International Journal for Uncertainty Quantification, 4 (2): 111–132 (2014)

DATA-FREE INFERENCE OF UNCERTAIN PARAMETERS IN CHEMICAL MODELS

Habib N. Najm, **Robert D. Berry, Cosmin Safta, Khachik Sargsyan, & Bert J. Debusschere***

P.O.Box 969, MS 9051; Sandia National Laboratories, Livermore, California 94551, USA

Original Manuscript Submitted: 06/26/2012; Final Draft Received: 01/17/2013

We outline the use of a data-free inference procedure for estimation of uncertain model parameters for a chemical model of methane-air ignition. The method involves a nested pair of Markov chains, exploring both the data and parametric spaces, to discover a pooled joint posterior consistent with available information. We describe the highlights of the method, and detail its particular implementation in the system at hand. We examine the performance of the procedure, focusing on the robustness and convergence of the estimated joint parameter posterior with increasing number of data chain samples. We also comment on comparisons of this posterior with the missing reference posterior density.

KEY WORDS: *uncertainty quantification, data-free inference, Bayesian, ignition, chemistry*

1. INTRODUCTION

Analysis and design of a wide variety of energy systems rely on computational predictions employing chemical kinetic models. These models are composed of reaction networks describing interactions among chemical species via elementary reaction steps. Relevant models typically employ many reaction steps, particularly so for complex reactant molecules. Broadly, such chemical kinetic models are relevant in a wide range of applications, including e.g., combustion, chemical processing, biochemistry, and geochemistry, and can involve heterogeneous reactions among multiple phases. There are many sources of uncertainty in this broad landscape, including chemical model structure, model parameters, initial and boundary conditions, general modeling assumptions such as homogeneity and transport, as well as underlying stochastic dynamics. In the present context, and as a demonstration case with select challenges corresponding to exothermic ignition, we focus particularly on the inference of uncertain parameters in homogeneous gas-phase chemical kinetic models relevant in combustion.

It is noteworthy that, in the combustion chemistry context, detailed kinetic models for the oxidation of complex hydrocarbons can involve $\mathcal{O}(10^3)$ species, and $\mathcal{O}(10^4)$ reversible reactions. The minimal specification of an elementary reaction step involves the reaction partners, stoichiometric coefficients, and a reaction rate [1]. Elementary reaction rates can be, in general, temperature dependent, and are commonly described using Arrhenius rate expressions [1], $k(T) = AT^n \exp(-E/R^oT)$, where R^o is the universal gas constant, and (A, n, E) are the Arrhenius rate parameters, being defined as the pre-exponential rate constant, temperature exponent, and activation energy, respectively. Most commonly, these parameters are either determined experimentally using (e.g., least-squares) fitting [2, 3], or are derived, using rate-rules [4–7], from other measured Arrhenius parameters. Accordingly, they are known only to within a certain degree of uncertainty.

Given that the input space of chemical models, namely the set of model parameters, is uncertain, it is important to analyze the impact of this uncertainty on model predictions. There have been great advances in uncertainty quantification (UQ) over the past couple of decades, particularly in the context of probabilistic UQ methods, where uncertain quantities are defined as random variables (or fields) [8–20]. In contrast with local error propagation or moment

*Correspond to Habib N. Najm, E-mail: hnnajm@sandia.gov

2152–5080/14/$35.00

methods [21–23], probabilistic methods can account accurately for the full range of uncertainty in model parameters, allowing handling of systems with large and arbitrarily distributed input uncertainties, or strongly nonlinear systems with large amplification of input uncertainties. The use of probabilistic forward UQ methods, on the other hand, does necessitate a similarly probabilistic characterization of the uncertain input space. This requires the use of statistical methods for inference of model parameters based on experimental measurements, where the outcome is a joint probability density function (PDF) on the uncertain input parameters, which is not available from, say, least-squares fitting methods. In particular, Bayesian inference methods [24, 25] provide a convenient framework for estimating the joint PDF on the input space given experimental data.

