Collaborative Data Processing in Developing Predictive Models of Complex Reaction Systems

MICHAEL FRENKLACH, ANDREW PACKARD, PETE SEILER, RYAN FEELEY

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

ABSTRACT: The subject of this report is a methodology for the transformation of (experimental) data into predictive models. We use a concrete example, drawn from the field of combustion chemistry, and examine the data in terms of precisely defined modes of scientific collaboration. The numerical methodology that we employ is founded on a combination of response surface technique and robust control theory. The numerical results demonstrate that an essential element of scientific collaboration is collaborative processing of data, demonstrating that combining the entire collection of data into a joint analysis extracts substantially more of the information content of the data.

Contract Grant Sponsor: National Science Foundation, Information Technology Research Program, Grant No. CTS-0113985.

INTRODUCTION

Development of predictive models for complex natural phenomena and industrial processes is at the core of scientific activity. With the present problems facing our society—threat of terrorist attacks, global warming, earthquake preparedness, safety of transport of nuclear waste, pollutant emission from automobile engines, etc.—one has to have a certain degree of confidence to rely on model predictions for political decisions, economic forecast, or design and manufacturing of automotive engines. Models of such complexity call for integration of large amounts of information, collected by numerous researchers and often from different disciplines. While collaboration among scientists is widely accepted as truism, it usually takes the form of a simple exchange of data and merging of computer codes. The subject of the present report is a quantitative demonstration that collaborative processing of the entire data base leads to systematic development of predictive models.

To make the presentation of the ideas clearer and results more concrete, we focus on a "real-world" example: chemical kinetics of pollutant formation in combustion of natural gas. The level of complexity, the knowledge base available, the degree of uncertainty, and the societal importance of the subject all make this system a suitable subject for the analysis.

The purpose of a natural gas chemical kinetics model is to predict concentration of combustion products, such as NO and CH_2O, which are major atmospheric pollutants yet minor chemical species in the context of the combustion chemistry. An accurate description of these products necessitates inclusion of many (hundreds) reaction steps. Mathematically, the evolution of chemical concentrations is governed by coupled first-order ordinary differential equations (ODEs), a nonlinear system that does not possess a closed-form solution. The model's authenticity relies on the knowledge of the reaction network, which sets the structure of the ODEs, and the model parameters, in this case reaction rate coefficients.

How does such a complex model come into being? In the case of natural gas combustion, half-a-century of research across five continents has established consensus for the reaction

steps of the system to a degree that the mathematical structure of the model can be assumed known. Reaching consensus on the "correct" parameter values has been more problematic. Scrupulous efforts on the part of chemical kineticists have led to isolation of small groups of reactions at well defined experimental conditions. The collected experimental observations are interpreted in isolation, focusing typically on one or two chemical reaction steps and making assumptions for the rest of the chemistry. These *derived* data are reported in the form of *conclusions* on the presence or absence of reaction pathways and the values of the corresponding rate parameters. The working paradigm of the community has been that combining information collected from such individual studies should create a "comprehensive" model of the process. This strategy, indeed, advanced the understanding of reaction networks, their interworking, and had general implications to practical combustion. Yet, predictive capabilities of such models did not materialize—assigning literature recommendations to model parameters seldom, if at all, results in adequate quality of model prediction on the whole [1]. This failure was attributed to compound uncertainty spanning multi-dimensional space of model parameters, the presence of intrinsic correlations among them, and restriction of the "in-isolation" data analysis to only one or two model parameters keeping the rest at literature recommendations [1]. The effectiveness of the compartmentalized data processing in developing predictive models thus is brought into question.

An alternative approach, fostered by the GRI-Mech project [2], is to carry out the analysis using the entire knowledge base available on the system, as a collaboration among all the data. The latest GRI-Mech release is comprised of 325 reversible reactions among 53 chemical species [2]. It was trained on 77 experimental targets, well-documented and expert-evaluated experimental observations. The training was performed using the Solution Mapping (SM) method [1, 3], which included the following steps. For each of the targets, a sensitivity analysis identified a set of *active* model parameters, i.e., those having measurable effects on the modeled outcome. After that, 77 statistical surrogate models were developed with a response-surface technique, through a set of computer simulations arranged in a factorial design [4, 5].

Each surrogate model approximated the numerical solution of the ODE model for a given target by a second-order polynomial in the respective active model parameters over their ranges of uncertainties. A joint optimization was then performed, adjusting a small subset of the overall 102 active model parameters to best fit the 77 surrogate models to their respective experimental targets.

Recently, we found that the statistical surrogates can be useful not just in model optimization, but also in prediction of model uncertainty. The latter was demonstrated through application of constrained polynomial optimization from Robust Control theory [6]. In so doing the paradigm of a model is shifted from considering model parameters as "unique", predetermined values with individual, uncorrelated uncertainties to including the actual experimental data, along with the physicochemical theoretical constraints if available, as the integral part of the model, with model parameters playing a role of internal variables. In other words, instead of the two-stage approach—i.e, estimation of model parameters from fitting experimental data followed by model predictions using the obtained estimates—we transfer the uncertainties of the "raw" data into model prediction *directly*.

The present report pursues these ideas further, showing wider applications and more general implications for collaborative data analysis. We focus on the benefits of what we call *collaboration of data*. It does not change the way experimentation is done, but requires a different approach to analyzing even one's own observations and, as a consequence, places new standards on data reporting. In this approach, "reporting an experiment" consists of documenting three items: measured outcome, its estimated uncertainty, and a model of the experimental system. Taken together, these three pieces of "data" assert, albeit implicitly, what was learned from one specific experimental effort. A centralized data repository, actual or virtual, collects these assertions from scientists all over the world. Anyone, from a contributing scientist to practicing engineer to policy-maker can query the *community data*. Questions can be posed such as "Is a given set of assertions self-consistent?," "What is the best least-squares fit to reconcile the assertions," and "For a given model of global warming, what range of annual temperature

changes are consistent with the assertions?". Once such questions are translated into a suitable mathematical formalism, automated algorithms can use all available assertions to infer an answer to the posed question. The purely mathematical task of extracting desired information from all reported experiments is relegated to algorithms. We compare this proposed methodology to that of common practice, where reported data are of a more derived nature, often with additional assumptions, and where analysis combines one's own experimental results with the derived reported data of others producing and reporting new derived data. We demonstrate the methodology with the GRI-Mech dataset—an existing data repository of the type required for collaborative data analysis. We employ rigorous mathematical tools and compare in quantitative terms different modes of scientific data collaboration.

