol{\lambda}$. Because most components of $\boldsymbol{\lambda}$ are zero, only a small number of uncertainty bounds might improve dataset consistency upon revision.

Our next observation is that the sensitivity values in the top two panels are very small—much smaller than those in the bottom panels. Evidently, the prior information (i.e., the size of the initial hypercube and, in turn, the presumed knowledge of the model active parameters) does not have a significant influence on the consistency measure in the present case. As a numerical test, we recomputed $\underline{C}_{\mathcal{D}}$ and $\bar{C}_{\mathcal{D}}$ with a 10% enlargement of the prior information intervals. This modification increased $\bar{C}_{\mathcal{D}}$ from -0.0033 to -0.0010. Although an improvement, as judged by the increase in the value of $\bar{C}_{\mathcal{D}}$, the dataset remained inconsistent, because $\bar{C}_{\mathcal{D}}$ is still negative. Thus, from the low sensitivity to all components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we conclude that the uncertainties in the model parameters are unlikely to dominate the present inconsistency in the GRI-Mech dataset.

On the other hand, the peak sensitivities in the two bottom panels are large. The two largest peaks are the sensitivities of the consistency measure, with respect to the lower bound l_{57} of the experimental uncertainty of dataset unit 57 and the upper bound u_{58} of the experimental uncertainty of dataset unit 58 (bottom left and right panels of Figure 4, respectively). These same two dataset units surfaced in the pairwise test. Both facts gave us reason to suspect $\mathscr{U}_{57}$ and/or $\mathscr{U}_{58}$ as possible outliers. It turned out, as described next, this suspicion was justified.

4.3. Resolution of the Inconsistency. The two suspected experimental values are both reaction times to reach half of the maximum in OH concentration determined in the shock tube oxidation of methane.[20] The dataset unit $\mathscr{U}_{57}$ designates the experimental target labeled "OH.1a" for the initial conditions of 0.1% methane−0.2% oxygen−argon mixture at a pressure of 1 atm and a temperature of 2000 K, and $\mathscr{U}_{58}$ designates the target labeled "OH.1b" for the same mixture at a temperature of 2200 K. The respective experimental values[20] are $d_{57} = 970$ and $d_{58} = 218$ μs. These were not actual measurements but two "representative" points, at 2000 and 2200 K, extrapolated from a fit to a series of measurements performed over a range of temperatures.[23]

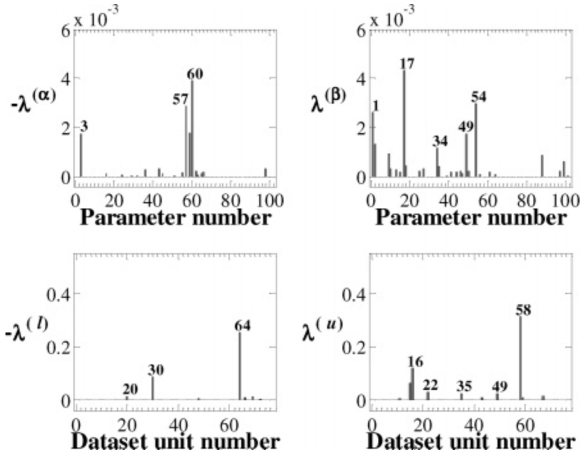

Figure 5. Lagrange multipliers after modifying d_{57}. $\lambda_{58}^{(u)}$ remains relatively large, but $\lambda_{57}^{(l)}$ and $\lambda_{57}^{(u)}$ are zero, indicating that the consistency measure is no longer sensitive to the uncertainty bounds of $\mathscr{U}_{57}$.

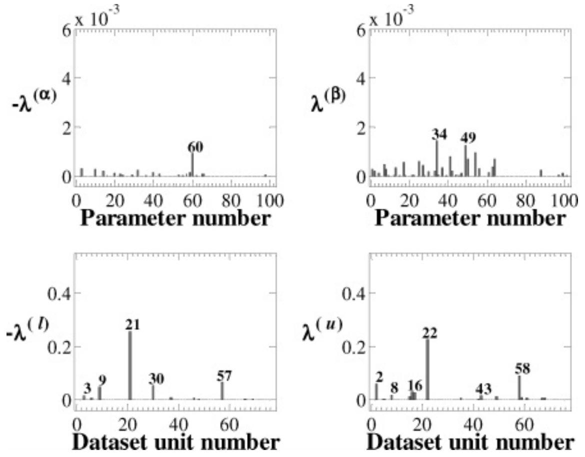

Figure 6. Lagrange multipliers after modifying d_{58}, but leaving d_{57} at the nominal value. Now, both the $\lambda_{57}^{(l)}$ and $\lambda_{58}^{(u)}$ peaks have decreased and additional sensitivity peaks have developed.

When the initial results of the GRI-Mech consistency analysis became known, we contacted the researchers[24] who originated the data for d_{57} and d_{58}. The only information we provided was that the results had been flagged by our analysis, giving no indication of which direction the measurements might be in error. Re-examination of the original observations led the researchers to modify the fit with new extrapolated values of $d_{57} = 700$ and $d_{58} = 255$ μs.[24] The directions of these changes, the decrease in d_{57}, and the increase in d_{58} are precisely those indicated by the signs of the sensitivity values: $\lambda_{57}^{(l)} = -0.4643$ and $\lambda_{58}^{(u)} = 0.4805$.