Prior UQ work in chemical systems has explored the forward propagation of uncertainty in chemical ignition, where the parameters were presumed independent and identically distributed (*i.i.d.*) for lack of information on any correlation/dependence among them [26–32]. The extent of each parameter PDF was defined according to its known error bars, where both lognormal and uniform PDF structures were employed in different works. On the other hand, Najm et al. [33] have established the importance of knowing the correlation structure of the joint PDF on the parameters. Specifically, depending on the chemical model at hand, and parameters of interest, the slope of the dependence relationship among model parameters can have a large impact on the resulting uncertainty in model predictions [33]. Yet published chemical models are typically specified with nominal values of rate parameters, and with at most error bars on the pre-exponential rate constant [34, 35]. Neither the uncertainties on the other model parameters (n, E), nor the correlations among any of the parameters, are typically specified in the literature. Therefore, one may well say that forward UQ studies with this published information are largely exercises in the art, but do not deliver reliable uncertainty estimates. Clearly, there is a great need for reanalysis of available data, as well as new data, to establish joint PDFs on chemical model parameters. However, raw data, even from recent experiments, are rarely available, and repeating experiments is both costly and very time-consuming. Therefore, it will be a long time before the fully specified probabilistic structure of the input space of complex chemical models is well characterized based on measurements.

In the meantime, however, it is of great interest to explore what we *do* know from previously processed data, and to constrain input uncertainties of chemical models to be consistent with this information. The key idea can be illustrated in a simple example as follows. Consider that investigator $\mathcal{A}$ has collected data, fitted it with a straight line using least-squares fitting, published the experimental details along with the observed nominal values and $\pm 3\sigma$ uncertainty ranges in the slope α and intercept β of the line, but did not publish the raw data itself, and then duly destroyed the data. While the two parameters *are* correlated by the fitting, this information is not reported. Then, investigator $\mathcal{B}$ comes along, requires the joint PDF on (α, β), but there are no available data to redo the analysis. One option is to presume (α, β) independent, assigning each a PDF consistent with its published mean and standard deviation values. However, such a PDF would assign finite probability to some parameter values that clearly do not fit the missing data, given that the true PDF exhibits a degree of correlation. In this situation, it is clear that, while the data are missing, other information is in fact available and can implicitly provide constraints on the PDF. Specifically, $\mathcal{B}$ knows that the fitting was done over some range of the data in the experiment, and that the fitting employed a straight line. Further, given $\mathcal{B}$'s knowledge of the nominal line, it is clear that, if, say, the intercept is increased beyond its nominal value, the slope would have to be changed in a manner that gives a line that is still in the vicinity of the nominal line. In other words, much of the information in the missing data, that would inform the sought-after correlation structure of the PDF, is in fact available in the other reported information. The question is, how to make this information *explicit* in the PDF structure in a manner that provides a PDF that is consistent with other given information.

While this problem is related to the missing data problem [36–40], it is different in the sense that *all* the data are missing. Recently, Berry et al. [41] provided a general "Data-Free" Inference (DFI) methodology, founded on maximum entropy arguments, to handle this problem. DFI involves a two level random walk procedure, one on the data space and another on the parametric space. The algorithm proposes hypothetical data sets, retaining those that, when employed for fitting the given model parameters, provide nominal values and bounds that are consistent with the given information. The retained/consistent data sets are then used to provide a pooled/averaged joint PDF on the parameters. While this is not generally expected to converge to the original PDF in any sense, it is at least consistent with the given information. The procedure was demonstrated, and shown to be successful, on algebraic function fits. Much needs to be done to improve its efficiency, and to demonstrate its accuracy in more complex situations. The goal

of this paper is to explore its utility and performance in the context of a highly nonlinear ordinary differential equation (ODE) system, being the set of governing equations for chemical ignition of a hydrocarbon fuel.

In the following, we outline the problem setup, providing an artificial data set on the ignition of a methane-air mixture, and then use it to fit parameters of a simple chemical model. Discarding the posterior PDF on the parameters, while retaining its nominal and marginal bounds, we then employ DFI to discover a consistent pooled posterior. We evaluate the pooled posterior, averaging information from a range of data-space samples, against the reference posterior. We outline the degree to which the two are similar, highlighting the degree of information, and its utility for constraining the parametric PDF without recourse to the discarded original data.

2. PROBLEM SETUP

We begin by generating synthetic data using an ignition solution computed with a detailed chemical model, with added noise. We will use these data to calibrate a simple global chemical model, providing the reference posterior as the target for the subsequent use of DFI below.