PROBLEM FORMULATION

We base our analysis on the 77 surrogate models together with the training data (see Appendix A). The active model parameters form a 102-dimensional real vector, $x = [x_1, \ldots, x_{102}]$. A priori knowledge, with centering and normalization, confines their values to the unit hypercube, $\mathcal{H}$, in 102-dimensional space, $-1 \leq x_k \leq +1$. The hypercube represents the current "literature recommendation": the center of $\mathcal{H}$ is treated as the set of "recommended values" and each of its [-1 +1] sides as the "community evaluated uncertainty" for the corresponding active parameter. Associated with an experiment E is a dataset unit: a measured outcome, experimental uncertainty, and model, written (D_E, U_E, M_E). In the present case, M_E is the surrogate model, a polynomial which models the effect of active parameters on the outcome of the experiment E. Parameters consistent with E and a priori information are those which satisfy

$$|M_E(x) - D_E| \leq U_E \qquad \text{and} \qquad x \in \mathcal{H}.$$

We will refer to them as the *feasible* set, F_E. The above forms an *assertion*, a set of constraints that the active parameters must satisfy. There are 77 different experimental dataset units under consideration, drawn from different physical arrangements and conditions. We identify these

dataset units by an integer subscript, i.e., (D_i, U_i, M_i). The feasible set implied by all dataset units is the intersection $\mathcal{F} = \bigcap_{i=1}^{77} F_i$.

Given the GRI-Mech dataset—$(D_i, U_i, M_i),\ i = 1, \ldots, 77$— our goal is to make a *prediction* for a new target, E_0, still within the domain of the accepted physical model. For instance, having a series of species concentration measurements under various laboratory apparatuses (the training data), one's objective is to make a prediction for the extent of global warming. In mathematical terms, the target E_0 has a corresponding model M_0. Any $x \in \mathcal{F}$ yields a possible outcome of E_0, namely $M_0(x)$. The analysis objective is to determine the minimum and maximum values of M_0 as x takes on values over the feasible set $\mathcal{F}$. These limits are an explicit manifestation of the assertions, i.e., the training data constraints and uncertainties. The extent to which all information is effectively and consistently used affects the quality and correctness of the prediction. It is the difficulty (logistical and mathematical) in solving this general problem that differentiates among approaches and, as a consequence, dictates different modes of collaboration.

With this in mind, we consider the following modes:

Mode A, Frozen Core—the extreme of non-collaboration: an individual experiment is analyzed in isolation, with the range of a single (usually most influential) parameter determined by fitting the measurement while having fixed the remaining parameters at the literature recommended values. This is currently the most typical mode of data processing.

Mode B: Free Core—possibly the best form of non-collaboration: again, analysis is carried out in isolation, focusing on the most influential parameter. In contrast to Mode A, the remaining parameters are not "frozen" at specified values but assumed to lie in the unit hypercube. This mode, as will be shown below, is not of much practical interest. We devised it solely for the purpose of bridging between Modes A and C, which are not directly comparable.

By themselves, Modes A and B are not explicitly aimed at prediction. They focus on parameter identification, with the presumption that once all model parameters are known precisely, any prediction using the identified parameters will be accurate.

Mode C: Full Data Collaboration—the extreme of collaboration: The entire knowledge base for the system is used for prediction.

We considered all three modes. For simplicity, and in light of insufficient records of experimental uncertainties even for such a well-documented case as GRI-Mech, an artificial but realistic uniform level of experimental uncertainties, $U_i = 0.1$, was assumed in all cases tested.

ANALYSIS

We begin with the Frozen Core scenario, Mode A. For experiment i, we select the parameter x_k, which has the largest impact on M_i (see Appendix A), and freeze the rest at $x_j = 0$ for $j = 1, \ldots, 102$, $j \neq k$. In 38 of the 77 cases, it is not possible to find a solution that fits the experimental observation, i.e., there does not exist $x_k \in [-1, 1]$ such that $M_i([0 \ldots 0\, x_k\, 0 \ldots 0]) = D_i$. If we allow for uncertainty in the experimental value, then for 5 of the 77 experiments there does not exist $x_k \in [-1, 1]$ such that $|M_i([0 \ldots 0\, x_k\, 0 \ldots 0]) - D_i| \leq U_i$. With this result, the experimenter i is forced to conclude (in this case incorrectly) that an error exists in either the model or the data.

Mode A analysis, which is typical in many fields, also can lead to "controversies" among researchers. For instance, based on E_{66} (GRI-Mech target SCH.C12), x_{44} (rate constant of reaction 126, see Appendix A) is reported to lie in the interval [0.38, 1.0]. However, based on E_{67} (GRI-Mech target SCH.C13), x_{44} is reported to lie in the interval [$-$1.0, 0.22]. x_{44} is the highest ranking impact parameter for both SCH.C12 and SCH.C13 targets, and it would appear (again incorrectly) that these experimental targets are in conflict with each other. We find similar inconsistencies affecting 9 of the 102 parameters. Creating collaborations among several

researchers by combining their measurements into a joint analysis yet still within the framework of the Frozen Core mode does not remove the principal difficulties with this approach.