We repeated the consistency analysis using the revised values of d_{57} and d_{58}, one at a time and then both together. The results are displayed in Figures 5−7 and Table 2. The increased consistency measures and the corresponding decrease in the threshold uncertainties listed in Table 2 both indicate that the revision improves the consistency of the GRI-Mech 3.0 dataset. When both d_{57} and d_{58} are updated, $C_{\mathcal{D}}$ becomes strictly positive, indicating that the dataset initially inconsistent at $-l_e = u_e = 0.08$ for each index e becomes consistent. Another way of looking at this result is to note that the threshold uncertainty, i.e., the level of experimental uncertainty at which $C_{\mathcal{D}}$ changes its sign, decreases from the initial average of 0.0849 to 0.0673, which is a 20% improvement in the dataset consistency.

TABLE 2: Results of GRI-Mech Dataset Update: Upper and Lower Bounds

	nominal dataset	d_{57} update	d_{58} update	d_{57} and d_{58} update
$C_{\mathscr{D}}$ ($u = 0.08$)	[−0.0065, − 0.033]	[−0.0017, 0.0085]	[0.0084, 0.0213]	[0.0056, 0.0198]
u_{thresh}	[0.0833, 0.0865]	[0.0715, 0.0817]	[0.0587, 0.0716]	[0.0602, 0.0744]

A closer examination of Figures 5−7 reveals some interesting trends—the type of trends that may assist the researcher in gaining a deeper understanding of the data and the model. Removing the primary source of inconsistency by revising the two dataset units reduced the maximum values of sensitivities to experimental uncertainties $\boldsymbol{l}$ and $\boldsymbol{u}$ (compare bottom panels of Figures 4−7). Updating only one dataset unit, namely d_{57} of $\mathscr{U}_{57}$, increased the sensitivities to parameter bounds, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (compare top panels of Figures 4 and 5). Comparing Figures 5 and 6, we notice that updating d_{58} alone has a deeper improvement than updating d_{57} alone, and updating the two together more resembles the former case, thereby hinting at a bigger problem with d_{58}. After the dataset is updated, the peak sensitivities decrease and the number of (now smaller) peaks increases, which implies that the dataset consistency measure becomes less sensitive to individual constraints. This indicates that the consistency level of the dataset no longer hinges on a few outliers; instead, the dependence becomes distributed over the content of the dataset.

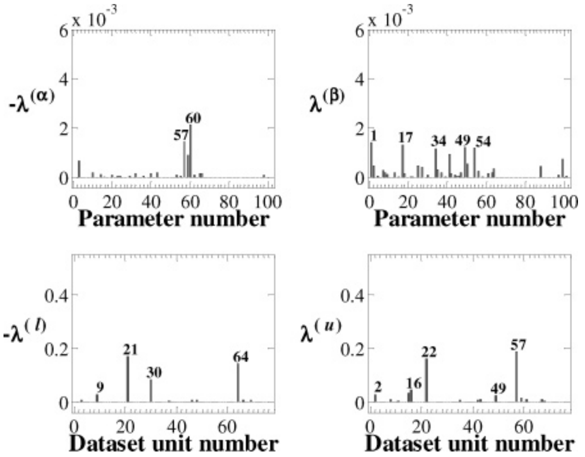

Figure 7. Lagrange multipliers after modifying both d_{57} and d_{58}. Overall, the sensitivity levels have decreased and the consistency measure is now sensitive to a larger number of dataset units. This implies that the dataset consistency measure is no longer dominated by just a few outliers.

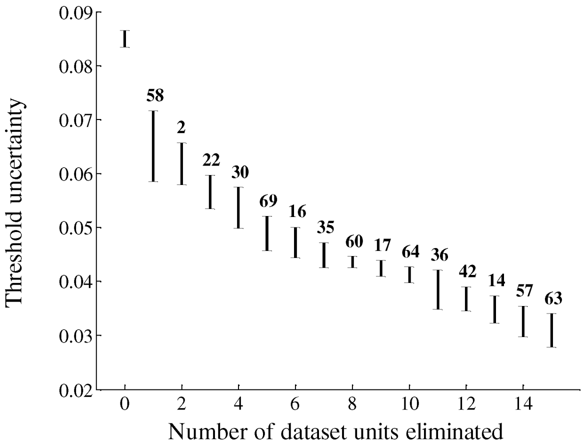

Figure 8. Ranges of the threshold uncertainty (vertical bars) computed by sequential and cumulative elimination of dataset units from the dataset. At each step, the listed dataset unit, which is the one that most inhibits dataset consistency, is removed. The lower and upper bounds of the range are the threshold uncertainty levels at which the dataset can be proven to become inconsistent, $\bar{C}_{\mathscr{D}} < 0$, or consistent, $\underline{C}_{\mathscr{D}} \geq 0$, respectively.