2.1 Ignition with a Detailed Model

We consider a preheated stoichiometric methane-air mixture, at atmospheric pressure. We compute the constant-pressure homogeneous ignition of this system using GRImech3.0 [42], over a range of initial temperature $T^o \in [1000, 1300]$ K. The system evolves through a preheat phase during which a radical pool develops and the mixture slowly heats up, before going through a fast ignition characterized by a rapid rise in temperature, consumption of reactants, and formation of combustion products. Defining the ignition delay time τ as the time instant at which $T = 1500$ K, we observe the expected decrease in τ with increasing T^o, as shown in Fig. 1.

Then, using $N = 11$ data points, defined on a uniform mesh over this range, $\{T_i^o\}_{i=1}^N$ with the associated set of computed ignition delay times $\{\tau_i\}_{i=1}^N$, and employing a Gaussian multiplicative noise term to represent measurement error, we generate a synthetic data set $\{T_i^o, \tau_i^d\}_{i=1}^N$, where

$$\tau_i^d \equiv \tau_i \, (1 + \sigma \epsilon_i), \qquad i = 1, \ldots, N, \tag{1}$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$, and $\sigma^2 = 0.02$. Note that this multiplicative noise model is equivalent to an additive noise with a standard deviation that is proportional to the signal level. The variation of τ, and the noisy τ^d, with T^o is also shown in Fig. 1

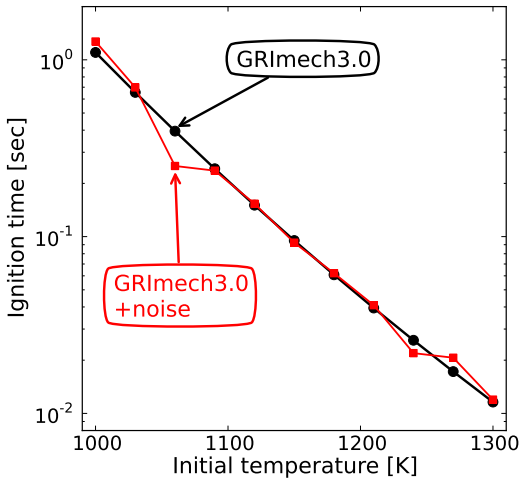

FIG. 1: Ignition time computed with GRImech3.0 over a range of variation of initial temperature, including also the corresponding noisy synthetic data points used for inference.

2.2 Fitting of a Simple Model

We employ a simple chemical model, to be calibrated per the above data, based on a global single-step irreversible methane mechanism,

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O \tag{2}$$

with the forward rate of progress[1]

$$\mathcal{R} = [CH_4][O_2]k(T), \tag{3}$$

where $[\cdot]$ is concentration (mol/cm^3), $k(T)$ is the forward rate, modeled using the Arrhenius rate expression

$$k(T) = A\exp(-E/R^oT). \tag{4}$$

where R^o is the universal gas constant, T is the temperature, E is the activation energy (cal/mol), and A is the pre-exponential constant (cm^{3}mol^{-1}s^{-1}). The goal is to infer the pair of parameters (A,E).

We use Bayesian inference to estimate (A,E) given the above model and the synthetic noisy data. This fitting presumes knowledge of the above form of the error model [Eq. (1)], but leaves σ as a hyperparameter to be inferred. More specifically, we infer the parameters $(\ln A, \ln E)$, and the hyperparameter $\ln\sigma$. Since A and σ are strictly positive quantities, it is natural to consider their logarithms as the object of inference, as this enforces their positivity. The use of $\ln E$, versus E, is a matter of convenience in the present setting, as it proved easier to attain good mixing in the Markov Chain Monte Carlo (MCMC) [43] procedure with $\ln E$. Further, this is acceptable in the present context because the range of feasible values of E is far above zero. In a situation where values of $E \leq 0$ provide feasible fits of the data, it is necessary to infer E directly rather than its logarithm.