As it happens, the feasible set $\mathcal{F}$ of the GRI-Mech dataset is not empty. The apparent dilemma arises because, in the interest of "simplifying" the analysis, the Mode A data processing freezes many active parameters. The general problem with parameter freezing is that only a subset of feasible points are actually considered. As a simple visualization, consider a typical banana-shaped confidence region [3] as the feasible set, $\mathcal{F}$. A one-dimensional cross-section, e.g. along a single dimension x_k, comprises only a subset of all feasible points. A collection of low-dimensional cross-sections may have no points in common, yet, in reality, all belong to the same $\mathcal{F}$. The controversies revealed in the Mode A analysis have similar origins, namely unnecessary (and actually unjustified) overconstraining. Mode A is further complicated by the fact that each cross section is derived from a different superset, $\mathcal{F}_i$, of $\mathcal{F}$.

To visualize further the multi-dimensional geometry of the feasible set, let us continue the example discussed above, namely that determination of x_{44} from E_{66} and E_{67} in separate procedures (each according Mode A) gives mutually inconsistent results: the respective ranges [0.38, 1.0] and [$-$1.0, 0.22] do not overlap. We now expand the analysis of each of the experiments to two dimensions in optimization variables, by including x_{45}, the second highest impact parameter for E_{67} and the fourth highest for E_{66} (A_{127} of reaction CH + $H_2O \rightarrow$ H + CH_2O). The feasible sets obtained, still under Mode A, are displayed in Fig. 1.

The thick horizontal lines at $x_{45} = 0$ in the two panels of Fig. 1, labelled A, are the non-overlapping ranges [0.38, 1.0] and [$-$1.0, 0.22]: the feasible set for x_{44} determined solely from E_{66} (top panel) and the feasible set for x_{44} from E_{67} (bottom panel), respectively. In other words, with $x_1 = 0, x_2 = 0, \ldots, x_{43} = 0, x_{45} = 0, \ldots, x_{102} = 0$ (i.e., with all rate coefficients but k_{126} frozen at their respective literature values), for any value of x_{44} from $-$1.0 to 0.22 the GRI-Mech 3.0 model prediction for $[CH]_{max}$ lies within the respective error bounds (i.e., within interval $D_{66} \pm U_{66}$).

The total shaded area in each panel of Fig. 1 represents a feasible set determined, again

solely from an individual experiment, but now for x_{44} and x_{45} simultaneously (while keeping the remaining 100 x's frozen at zero, and hence within Mode A). Thus, for every pair of the x_{44} and x_{45} values located within the shaded area of the top panel, GRI-Mech 3.0 predicts D_{66} within $D_{66} \pm U_{66}$. A similar interpretation applies to the bottom panel. Comparing the one-dimensional analysis with that extended to two-dimensions, we observe that whereas the one-dimensional feasible sets (thick horizontal lines at $x_{45} = 0$) do not overlap, the two-dimensional feasible sets do have a common set of points, within the region shaded in darker grey (labelled C). In other words, while determination of a single parameter from a single experiment led to an apparent "controversy" (as if the two results disagree with each), adding just one more parameter into consideration resolves it by finding a "mutually agreeable" set of acceptable values. Including more optimization variables expands the feasible set.

To illustrate the difference among the Modes A, B, and C of analysis, imagine now that the entire dataset is comprised of just one measurement, E_{66}, and the model $M_{66}(x)$ has only two active parameters, x_{44} and x_{45}. The results for this case are depicted in the top panel of Fig. 1. Determination of x_{44} by matching $M_{66}(x_{44}, x_{45})$ to D_{66} while keeping $x_{45} = 0$ constitutes, as before, the Mode A analysis. The thick solid line at $x_{45} = 0$ (labelled A) represents the corresponding feasible set: all values of x_{44} that assure the model prediction $M_{66}(x_{44}, x_{45} = 0)$ to lie within interval $D_{66} \pm U_{66}$. Allowing both x_{44} and x_{45} to vary while matching $M_{66}(x_{44}, x_{45})$ to D_{66} constitutes the Mode C, and the entire shaded area represents the corresponding feasible set: all points (x_{44}, x_{45}) for which $M_{66}(x_{44}, x_{45})$ lies within interval $D_{66} \pm U_{66}$. By comparison, Mode B has as its goal determination of only x_{44}, by matching $M_{66}(x_{44}, x_{45})$ to D_{66}, while allowing x_{45} to be within its uncertainty range, [-1, +1]. This results in the interval labelled B, obtained by the projection of the entire shaded area onto the x_{44} axis. The outcome of Mode B, reported as x_{44} belonging to interval B and x_{45} to interval [-1, +1], implies a rectangular area circumscribing the entire shaded area, thus overpredicting the true feasible region. In another example of a one-experiment dataset, with E_{67}, the Mode B analysis results in no reduction of the initial uncertainty at all (bottom panel of Fig. 1).

Imagine further that the above examples signify data analysis performed by two researchers, each individually, one on experiment E_{66} and the other on experiment E_{67}. Both researcher would report the same intervals for x_{45} but differing intervals for x_{44}. By comparison, the two researchers could combine their measurements forming a single dataset, comprised of two experiments E_{66} and E_{67} and the model with two active parameters $M_{66}(x_{44}, x_{45})$. The Mode C feasible set in this case is designated as the darker shaded area in both panels of Fig. 1, which contains all points (x_{44}, x_{45}) for which $M_{66}(x_{44}, x_{45})$ lies within interval $D_{66} \pm U_{66}$ and $M_{67}(x_{44}, x_{45})$ within interval $D_{67} \pm U_{67}$.