4.4. Sequential Analysis. The aforementioned observations suggest a possible approach to dataset analysis through sequential identification and removal of the top outliers. Figure 8 depicts several steps of such analysis for the GRI-Mech 3.0 dataset. Shown are intervals of threshold uncertainties obtained in the following sequential steps: the original dataset; removing $\mathscr{U}_{58}$ from the dataset; removing $\mathscr{U}_{58}$ and $\mathscr{U}_{2}$, identified to be the next top outlier at this stage; removing $\mathscr{U}_{58}$, $\mathscr{U}_{2}$, and $\mathscr{U}_{22}$; and so on, as indicated in the figure.

The results in Figure 8 seem to exhibit the effect sparsity: after several steps, with a significant decline at each step, the threshold uncertainty begins to level off. Such an outcome implies that the consistency of the dataset can be dramatically improved by revisiting just a few dataset units. Whether this eventually identifies a possible error in data summary and transfer, revisiting the experimental conditions, performing new experiments, or modifying the kinetic model will be determined on a case-by-case basis. The important result of the present study is that our analysis identifies a specific direction to follow for improving dataset consistency and provides an estimate of the extent of possible improvement.

4.5. Computational Details. It is pertinent to mention the computational expense of our method. The MATLAB programming language was used to develop software that implements the techniques described in this work. Using a 1.7 GHz Pentium IV processor, 8.5 min of CPU time were needed to produce the information for Figure 3 and 45 s of CPU time were needed to obtain the upper and lower bounds on $C_{\mathscr{D}}$ and the accompanying sensitivities. The data in the former required a significantly longer time to produce because the solution of an optimization was required to determine u_{ef} for each pair of dataset units.

5. Summary

We have demonstrated that the technique of data collaboration that we developed recently[15,16] can be extended to determine if a collection of related experimental results are consistent with each other, within a specified chemical kinetics model. A key requirement for our analysis is the formulation of a dataset, which entails creation of dataset units from experimental observations and a common kinetic model. A dataset unit should consist of the measured observation, uncertainty bounds on the measurement, and a model that transforms active parameter values into a prediction for the measurement. Organized in this manner, the dataset can be subjected to rigorous numerical analysis, addressing questions of practical significance.

The technique of data collaboration rests on optimization constrained to a set of assertions. In this work, these assertions are the confinement of model parameters to the hypercube $\mathscr{H}$ and the requirement that the model predictions remain within the experimentally determined bounds. The new development presented in this paper is the use of constrained optimization to rigorously determine dataset consistency. Lagrange multiplier methods determine the sensitivity of the dataset consistency to the experimental results and prior information, providing researchers with insights into the quality of the data and the model.

The capability of the new numerical analysis was demonstrated on a real-world example, the GRI-Mech 3.0 dataset.[20]

As it actually happened, unintentionally and unexpectedly, the new procedure immediately identified two major outliers of the dataset, which were subsequently corrected upon re-examination of the raw experimental data. The results of the analysis suggest a sequential procedure with step-by-step identification of outliers and inspection of the causes. Altogether, the new numerical approach offers an important tool for assessing experimental observations and model building.

Acknowledgment. The material is based upon work supported by the National Science Foundation, under Grant No. CTS-0113985.

Appendix

The consistency measure $C_{\mathscr{D}}$ of a dataset is defined by constrained optimization. Below, we derive the relationship between the consistency measure and perturbations to the constraints. Most importantly, we provide results specific to quadratic models.

1. Determination of $C_{\mathscr{D}}$. Express the constraints from the prior information $\boldsymbol{x} \in \mathscr{H}$ as $f_j^{(\alpha)}(\boldsymbol{x}) \le 0$ and $f_j^{(\beta)}(\boldsymbol{x}) \le 0$, where

$$f_j^{(\alpha)}(\boldsymbol{x}) = -x_j + \alpha_j \qquad (\text{for } j = 1, 2, ..., n) \tag{A1a}$$

$$f_j^{(\beta)}(\boldsymbol{x}) = x_j - \beta_j \qquad (j = 1, 2, ..., n) \tag{A1b}$$

To express the constraints from the experimental data, define

$$f_e^{(l)}(\boldsymbol{x},\gamma) = \gamma + l_e - M_e(\boldsymbol{x}_e) + d_e \tag{A2a}$$

$$f_e^{(u)}(\boldsymbol{x},\gamma) = M_e(\boldsymbol{x}_e) - d_e - u_e + \gamma \tag{A2b}$$

for each $e \in \mathscr{E}$. Obviously, $f_e^{(l)}(\boldsymbol{x},\gamma) \le 0$ is the same as $\gamma + l_e \le M_e(\boldsymbol{x}_e) - d_e$, and the analogous relationship is true for $f_e^{(u)}(\boldsymbol{x},\gamma) \le 0$. With this notation,

$$C_{\mathscr{D}} = \max_{\boldsymbol{x},\gamma} \gamma, \text{ subject to}$$
$$\begin{cases} f_j^{(\alpha)}(\boldsymbol{x}) \le 0, & f_j^{(\beta)}(\boldsymbol{x}) \le 0 \quad (\text{for } j = 1, 2, ..., n) \\ f_e^{(l)}(\boldsymbol{x},\gamma) \le 0, & f_e^{(u)}(\boldsymbol{x},\gamma) \le 0 \quad (\text{for } e \in \mathscr{E}) \end{cases} \tag{A3}$$