There is a range of MCMC algorithms that can be employed in general. In the present context, we have found that the degree of strong correlation in the posterior density renders the choice of a good proposal distribution difficult. Accordingly, it is useful to employ an MCMC algorithm that is adaptive, in the sense that it learns and adapts the proposal distribution based on existing samples. To this end, we employ an adaptive MCMC (AMCMC) [44, 45] algorithm to estimate the posterior density on the parameters $(\ln A, \ln E, \ln\sigma)$. Starting from an initial specified state, and an initial guess at a covariance matrix for the multivariate normal proposal distribution, this method proceeds for a specified number of steps with this proposal distribution, before beginning the adaptive phase. In this phase, the covariance matrix of the proposal distribution is updated based on the continuously increasing sample set. In the present context, the method generates an MCMC chain with good mixing despite the sharp posterior density.

For a prior, we consider the three parameters to be *i.i.d.* uniform, with $\ln A \in [5,100]$, $\ln E \in [2.5,50]$, and $\ln\sigma \in [-5,0]$. The prior choice is generally based on available information prior to taking data. This obviously depends on the practitioner's knowledge about the chemical system at hand, i.e., of $(\ln A, \ln E)$, and of the instrument noise, i.e., of $\ln\sigma$. This being an artificial situation, with simulated data, our prior bounds on each of the three parameters are simply chosen to be sufficiently wide to represent a significant degree of ignorance about each. The intention, for all three parameters, is to specify ranges that do not in fact impact the MCMC chain, which we checked *a posteriori*. We could have also simply used improper, i.e., unbounded, uniform priors on each of $(\ln A, \ln E, \ln\sigma)$, as the present fitting is sufficiently constrained by the data, and there is no requirement for regularization through the prior.

As for the likelihood function, it is constructed as follows. With $\beta \equiv (\ln A, \ln E)$, and given the known noise model structure for the ignition time τ, as shown in Eq. (1) above, we have $\tau = \tau^m(T^o,\beta)(1+\sigma\epsilon)$, where $\tau^m()$ is τ as computed by the fit model at the given (T^o,β), we have

$$\tau = \tau^m(T^o,\beta) + \tau^m(T^o,\beta)\sigma\epsilon, \tag{5}$$

such that

$$\tau\,|\,T^o,\beta,\sigma \;\sim\; N\left(\tau^m(T^o,\beta), [\tau^m(T^o,\beta)\sigma]^2\right). \tag{6}$$

[1]Note that, if the reaction in Eq. (2) were an elementary reaction step, its rate of progress would involve the square of $[O_2]$. However, with the present *global* reaction the exponents on $[CH_4]$ and $[O_2]$ are largely a convenient modeling choice, and are chosen here to be unity.

Considering the full data set $z = \{T_i^0, \tau_i^d\}_{i=1}^N = \{u, v\}$, where $u = \{T_i^0\}_{i=1}^N$ is the independent data, and $v = \{\tau_i^d\}_{i=1}^N$ is the dependent data, with the above *i.i.d.* noise model, the corresponding likelihood for v is given by

$$p(v\,|\,u,\beta,\ln\sigma) = \frac{1}{(2\pi\sigma^2)^{N/2}} \frac{1}{\prod_{i=1}^N \tau^m(T_i^o,\beta)} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^N \left[\frac{\tau_i^d}{\tau^m(T_i^o,\beta)} - 1\right]^2\right\}. \tag{7}$$

It is important to start the AMCMC procedure at a point that is close to the posterior maximum, because of the strong correlation of the posterior density, as will be seen below. In this low-dimensional parameter space, we discover the posterior maximum by inspection of the likelihood surface, and start the chain there. More generally, least-squares fitting can be employed to find the maximum likelihood estimate. Further, we start the chain with an initial proposal covariance matrix with only diagonal nonzero terms. Adequate initial proposal widths in each dimension were found by trial and error. Good mixing is achieved after initial transients on all three parameters, as can be seen in Fig. 2, thereby providing reliable estimation of the posterior.