Our next dataset example, which is still accessible to graphical visualization, is obtained by expanding the combined dataset to include an additional optimization variable, x_{34}, representing A_{97} of reaction OH + CH_3 $\rightarrow$ CH_2(S) + H_2O, the second highest impact parameter for E_{66} and the fourth highest for E_{67}). The feasible set obtained in the Mode C analysis is shown in Fig. 2. A cross-section of this feasible set by the plane $x_{34} = 0$ forms the feasible set C in Fig. 1. With the addition of experiments and active variables, the geometry of the feasible set grows in complexity, and, while difficult to visualize, the basic features should be similar to those exhibited by the above two- and three-dimensional examples.

By their nature, Modes B and C eliminate artificial controversies, as no feasible points are ruled out. In Mode B, the intersection of all derived parameter intervals results in a hyperparallelogram, $\hat{\mathcal{H}} \subseteq \mathcal{H}$, and the analysis insures that $\hat{\mathcal{H}}$ contains the feasible set, $\mathcal{F} \subseteq \hat{\mathcal{H}}$. This allows us to express a range of prediction, $R_{\text{posterior}}$, as

$$R_{\text{posterior}} := [\min_{x \in \hat{\mathcal{H}}} M_0(x) \quad \max_{x \in \hat{\mathcal{H}}} M_0(x)].$$

We use the subscript *posterior* to distinguish this estimate from the one obtained without constraining x to the feasible region, as will be discussed later. Mode C, Full Data Collaboration, is naturally expressed in the same mathematical form, yet now using the information implied by the reported experiments *all at once*. This replaces the constraint $x \in \hat{\mathcal{H}}$ with $x \in \mathcal{F}$,

$$R_{\text{posterior}} := [\min_{x \in \mathcal{F}} M_0(x) \quad \max_{x \in \mathcal{F}} M_0(x)].$$

Since $\mathcal{F}$ is a subset of $\hat{\mathcal{H}}$, this guarantees tighter predictions than Mode B.

To determine the posterior predictions, the constrained minima and maxima are computed by bounding each quantity from above and below. Evaluating $M_0(x)$ at any feasible x yields a lower bound on the maximum and an upper bound on the minimum, called *inner bounds*, and they can be improved with local search. Upper bounds on the maximum and lower bounds on the minimum yield *outer bounds* of the predicted range. Outer bounds are hard limits on the value of the objective function over the feasible set. Obtaining them, in principle, requires considering every point of the feasible set. In the context of computational complexity theory [7], for general functions and constraints, this task is difficult. Even deciding that the minimum of an indefinite quadratic function subject to box constraints (each x_i lies in $[0\ 1]$) is greater than a given fixed number is known to be a "hard" (NP-complete) problem [8, 9]. In more descriptive terms, an NP-complete problem has two characteristics: no algorithm is known to solve all problems of this type efficiently [and algorithm is efficient if the computation time scales as O(n^p), where n is the problem's "size"]; and it is unlikely that such an algorithm exists, as if it were to exist for one problem of this type, it would be immediately transformable to all other such problems (the NP-complete class of problems).

NP-completeness, however, does not rule out the existence of good algorithms which work well on some specific instances of a problem. Our present study is an example of such a case. Analysis methods in Robust Control (RC) offer practical outer bound computational schemes for the problem we consider—minimization of quadratic functions subject to quadratic inequality constraints. These only yield outer bounds, and on any given problem the computed bound may be far below the true minimum. When coupled with the inner bound, a concrete (if not tight) statement about the value of the constrained optimum may be made.

In the present formulation, it is the surrogate response models that play the role of constraints, and the fact that these models are quadratic functions lends naturally to the employment of the RC techniques. In all calculations that follow, we computed both inner and outer bounds. In each case, the inner and outer bounds were within a few percent of each other, indicating

close approximation to the actual range. Results quoted refer to outer bounds.

The general ideas of the RC algorithms that we use are briefly outlined in Appendix B. We remark that while the proposed methodology yields a new approach to analysis of dynamic models, neither the RC algorithms nor the development of surrogate models through Solution Mapping are new or unique. Rather, our approach casts the prediction problem as a constrained optimization, drawn on the entire knowledge base available on the system of interest. Our combined RC-SM technique is just one possible method of approaching the optimization problem. The principal conclusion of the present study should not change if other numerical methods are employed for this purpose. At the present time, however, we are unaware of methods comparable in computational efficiency (see below) to our approach, or that the problem we pose—assessment of Modes B and C—has been addressed in a numerically rigorous manner.

To make the analysis of Modes B and C more general, we chose an arbitrary polynomial form for M_0, with qualitative features similar to a typical experiment of the GRI-Mech dataset: a quadratic function that depends on the 20 active parameters which occur most frequently in the GRI-Mech dataset. The norms of the linear and quadratic terms of these functions were scaled to be equal to the average values of the linear and quadratic terms of the GRI-Mech surrogate models. For each case, we generated 100 random targets.

To compare effectiveness of Modes B and C, we introduce a measure of information gained as a result of the data processing in these modes,

$$I := 1 - \frac{\Delta R_{\text{posterior}}}{\Delta R_{\text{prior}}}.$$

where ΔR is the length of an interval R. This definition characterizes the information gain I as a relative decrease in the predicted range of model M_0. Prior to data processing the variations of x's are confined to the initial hypercube $\mathcal{H}$, and R_{prior} is the range of model prediction on $\mathcal{H}$,

$$R_{\text{prior}} := [\min_{x \in \mathcal{H}} M_0(x) \quad \max_{x \in \mathcal{H}} M_0(x)]$$

If processing the data results in no reduction in the predicted range, i.e., $\Delta R_{\text{posterior}} = \Delta R_{\text{prior}}$,

we have gained no information and $I = 0$. If $\Delta R_{\rm posterior} = 0$, we have extracted the maximum information possible for target E_0 and $I = 1$.