We compute a lower bound, $\underline{C}_{\mathscr{D}}$, by attempting to solve eq A3 using nonlinear optimization software, with the resulting (local) maximum yielding a lower bound on $C_{\mathscr{D}}$. An upper bound, $\bar{C}_{\mathscr{D}}$, is generated by the Lagrange dual to eq A3:

$$\bar{C}_{\mathscr{D}} = \min_{-\nu_j^{(\alpha)},\nu_j^{(\beta)},-\nu_e^{(l)},\nu_e^{(u)} \ge 0} L(\boldsymbol{\nu})$$

where

$$L(\boldsymbol{\nu}) = \max_{\boldsymbol{x},\gamma} \{\gamma - \sum_{j=1}^{n} (-\nu_j^{(\alpha)} f_j^{(\alpha)}(\boldsymbol{x}) + \nu_j^{(\beta)} f_j^{(\beta)}(\boldsymbol{x})) - \sum_{e \in \mathscr{E}} (-\nu_e^{(l)} f_e^{(l)}(\boldsymbol{x},\gamma) + \nu_e^{(u)} f_e^{(u)}(\boldsymbol{x},\gamma))\}$$

and

$$\boldsymbol{\nu} = (\nu_1^{(\alpha)}, ..., \nu_n^{(\alpha)}, \nu_1^{(\beta)}, ..., \nu_n^{(\beta)}, \nu_{e_1}^{(l)}, ..., \nu_{e_m}^{(l)}, \nu_{e_1}^{(u)}, ..., \nu_{e_m}^{(u)}) \tag{A4}$$

The optimal $\boldsymbol{\nu}$ value gives the Lagrange multipliers discussed in the main text, $\boldsymbol{\nu} = \text{argmin } L(\boldsymbol{\nu})$. The objective function $L(\boldsymbol{\nu})$ in eq A4, which is called the Lagrangian, is a convex function of $\boldsymbol{\nu}$; however, for each $\boldsymbol{\nu}$, evaluating the Lagrangian requires the solution of a maximization, so this function may be difficult to evaluate. This presents an obstacle because the reason to compute an upper bound in the first place is to avoid solving a difficult problem. However, the maximization that determines $L(\boldsymbol{\nu})$ is unconstrained, so this function can be readily evaluated in special cases. An instance of this is treated in the next subsection, where each constraint function, e.g., $f_j^{(\alpha)}(\boldsymbol{x})$, is quadratic in its arguments.

2. Quadratic Models. For quadratic models, eq A4 is readily solved. In this case, for each $e \in \mathscr{E}$, there exist a symmetric $n \times n$ matrix A_e, an $n \times 1$ vector b_e, and a scalar c_e such that M_e is of the form $M_e(\boldsymbol{x}_e) = \boldsymbol{x}^{\mathrm{T}} A_e \boldsymbol{x} + 2b_e^{\mathrm{T}}\boldsymbol{x} + c_e$. The constraints $f_e^{(l)}(\boldsymbol{x},\gamma) \le 0$ are then

$$\begin{bmatrix} 1 \\ \gamma \\ \boldsymbol{x} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -c_e + d_e + l_e & 0.5 & -b_e \\ 0.5 & 0 & 0 \\ -b_e^{\mathrm{T}} & 0 & -A_e \end{bmatrix} \begin{bmatrix} 1 \\ \gamma \\ \boldsymbol{x} \end{bmatrix} \le 0 \tag{A5}$$

Similarly, $f_e^{(u)}(\boldsymbol{x},\gamma) \le 0$ becomes

$$\begin{bmatrix} 1 \\ \gamma \\ \boldsymbol{x} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} c_e - d_e - u_e & 0.5 & b_e \\ 0.5 & 0 & 0 \\ b_e^{\mathrm{T}} & 0 & A_e \end{bmatrix} \begin{bmatrix} 1 \\ \gamma \\ \boldsymbol{x} \end{bmatrix} \le 0 \tag{A6}$$

We use the following procedure to express the prior information constraints. Fix $\epsilon > 0$. Observe that, for $j = 1, 2, ..., n$, $\alpha_j \le x_j \le \beta_j$ is equivalent to the two constraints $\alpha_j \le x_j \le \beta_j + \epsilon$ and $\alpha_j - \epsilon \le x_j \le \beta_j$, which, in turn, are equivalent to

$$(\alpha_j - x_j)(\beta_j + \epsilon - x_j) \le 0 \qquad (j = 1, 2, ..., n) \tag{A7a}$$

$$(\alpha_j - \epsilon - x_j)(\beta_j - x_j) \le 0 \qquad (j = 1, 2, ..., n) \tag{A7b}$$