Employing a 10^5-step-long AMCMC chain, and neglecting a burn-in chain length of 5000 steps, the resulting marginals for $\ln A$ and $\ln E$, produced with kernel density estimation (KDE) employing the chain samples, are shown in Fig. 3. The corresponding means are $(32.17, 10.73)$, and the marginal standard deviations are $(0.58, 0.031)$, respectively. Further, the marginal joint posterior pdf for $(\ln A, \ln E)$ is shown in Fig. 4, again estimated using KDE. Clearly there is a strong correlation between the two parameters. The posterior is a very narrow ridge with a well defined slope as observed in other settings [33]. Further, the posterior density is sufficiently far from the edges of the uniform prior ranges, such that there is no direct impact due to the uniform prior bounds on the posterior. Finally, we note that the global chemical model evaluated with the posterior mean parameter values provides a good fit with respect to the original detailed GRImech3.0 ignition time predictions, as seen in Fig. 5.

Having employed Bayesian inference to estimate the uncertain parameters using synthetic data, we now purposely discard these data, and retain only the following summary statistics and range information, for the subsequent application of DFI:

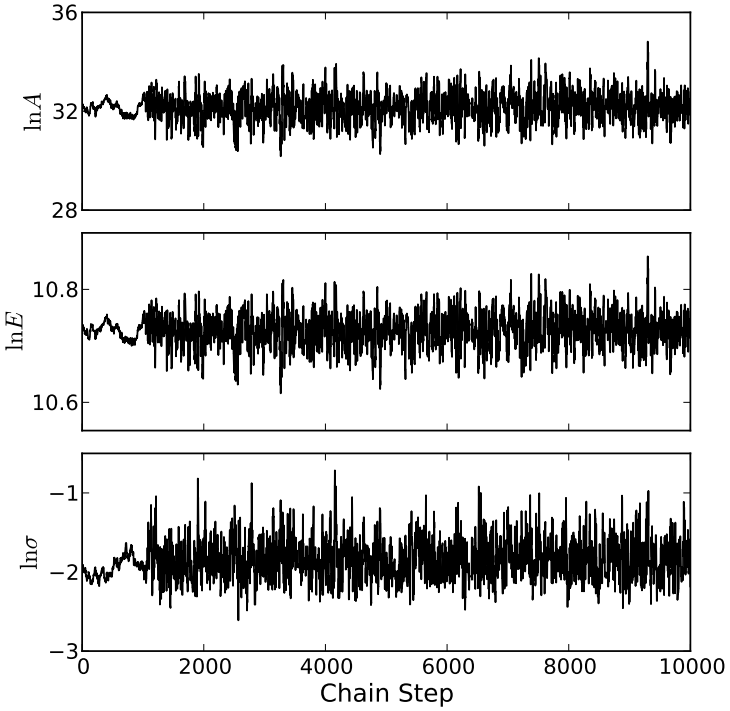

FIG. 2: Mixing in the AMCMC chain over the first 10^4 steps.

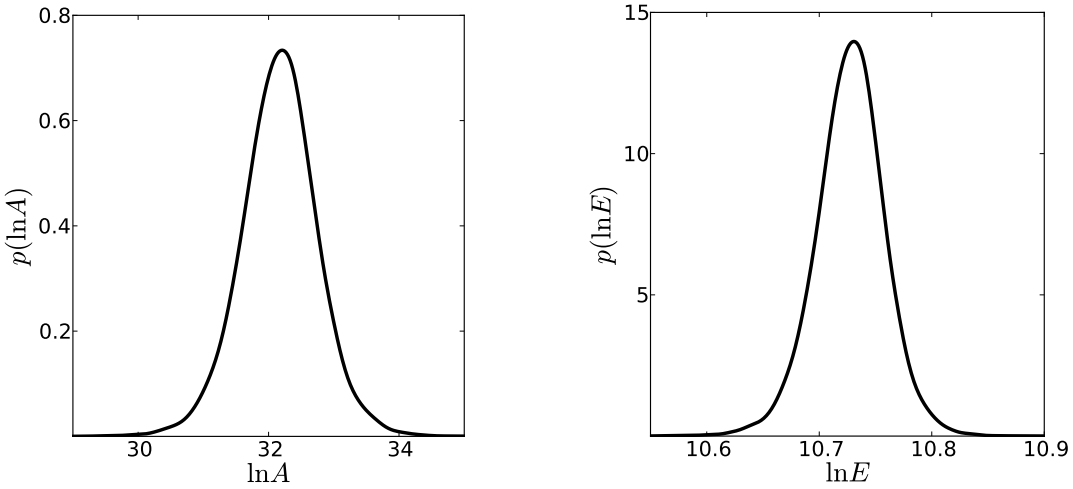

FIG. 3: One-dimensional marginal posteriors on $\ln A$ (left) and $\ln E$ (right).