The frequency of I is displayed in the left panels of Fig. 3. The Mode B (top) resulted in a rather small information gain, with an average value of $\overline{I}_{\rm B} = 0.02$. Applied to the same data, the full-data-collaboration mode (Mode C, bottom) resulted in $\overline{I}_{\rm C} = 0.27$, an order of magnitude increase from Mode B. The improvement from Mode B to C is due to the fact that $\hat{\mathcal{H}}$ is usually a crude approximation of $\mathcal{F}$ and many regions of $\hat{\mathcal{H}}$ are inconsistent with at least one experiment. The essence of this numerical exercise is precisely our main point: combining the entire collection of data into a joint analysis extracts substantially more of the information content of the data.

The computational efficiency of our analysis is also noteworthy. For instance, it took about 100 minutes on a 1.4 GHz Pentium III processor to compute the data displayed in the bottom left panel of Fig. 1. This translates into roughly 1 min for two inner and two outer bounds with a system of 102 variables and 77 experimental dataset units.

We repeated the analysis on other data sets. Taking one of the GRI-Mech experiments as E_0 with the remaining 76 forming the training/assertion set, we obtained, as an average over the 77 such tests, $\overline{I}_{\rm B} = 0.06$ and $\overline{I}_{\rm C} = 0.48$. These larger values of $\overline{I}_{\rm B}$ and $\overline{I}_{\rm C}$ signify that a prediction for a target similar to the GRI-Mech dataset is more informative than for an arbitrary target. While this conclusion is rather trivial, the ability to arrive at it in a systematic way with rigorously-quantifiable measures is not.

To test further, we created a "toy" training set, mimicking the GRI-Mech dimensions but with linear surrogate models, $M_k = \sum_{i=1}^{102} a_{ki} x_i, \quad k = 1, \ldots, 77$. The coefficients a_{ki} were selected randomly, with the singular values of the $[a_{ki}]$ matrix equal to those of the linear part of the GRI-Mech surrogate models. Performing the Mode B and C analysis with the toy dataset resulted in $\overline{I}_{\rm B}^{\rm toy} = 0.00$ (Fig. 3, top right) and $\overline{I}_{\rm C}^{\rm toy} = 0.20$ (Fig. 3, bottom right). The outcome is very similar to the GRI-Mech cases. This demonstrates that the reason for the improvement from B to C is not necessarily the nonlinearity of the individual models M_i, but the correlation

of model parameters implicitly revealed by simultaneously considering all of the constraints.

In no case of collaborative data processing of the GRI-Mech dataset at the $U_i = 0.1$ level of uncertainty with Modes B and C did we see inconsistencies that plagued Mode A. Qualitatively similar results were obtained at uncertainty levels above $U_i = 0.08$. Below this level, our analysis detected that the feasible set $\mathcal{F}$ is empty, indicating that the 77 assertions, coupled with the a priori knowledge, are inconsistent at $U_i < 0.08$. The numerical approach proposed here offers further analysis for such situations, and we will pursue this direction in a future study.

The mode of data processing portrayed by Mode A is not uncommon. On the contrary, variants of Mode A are for the most part the current *modus operandi*, certainly in scientific fields where the progress is built on integration of large amounts of data with the goal of producing accurate forecast; examples may include chemical model building in combustion, atmospheric pollutants, astrophysics, and material synthesis. In fact, our choice of natural gas combustion chemistry for the present work was motivated by the advanced state of science in combustion chemistry, where enough is known to define an ODE model and the parameter uncertainties are confined to relatively small ranges yet are substantial enough to impede further progress.

CONCLUSIONS

The GRI-Mech project fosters the paradigm of collaborative data processing for large-scale, distributed scientific research. The present analysis provides evidence, demonstrated on rigorous mathematical grounds, of the usefulness of this approach. One of the key features of the present analysis is the distinction between the loosely defined, colloquial meaning of scientific collaboration and the term *collaboration of data* introduced here. Our numerical results show that an essential element of scientific collaboration is collaboration of data, demonstrating that combining the entire collection of data into a joint analysis extracts substantially more of the information content of the data.

While the present results were obtained with a particular system, drawn from combustion

chemistry, we see no principal limitations in either the approach or implications for other scientific fields, such as atmospheric chemistry, material synthesis, astrophysics, and biology. Also, active parameters are not limited to rate constants but can be the initial and boundary conditions, thermodynamic or molecular properties, and the like. The logistics enabling the full collaboration of data we advocate can be addressed by the rapidly developing technology of informatics.

APPENDIX A: GRI-MECH DATASET

The purpose of this Appendix is to help the Reader in relating to mathematical definitions of data analysis used in the present study. The more formal language of the main text is designed to address a general situation in a rigorous way, not just chemical kinetics of natural gas combustion. In what follows, we address specifically the GRI-Mech 3.0 release [2].

We define a *dataset* as a collection of dataset units. A *dataset unit* is a numerical summary of an experiment (E): the measured value (D_E), its uncertainly (U_E), and the chemical kinetics model (M_E) that is presumably capable of predicting the measurement D_E. The underlying presumption, and hence the principle on which premises the dataset is organized, is that a *single* chemical kinetics model is capable, on physical grounds, of predicting all experimental values of the dataset. For instance, the present GRI-Mech dataset is built on 77 experiments ($E = 1, 2, \ldots, 77$). Having said that, however, the meanings we assign to "experiment" and "model" need to be qualified further.

The objective of the GRI-Mech project was to develop a "single mechanism" for natural gas combustion. The starting point, as typical for chemical kinetics, was a set of chemical reactions with initially assigned literature rate coefficients. It is a common experience, though, that a model composed from the "best literature values" does not predict equally well *all* experimental data available. The question, and this is the "heart" of the present paper, is how one goes about "tuning" the model parameters (say, reaction rate coefficients). In the GRI-Mech 3.0 release, 77 experimental targets were chosen to serve as a *training set*, i.e., the model was required to fit (or

to be *trained* on) a set of 77 targets. These targets were comprised of species concentrations, ignition delays, flame velocities, shifts in peak positions, etc. Some of these targets were the actual measurements while others were averages of a group or series of measurements. In the context of the present discussion, an experiment E, and hence a dataset unit E, is one of these experimental targets, its value being the measured outcome D_E and its uncertainty U_E.