Let $Z_e^{(l)}$, $Z_e^{(u)}$ for $e \in \mathscr{E}$, and $Z_j^{(\alpha)}$, $Z_j^{(\beta)}$ (for $j = 1, ..., n$) denote the $(2 + n) \times (2 + n)$ symmetric matrixes associated with the quadratic functions in eqs A5, A6, and A7. Let Z_0 be the symmetric matrix that satisfies

$$\begin{bmatrix} 1 \\ \gamma \\ \boldsymbol{x} \end{bmatrix}^{\mathrm{T}} Z_0 \begin{bmatrix} 1 \\ \gamma \\ \boldsymbol{x} \end{bmatrix} = \gamma \tag{A8}$$

Using these matrixes, eq A4 may be expressed as

$$\bar{C}_{\mathscr{D}} = \min_{-\nu_j^{(\alpha)},\nu_j^{(\beta)},-\nu_e^{(l)},\nu_e^{(u)} \ge 0} L(\boldsymbol{\nu})$$

where

$$L(\boldsymbol{\nu}) = \max_{\boldsymbol{x},\gamma} \begin{bmatrix} 1 \\ \gamma \\ \boldsymbol{x} \end{bmatrix}^{\mathrm{T}} Z_0 - \sum_{j=1}^{n} (-\nu_j^{(\alpha)} Z_j^{(\alpha)} + \nu_j^{(\beta)} Z_j^{(\beta)}) - \sum_{e \in \mathscr{E}} (-\nu_e^{(l)} Z_e^{(l)} + \nu_e^{(u)} Z_e^{(u)}) \begin{bmatrix} 1 \\ \gamma \\ \boldsymbol{x} \end{bmatrix} \tag{A9}$$

Recall that, for a symmetric $(1 + n) \times (1 + n)$ matrix M, $\max_{x \in \mathrm{R}^n} [{}^{1}_{x}]^{\mathrm{T}} M [{}^{1}_{x}] \le 0$ *if and only if* M is negative semidefinite (denoted $M \le 0$). Hence, eq A9 is equivalent to

$$\bar{C}_{\mathscr{D}} = \min_{-\nu_j^{(\alpha)},\nu_j^{(\beta)},-\nu_{\mathrm{e}}^{(l)},\nu_e^{(u)} \ge 0;\rho} \rho \qquad \text{subject to } Z(\boldsymbol{\nu}) \le 0$$

where

$$Z(\boldsymbol{\nu}) = Z_0 - \begin{bmatrix}\rho & 0\\ 0 & 0\end{bmatrix} - \sum_{j=1}^{n}(-\nu_j^{(\alpha)} Z_j^{(\alpha)} + \nu_j^{(\beta)} Z_j^{(\beta)}) - \sum_{e\in\mathcal{E}}(-\nu_e^{(l)} Z_e^{(l)} + \nu_e^{(u)} Z_e^{(u)}) \tag{A10}$$

This type of problem is called a semidefinite program (SDP) and is readily solved with special-purpose convex programming algorithms.[25]

The value of ϵ used in $Z_j^{(\alpha)}$ and $Z_j^{(\beta)}$ should be small. Note that

$$Z_j^{(\alpha)} = \begin{bmatrix}\alpha_j\beta_j + \epsilon\alpha_j & 0 & -0.5(\alpha_j + \beta_j + \epsilon) & 0\\ 0 & 0 & 0 & 0\\ -0.5(\alpha_j + \beta_j + \epsilon) & 0 & 1 & 0\\ 0 & 0 & 0 & 0\end{bmatrix} \tag{A11a}$$

$$Z_j^{(\beta)} = \begin{bmatrix}\alpha_j\beta_j - \epsilon\beta_j & 0 & -0.5(\alpha_j + \beta_j - \epsilon) & 0\\ 0 & 0 & 0 & 0\\ -0.5(\alpha_j + \beta_j - \epsilon) & 0 & 1 & 0\\ 0 & 0 & 0 & 0\end{bmatrix} \tag{A11b}$$

where the dimension of the zero elements is determined by the index j. Large values of ϵ have a tendency to make $Z(\boldsymbol{\nu})$ more "positive" by contributing to the off-diagonal terms, so larger values of ρ are needed to ensure that $Z(\nu)$ remains negative semidefinite. This results in an unnecessarily large upper bound on $C_{\mathcal{D}}$; thus, the chosen value of ϵ should be small. However, for very small values of ϵ, $Z_j^{(\alpha)} \approx Z_j^{(\beta)}$; thus, the values of $\lambda_j^{(\alpha)}$ and $\lambda_j^{(\beta)}$ become unreliable indicators of the constraint sensitivity, because the **α**- and $\boldsymbol{\beta}$-constraints are represented by virtually identical matrixes. Our experience has been that a value of $\epsilon = 0.05(\beta_j - \alpha_j)$ provides accurate sensitivities without noticeably increasing $\bar{C}_{\mathcal{D}}$ from the value achieved for $\epsilon = 0$.