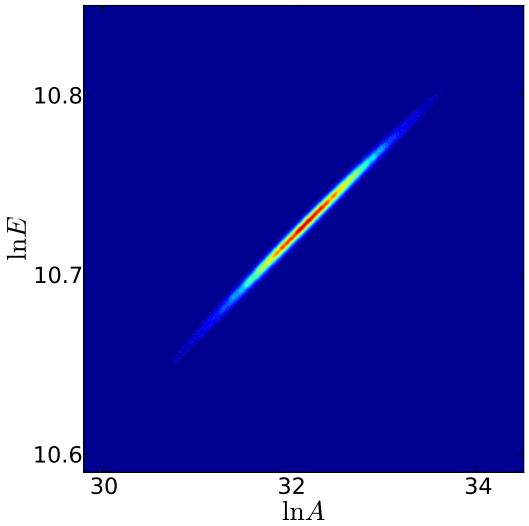

FIG. 4: Joint 2D marginal posterior on $(\ln A, \ln E)$.

1. Initial temperature $T^o \in [1000, 1300]$
2. Nominal parameters $\mathbb{E}(\ln A, \ln E) \equiv \beta_0 = (32.17, 10.73)$
3. 5% marginal bounds on $\ln A$ and $\ln E$: $\beta_5 = [(\ln A)_5, (\ln E)_5] = (31.22, 10.68)$
4. 95% marginal bounds on $\ln A$ and $\ln E$: $\beta_{95} = [(\ln A)_{95}, (\ln E)_{95}] = (33.09, 10.78)$

This replicates a typical situation where an experimental measurement was made, and such summary information is reported in the literature, but not the joint PDF or correlation among the parameters, and where the original raw data are not reported. Note that the specific choice of the 5% and 95% quantiles is simply for illustration. Other quantiles, or other statistics of the posterior density, can be similarly incorporated in the algorithm

We also do not presume knowledge of the *number* of data points used in the original fitting. It should be noted that this is a very common occurrence. On the other hand, we do presume knowledge of the instrument fitting model. This

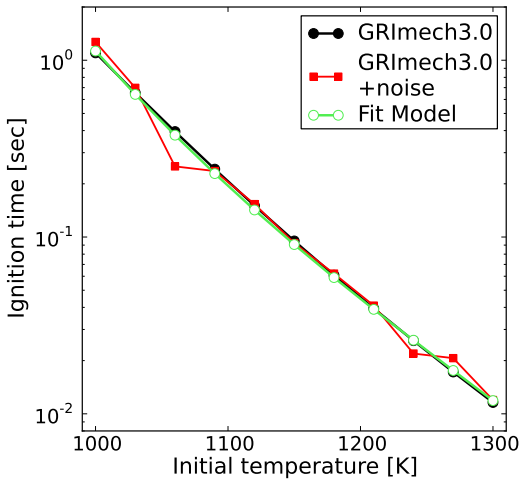

FIG. 5: Comparison of the GRImech3.0 ignition delay time predictions, and those from the posterior mean global model predictions. Also superposed is the synthetic data used in the inference.

requires some knowledge of the experimental system, in this case, the constant pressure ignition model, the initial reactants mixture composition, the definition of the ignition delay time, and the characteristics of the measurement noise model. We note that the latter is frequently well known for a typical experimental system, e.g., whether the noise is Gaussian or Poisson, or whether the noise amplitude is signal dependent.

With this setup in place, we now proceed to a description of the DFI algorithm, and its application in the present context to discover a joint PDF on the parameters of interest that is consistent with the given information.

3. DFI ALGORITHM OUTLINE

The DFI algorithm is outlined in significant detail in [41]. Here we give an essential outline for completeness.