Qualitatively, the model of a dataset unit can be thought of as the "single chemical kinetic mechanism" (such as the 325-reaction set of GRI-Mech 3.0): for the experimental conditions of experiment E it produces a prediction matching D_E. A "direct" numerical implementation would require solving a set of ordinary differential equations, which presents challenging and often unsurmountable numerical and algorithmic difficulties for model optimization and error propagation. In another approach, called Solution Mapping [3], one can express the relationships between the model input and model output in a parameterized form, referred to as a *surrogate* model. Each dataset unit E has such a surrogate model, M_E.

Let us for a moment think of M_E as a solution of the single ODE kinetic model at conditions of experiment E. In the context of model optimization, the initial conditions of ODE integration for a dataset unit 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, ratios of rate coefficients, and enthalpies of formation). GRI-Mech dataset has 102 optimization variables, $x_1, x_2, \ldots, x_{102}$. Only a small fraction of these is *active* at conditions of experiment E, i.e., has a measurable influence on the measured outcome D_E. Hence, our notation $M_E(x)$ implies a surrogate model of dataset unit E, a function of a set of optimization variables x active at conditions of experiment E.

In the present study we expressed optimization variables in a normalized and centered form, essentially keeping them as factorial variables. For instance,

$$x_{44} = \frac{\ln\left(A_{126}/A_{126,0}\right)}{\ln s_{126}}$$

is an optimization variable associated with reaction (numbered 126 in the GRI-Mech 3.0 dataset)

$$\mathrm{CH} + \mathrm{H}_2 \longrightarrow \mathrm{H} + \mathrm{CH}_2$$

where A_{126} is the pre-exponential factor of the reaction rate coefficient, $A_{126,0}$ its initial value taken from literature, and s_{126} the span of variation in A_{126}, thus implying variation of A_{126} from $A_{126,0} \times s_{126}$ to $A_{126,0}/s_{126}$. Expressed in this way, the optimization variable x_{44} varies from -1 to $+1$ and equals 0 at the center of this variation interval, i.e., at the literature recommendation for the associated rate coefficient. Defined in this manner, all optimization variables form a 102-dimensional hypercube, $\mathcal{H}$, with each side ranging from -1 to $+1$, and its center ($x_i = 0$, $i = 1, 2, \ldots, 102$) corresponding to the initial set, i.e., the "literature recommendation".

Defined in the normalized and centered way, the factorial representation of optimization variables in surrogate functions has an additional implication. Whereas a derivative

$$S_A = \frac{\partial \ln [\mathrm{CH}]_{\max}}{\partial \ln A_{126}}$$

provides a sensitivity of the CH peak concentration (the response of the SCH.C12 target) with respect to the actual parameter A_{126} [3], derivative

$$S_x = \frac{\partial y_{66}}{\partial x_{44}}$$

specifies an *impact* of A_{126} on the response of E_{66} ($y_{66} = \ln [\mathrm{CH}]_{\max}$). The parameter impact on a response is defined as $|sensitivity \times uncertainty|$, which in essence "scales" the sensitivity by the range of (allowed) variation in the parameter value, $S_x = S_A \times \ln s$.

APPENDIX B: SOS OPTIMIZATION AND $\mathcal{S}$-PROCEDURE

One of the key methodologies in Robust Control is the *sum of squares* (SOS) method of optimization [10, 11]. It exploits the polynomial nature of the constraints and objective. For instance, determination of the upper bound of the Mode C posterior range, $\max_{x \in \mathcal{F}} M_0(x)$, can be expressed as a set containment question: Given γ, do the constraints $x \in \mathcal{H}$ and

$\{-U_i \leq M_i(x) - D_i \leq U_i\}_{i=1}^{77}$ imply that $M_0(x) \leq \gamma$? To answer this question, we define the sublevel set for a polynomial p of n real variables, $L_p := \{x \in \mathbf{R}^n : p(x) \leq 0\}$, and pose the following question for an arbitrary set of polynomials: Given polynomials $p_0, p_1, \ldots, p_N$, is $\bigcap_{l=1}^{N} L_{p_l} \subseteq L_{p_0}$? If there exist globally nonnegative polynomials λ_l such that $-p_0 + \sum_{l=1}^{N} \lambda_l p_l$ is globally nonnegative then containment $\bigcap_{l=1}^{N} L_{p_l} \subseteq L_{p_0}$ holds; indeed, p_0 is nonpositive wherever all of the p_l are nonpositive.

In general, checking global nonnegativity of a polynomial is computationally complex [12], unless the polynomial is SOS, i.e., expressed as a sum of squares of other polynomials, $\sum_i f_i^2$. Determining whether a polynomial is SOS is accomplished within the optimization framework of Semidefinite Programming (SDP) [13, 14]. Our problem is reduced then to the manageable task of finding SOS λ_l such that $-p_0 + \sum_{l=1}^{N} \lambda_l p_l$ is SOS, and again is decided with SDP. If all p_l are quadratic functions and all λ_l are restricted to be nonnegative constants, the sufficient condition that $-p_0 + \sum_{l=1}^{N} \lambda_l p_l$ be a positive semidefinite quadratic function is known as "the $\mathcal{S}$ procedure" [10]. Given the quadratic form of the GRI-Mech surrogate models, M_i, we used the $\mathcal{S}$ procedure to determine outer bounds.

A more detailed account of the numerical procedure and optimization is given in [15].

References

[1] Frenklach, M.; Wang, H.,; Rabinowitz, M. J. Prog Energy Combust Sci 1992, 18, 47.

[2] 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/.

[3] Frenklach, M., in Combustion Chemistry; Gardiner, W. C., Jr. (Ed.); Springer-Verlag: New York, 1984, p 423.