Quadratic models provide an additional benefit. Consider a case where γ and the parameter vectorx are modeled as an $(n + 1) \times 1$ random variable $\boldsymbol{V}$. We can recast eq A3 as an optimization over the random variable $\boldsymbol{V}$, where the constraints are only required to be satisfied in expected values. For example, the constraint $f_j^{(\alpha)}(x) \leq 0$ is replaced with the constraint that the expected value is $E[f_j^{(\alpha)}(\boldsymbol{V})] \leq 0$. This problem also may be formulated as an SDP,[25] and its solution provides the mean and covariance of the optimal $\boldsymbol{V}$. In the present context, this interpretation is useful, because we may sample a distribution with this mean and covariance to initialize the general constrained optimization that solves for $\underline{C}_{\mathcal{D}}$.

3. Derivation of Equation 3. Using the notation of eq A3, the consistency measure $C_{\tilde{\mathcal{D}}} = C_{\mathcal{D}} + \Delta C_{\mathcal{D}}$, obtained after perturbing the constraint bounds, is

$$C_{\tilde{\mathcal{D}}} = \max_{x,\gamma}\gamma, \text{ subject to}$$
$$\begin{cases} f_j^{(\alpha)}(\boldsymbol{x}) \leq -\Delta\alpha_j, \quad f_j^{(\beta)}(\boldsymbol{x}) \leq \Delta\beta_j, & (\text{for } j = 1, 2, ..., n)\\ f_e^{(l)}(\boldsymbol{x},\gamma) \leq -\Delta l_e, \quad f_e^{(u)}(\boldsymbol{x},\gamma) \leq \Delta u_e & (\text{for } e \in \mathcal{E})\end{cases} \tag{A12}$$

For all Δt, we have the following compact version of eq 3 from the main text:

$$\Delta C_{\mathcal{D}} \leq \bar{C}_{\mathcal{D}} - C_{\mathcal{D}} + \boldsymbol{\lambda}^{\mathrm{T}}\Delta t \tag{A13}$$

where the Lagrange multipliers ($\boldsymbol{\lambda} = \text{argmin } L(\boldsymbol{\nu})$) and the upper bound on the nominal consistency measure ($\bar{C}_{\mathcal{D}}$) are obtained from eq A4. For proof of eq A13, take any $(\boldsymbol{x},\gamma)$ that satisfies the constraints in eq A12. From the definition of $\bar{C}_{\mathcal{D}}$ in eq A4,

$$\bar{C}_{\mathcal{D}} = L(\boldsymbol{\lambda}) \geq \gamma - \sum_{j=1}^{n}(-\lambda_j^{(\alpha)} f_j^{(\alpha)}(\boldsymbol{x}) + \lambda_j^{(\beta)} f_j^{(\beta)}(\boldsymbol{x})) - \sum_{e\in\mathcal{E}}(-\lambda_e^{(l)} f_e^{(l)}(\boldsymbol{x},\gamma) + \lambda_e^{(u)} f_e^{(u)}(\boldsymbol{x},\gamma))$$

Because $(\boldsymbol{x},\gamma)$ satisfies the constraints in eq A12, $f_j^{(\alpha)}(\boldsymbol{x}) \leq -\Delta\alpha_j$, $f_j^{(\beta)}(\boldsymbol{x}) \leq \Delta\beta_j$, $f_e^{(l)}(\boldsymbol{x},\gamma) \leq -\Delta l_e$, and $f_e^{(u)}(\boldsymbol{x},\gamma) \leq \Delta u_e$. Also, $-\lambda_j^{(\alpha)}$, $\lambda_j^{(\beta)}$, $-\lambda_e^{(l)}$, and $\lambda_e^{(u)} \geq 0$. Consequently, $\bar{C}_{\mathcal{D}} \geq \gamma - \boldsymbol{\lambda}^{\mathrm{T}}\Delta\boldsymbol{t}$. Maximizing both sides of this last equation over all $(\boldsymbol{x},\gamma)$ that satisfy the constraints in eq A12, so γ becomes $C_{\tilde{\mathcal{D}}}$ ($C_{\tilde{\mathcal{D}}} = C_{\mathcal{D}} + \Delta C_{\mathcal{D}}$), yields

$$\bar{C}_{\mathcal{D}} \geq C_{\mathcal{D}} + \Delta C_{\mathcal{D}} - \boldsymbol{\lambda}^{\mathrm{T}}\Delta\boldsymbol{t}$$

Algebraic manipulation then yields eq A13.

4. Derivation of Equation 8. Equation 8 indicates

$$C_{\mathcal{D}} \leq C_{\tilde{\mathcal{D}}} \leq \bar{C}_{\mathcal{D}} \tag{A14}$$

where $\tilde{\mathcal{D}}$ is generated by making the following changes to $\mathcal{D}$:

$$\tilde{\alpha}_j = \begin{cases}-\infty & (\text{if } \lambda_j^{(\alpha)} = 0)\\ \alpha_j & (\text{otherwise})\end{cases}$$
$$\tilde{\beta}_j = \begin{cases}\infty & (\text{if } \lambda_j^{(\beta)} = 0)\\ \beta_j & (\text{otherwise})\end{cases} \quad (\text{for } j = 1, 2, ..., n)$$
$$\tilde{l}_e = \begin{cases}-\infty & (\text{if } \lambda_e^{(l)} = 0)\\ l_e & (\text{otherwise})\end{cases}$$
$$\tilde{u}_e = \begin{cases}\infty & (\text{if } \lambda_e^{(u)} = 0)\\ l_e & (\text{otherwise})\end{cases} \quad (\text{for } e \text{ in } \mathcal{E}) \tag{A15}$$

The left inequality in eq A14 is clear, because, to determine $C_{\tilde{\mathcal{D}}}$, we maximize γ over a larger set than that which is used in determining $C_{\mathcal{D}}$. It remains to be seen that the right inequality holds as well.

Define a sequence of datasets by making the following modifications to $\mathcal{D}$. For $k = 1, 2, 3, ...$, generate $\tilde{\mathcal{D}}_k$ by

$$\tilde{\alpha}_{jk} = \begin{cases}\alpha_j - k & (\text{if } \lambda_j^{(\alpha)} = 0)\\ \alpha_j & (\text{otherwise})\end{cases}$$
$$\tilde{\beta}_{jk} = \begin{cases}\beta_j + k & (\text{if } \lambda_j^{(\beta)} = 0)\\ \beta_j & (\text{otherwise})\end{cases} \quad (\text{for } j = 1, 2, ..., n)$$
$$\tilde{l}_{ek} = \begin{cases}l_e - k & (\text{if } \lambda_e^{(l)} = 0)\\ l_e & (\text{otherwise})\end{cases}$$
$$\tilde{u}_{ek} = \begin{cases}u_e + k & (\text{if } \lambda_e^{(u)} = 0)\\ u_e & (\text{otherwise})\end{cases} \quad (\text{for } e \text{ in } \mathcal{E}) \tag{A16}$$

Observe that, in the limit, $\tilde{\mathcal{D}}_k$ approaches $\mathcal{D}$. Because the constraints in eq A12 become progressively weaker as k increases, $C_{\tilde{\mathcal{D}}_k} \leq C_{\tilde{\mathcal{D}}_{k+1}}$ for all natural numbers k. Invoking eq A13 and noting that the $\boldsymbol{\lambda}^{\mathrm{T}}\Delta\boldsymbol{t}$ term vanishes, we have $C_{\tilde{\mathcal{D}}_k} \leq \bar{C}_{\mathcal{D}}$ for all k. Consequently, the increasing sequence $(C_{\tilde{\mathcal{D}}_k})_{k=1}^{\infty}$

converges, for example, to $r \leq \bar{C}_{\mathscr{D}}$. In other words, $\lim_{k\to\infty} C_{\tilde{\mathscr{D}}_k} = r \leq \bar{C}_{\mathscr{D}}$.

To conclude the proof, we will show $r = C_{\tilde{\mathscr{D}}}$. By construction of $\tilde{\mathscr{D}}_k$ and the definition of the consistency measure, $C_{\tilde{\mathscr{D}}_k} \leq C_{\tilde{\mathscr{D}}}$ for all k; thus,

$$r \leq C_{\tilde{\mathscr{D}}} \tag{A17}$$

For the reverse inequality, fix $\epsilon > 0$ and take $\boldsymbol{x}' \in R^n$, $\gamma' \in R$ for which $C_{\tilde{\mathscr{D}}} - \gamma' < \epsilon$, and

$$\begin{gathered} -x'_j \leq -\tilde{\alpha}_j,\ x'_j \leq \tilde{\beta}_j\ (j = 1, 2, ..., n) \\ -M_e(\boldsymbol{x}'_e) + d_e \leq -\tilde{l}_e - \gamma',\ M_e(\boldsymbol{x}'_e) - d_e \leq \tilde{u}_e - \gamma'\ (e \text{ in } \mathscr{E}) \end{gathered} \tag{A18}$$

(We tacitly assumed that $C_{\tilde{\mathscr{D}}}$ was finite in choosing γ'; if $C_{\tilde{\mathscr{D}}} = \infty$, one can show $C_{\bar{\mathscr{D}}}$ must also be infinite, so eq A14 holds.) The left-hand side of each inequality in eq A18 is finite, so we may choose a natural number K such that

$$-x'_j \leq -\tilde{\alpha}_{jK} \qquad (j = 1, 2, ..., n) \tag{A19a}$$

$$x'_j \leq \tilde{\beta}_{jK} \qquad (j = 1, 2, ..., n) \tag{A19b}$$

$$-M_e(\boldsymbol{x}'_e) + d_e \leq -\tilde{l}_{eK} - \gamma' \tag{A19c}$$

$$M_e(\boldsymbol{x}'_e) - d_e \leq \tilde{u}_{eK} - \gamma' \text{ (for } e \text{ in } \mathscr{E}) \tag{A19d}$$

This gives $\gamma' \leq C_{\tilde{\mathscr{D}}_K} \leq r$, where the left-hand inequality follows because $(\boldsymbol{x}',\gamma')$ satisfies the constraints implied by $\tilde{\mathscr{D}}_K$, and the right-hand inequality holds because r is the limit of the increasing sequence of $C_{\tilde{\mathscr{D}}_k}$ values. We chose a γ' value for which $C_{\tilde{\mathscr{D}}} - \epsilon < \gamma'$, so we have $C_{\tilde{\mathscr{D}}} - \epsilon < r$. The positive constant ϵ was arbitrary, so $C_{\tilde{\mathscr{D}}} \leq r$. Combining the latter with eq A17, we have $C_{\tilde{\mathscr{D}}} = r$, so $C_{\tilde{\mathscr{D}}} \leq \bar{C}_{\mathscr{D}}$ as claimed. This concludes the proof of eq A14.