[4] Box, G. E. P.; Draper, N. R. Empirical Model-Building and Response Surfaces; Wiley: New York, 1987.

[5] Myers, R. H.; Montgomery, D. C. Response Surface Methodology; Wiley: New York, 2002.

[6] Frenklach, M.; Packard, A.; Seiler; P. In Proceedings of the American Control Conference, IEEE Catalog Number: 02CH37301C, 2002, p 4135.

[7] Papadimitriou, C. H.; Steiglitz, K. Combinatorial Optimization: Algorithms and Complexity; Dover: New York, 1998.

[8] Murty, K. G.; Kabadi, S. N. Math Programm 1987, 39, 117.

[9] Vavasis, S. A. Informat Process Lett, 1990, 36, 73.

[10] Boyd, S.; Ghaoui, L. E.; Feron, E.; Balakrishnan, V. Linear Matrix Inequalities in System and Control Theory; Studies in Applied Mathematics, Volume 15, SIAM, 1994.

[11] Parrilo, P. PhD thesis, California Institute of Technology, 2000 (available at http://www.aut.ee.ethz.ch/ parrilo/pubs/).

[12] Reznick, B. Some concrete aspects of Hilbert's 17th problem. Preprint available at http://www.math.uiuc.edu/Reports/reznick/98-002.html.

[13] Vandenberghe, L.; Boyd, S. SIAM Review 1995, 38, 49.

[14] Sturm, J. SeDuMi version 1.05, 2001; http://fewcal.kub.nl/sturm/software/sedumi.html.

[15] Seiler, P.; Frenklach, M.; Packard, A.; Feeley, R. "Numerical approaches for developing predictive models", in preparation.

FIGURE CAPTIONS

Fig. 1 Feasible sets obtained in analysis of Experiments 66 (top panel) and 67 (bottom panel) individually, with all x's but x_{44} and x_{45} set to 0 (see text).

Fig. 2 A feasible set obtained in a joint analysis of Experiments 66 and 67, with all x's but x_{44}, x_{45}, and x_{34} set to 0.

Fig. 3 Frequency of information gain, I, for the Free-Core (Case B, top panels) and Full-Collaboration (Case C, bottom panels) modes of data processing on the GRI-Mech (left panels) and Toy (right panels) datasets.

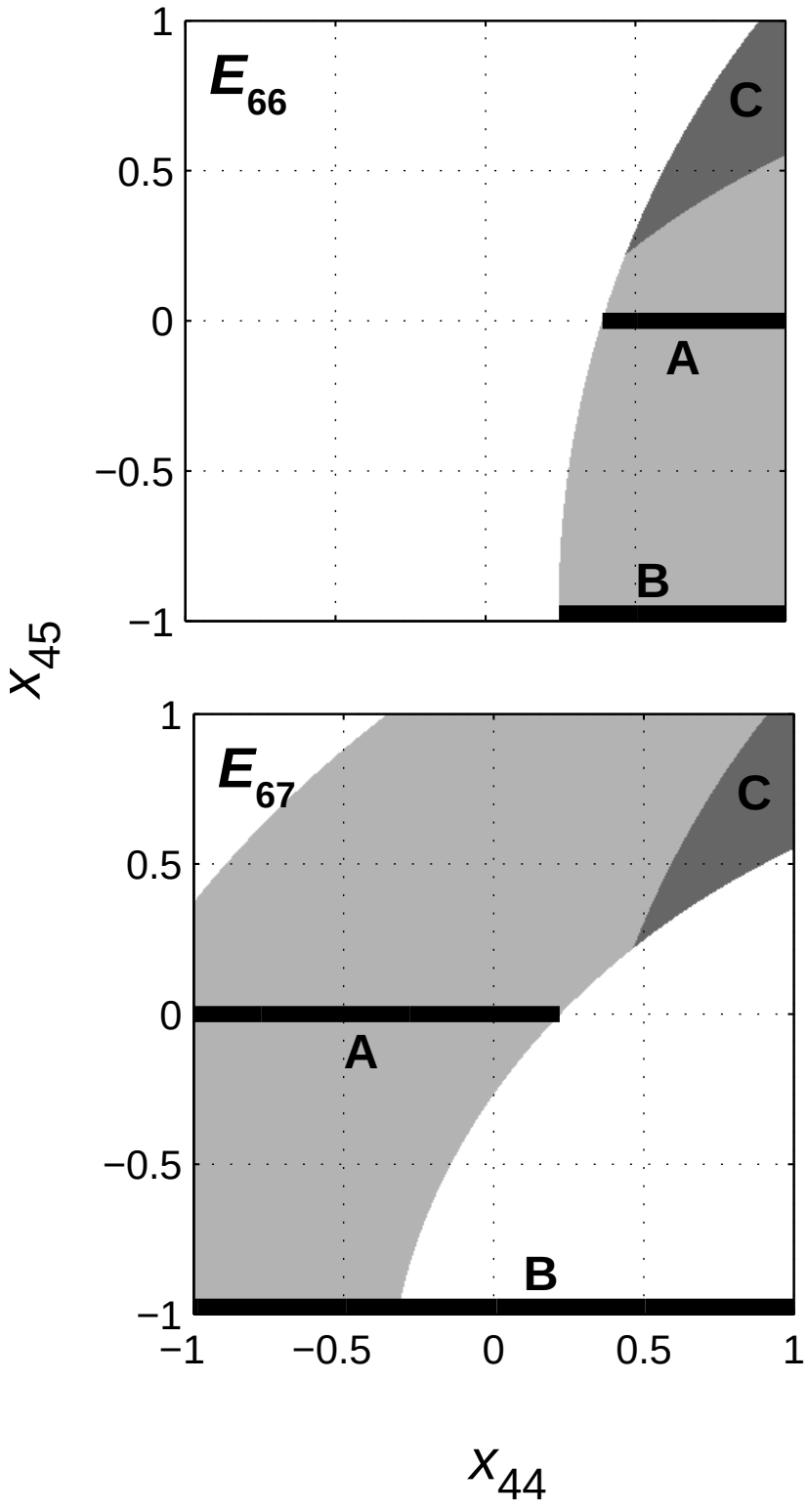

Figure 1: Feasible sets obtained in analysis of Experiments 66 (top panel) and 67 (bottom panel) individually, with all x's but x_{44} and x_{45} set to 0 (see text).