References and Notes

(1) Rosenbrock, H. H.; Storey, C. *Computational Techniques for Chemical Engineers*; Pergamon: New York, 1966.

(2) Polak, L. S. *Application of Numerical Methods to Chemical and Physical Kinetics* (in Russ.); Nauka: Moscow, Russia, 1969.

(3) Box, G. E. P.; Draper, N. R. *Empirical Model-Building and Response Surfaces*; Wiley: New York, 1987.

(4) Frenklach, M. In *Combustion Chemistry*; Gardiner, W. C., Jr., Ed.; Springer−Verlag: New York, 1984; p 423.

(5) Milstein, J. In *Modelling of Chemical Reaction Systems*; Ebert, K. H., Deuflhard, P., Jäger, W., Eds.; Springer−Verlag: Berlin, 1981; p 92.

(6) Bock, H. G. In *Modelling of Chemical Reaction Systems*; Ebert, K. H., Deuflhard, P., Jäger, W., Eds.; Springer−Verlag: Berlin, 1981, p 102.

(7) Scire, J. J., Jr.; Dryer, F. L.; Yetter, R. A. *Int. J. Chem. Kinet.* **2001**, *33*, 784−802.

(8) Miller, D.; Frenklach, M. *Int. J. Chem. Kinet.* **1983**, *15*, 677−696.

(9) Frenklach, M.; Wang, H.; Rabinowitz, M. J. *Prog. Energy Combust. Sci.* **1992**, *18*, 47−73.

(10) Qin, Z.; Lissianski, V.; Yang, H.; Gardiner, W. C., Jr.; Davis, S. G.; Wang, H. *Proc. Combust. Inst.* **2001**, *28*, 1663−1669.

(11) Harris, S. D.; Elliot, L.; Ingham, D. B.; Pourkashanian, M.; Wilson, C. W. *Comput. Methods Appl. Mech. Eng.* **2000**, *190*, 1065−1090.

(12) Elliot, L.; Ingham, D. B.; Kyne, A. G.; Mera, N. S.; Pourkashanian, M.; Wilson, C. W. *Combust. Sci. Eng.* **2003**, *175*, 619−648.

(13) Phenix, B. D.; Dinaro, J. L.; Tatang, M. A.; Tester, J. W.; Howard, J. B.; McRae, G. J. *Combust. Flame* **1998**, *112*, 132−146.

(14) Najm, H.; Reagan, M.; Debusschere, B.; Knio, O.; Ghanem, R.; Le Maître, O. Presented at the 19th International Colloqium on the Dynamics of Explosions and Reactive Systems; Hakone, Japan, July 27−August 1, 2003.

(15) Frenklach, M.; Packard, A.; Seiler, P. In *Proceedings of the American Control Conference*; IEEE: New York, 2002; pp 4135−4140.

(16) Frenklach, M.; Packard, A.; Seiler, P.; Feeley, R. *Int. J. Chem. Kinet.* **2004**, *36*, 57−66.

(17) Seiler, P.; Frenklach, M.; Packard, A.; Feeley, R. Numerical Approaches for Developing Predictive Models. Submitted to *Eng. Optim.*

(18) Gardiner, W. C., Jr. *Rates and Mechanisms of Chemical Reactions*; W. A. Benjamin: Menlo Park, CA, 1972.

(19) Voet, D.; Voet, J. G.; Pratt, C. W. *Fundamentals of Biochemistry;* Wiley: New York, 1999.

(20) Smith, G. P.; Golden, D. M.; Frenklach, M.; Moriarty, N. W.; Eiteneer, B.; Goldenberg, M.; Bowman, C. T.; Hanson, R. K.; Song, S.; Gardiner, W. C., Jr.; Lissianski, V. V.; Qin, Z. GRI-Mech 3.0; http://www.me.berkeley.edu/gri_mech/.

(21) Box, G. E. P.; Meyer, R. D. *J. Res. Natl. Bur. Stand.* **1987**, *90*, 494−500.

(22) Myers, R. H.; Montgomery, D. C. *Response Surface Methodology*; Wiley Series in Probability and Statistics; Wiley: New York, 2002.

(23) Chang, A.; Davidson, D.; DiRosa, M.; Hanson, R. K.; Bowman, C. T. Shock Tube Experiments for Development and Validation of Kinetic Models of Hydrocarbon Oxidation. In *25th Symposium (International) on Combustion*; 1994; Poster 3-23.

(24) Bowman, C. T. Personal communication, 2002.

(25) Vandenberghe, L.; Boyd, S. *SIAM Rev.* **1996**, *38*, 49−95.