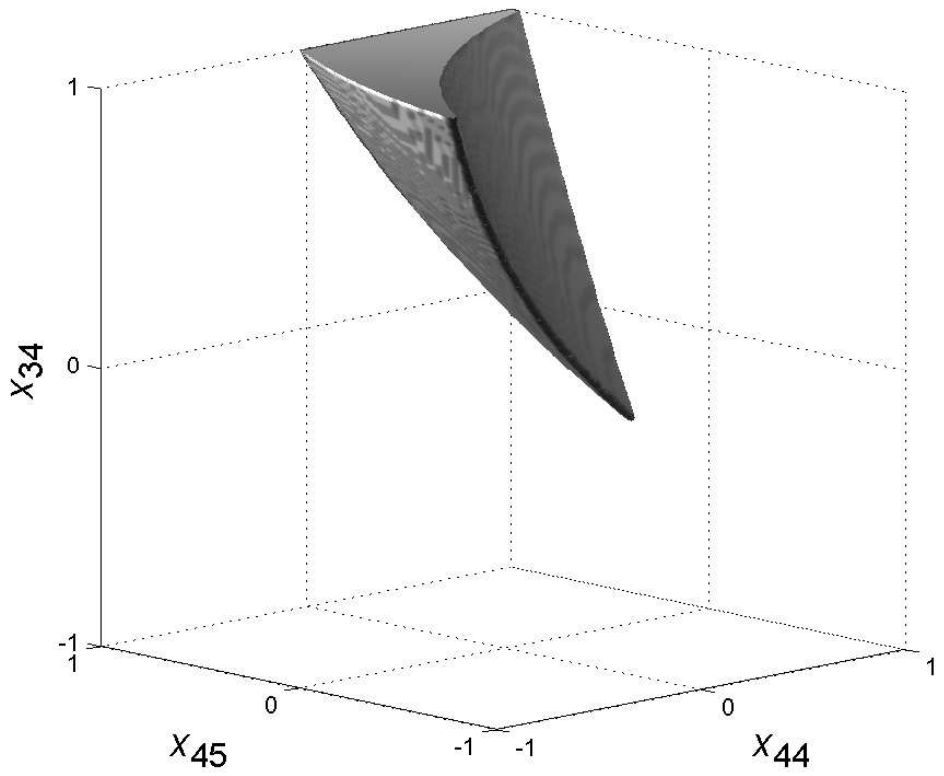

Figure 2: A feasible set obtained in a joint analysis of Experiments 66 and 67, with all x's but x_{44}, x_{45}, and x_{34} set to 0.

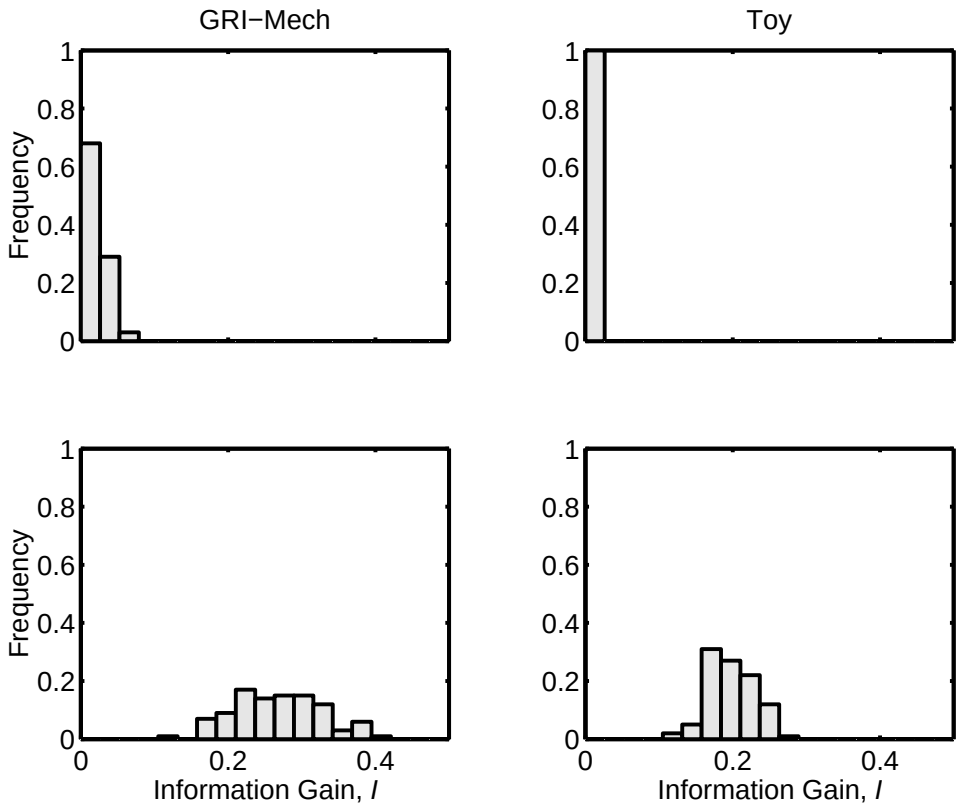

Figure 3: Frequency of information gain, I, for the Free-Core (Case B, top panels) and Full-Collaboration (Case C, bottom panels) modes of data processing on the GRI-Mech (left panels) and Toy (right panels) datasets.