The objective of the algorithm is to estimate a posterior density on model parameters that is consistent with the given constraints, in the absence of data. In the present context, the algorithm takes as inputs the specification of the fit model, the instrument noise model, the presumed number of data points, and the initial temperature range, nominal parameter values, and marginal bounds. Its key output is a posterior density on $(\ln A, \ln E)$. The algorithm explores the data space using an MCMC procedure, accepting those data sets that, when employed for inferring $(\ln A, \ln E)$, result in a posterior density that satisfies the stated nominals and marginal bounds. Each of the acceptable data sets is essentially equivalent to the missing data given the available information. Accordingly, acceptable data sets are pooled, providing an appropriately averaged posterior that represents our knowledge of $(\ln A, \ln E)$.

More specifically, the construction involves a nested pair of MCMC chains, with an outer chain on the data space, and an inner chain on the parameter space. At each step of the outer/data chain, a data set is proposed, which is evaluated using the inner/parameter chain for consistency with the given information. The inner chain estimates a posterior density on model parameters, given the data set provided by the outer chain. The consistency check, which makes use of parameter samples from the inner chain, constitutes the likelihood function for the outer chain. Each accepted data chain step provides, via the inner chain, a consistent posterior on the model parameters. We employ logarithmic pooling, following [41], to arrive at a consensus/averaged pooled posterior on the parameters. This pooled posterior is the essential output of the algorithm.

We can write the two Bayesian problems solved by the MCMC chains as follows. Let the given information be I, the state vector for the outer (data) chain be ζ, and the state vector for the inner (parameter) chain be λ. We then have the outer chain solving the Bayesian problem

$$p(\zeta|I) \propto p(I|\zeta)\pi(\zeta) \tag{8}$$

and the inner chain solving

$$p(\lambda|\zeta) \propto p(\zeta|\lambda)\pi(\lambda), \tag{9}$$

where the likelihood function of the outer chain is defined based on the posterior of the inner chain, thus $p(I|\zeta) = F(p(\lambda|\zeta), I)$. For each proposed ζ in the outer chain, the inner chain is run, providing the posterior $p(\lambda|\zeta)$. This then provides the estimate of $p(I|\zeta)$, the likelihood function of the outer chain, and the posterior $p(\zeta|I)$, which enables the jump to the next outer chain step.

Delving into more detail, we note that, along with the "data" set proposed at each step of the outer chain, the algorithm proposes the single nuisance parameter, namely the log standard deviation of the data set $\ln\sigma_d$, such that, with each data set comprised of N data points, each data point being a pair of (T_i^0, τ_i^d), the outer chain is $2N+1$-dimensional, where

$$\zeta = (T_1^0, \tau_1^d, \ldots, T_N^0, \tau_N^d, \ln\sigma_d) = (z, \ln\sigma_d), \tag{10}$$

where $z = (T_1^0, \tau_1^d, \ldots, T_N^0, \tau_N^d)$. The state vector of the inner chain is

$$\lambda = (\ln A, \ln E, \ln\sigma) = (\beta, \ln\sigma), \tag{11}$$

where $\beta = (\ln A, \ln E)$. Given the proposed ζ, particularly its subset z, the inner chain is run, arriving at a posterior density on $(\ln A, \ln E, \ln\sigma)$.

Recalling the above decomposition of the data vector, $z = \{u, v\}$, the formulation of the inner chain likelihood function for any given $(\beta, \ln\sigma)$ is given by

$$p(z\,|\,\beta, \ln\sigma) = p(u, v\,|\,\beta, \ln\sigma) = p(v\,|\,u, \beta, \ln\sigma)\, p(u\,|\,\beta, \ln\sigma). \tag{12}$$

Then, given that all we know about the independent data vector $u = \{T_1^o, \ldots, T_N^o\}$, is that it is within the interval $[1000, 1300]$, we rely on maximum entropy arguments, dictating a uniform density on this interval for u. Accordingly, $p(u|\beta, \ln\sigma) = 1/300$ for proposed data sets on this interval, and we have, per Eq. (7),

$$p(z\,|\,\beta, \ln\sigma) \propto p(v\,|\,u, \beta, \ln\sigma) = \frac{1}{(2\pi\sigma^2)^{N/2}} \frac{1}{\prod_{i=1}^N \tau^m(T_i^o, \beta)} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^N \left[\frac{\tau_i^d}{\tau^m(T_i^o, \beta)} - 1 \right]^2 \right\}. \tag{13}$$

This then is the inner chain likelihood function for any $(\beta, \ln\sigma)$.

The key step in the algorithm is the use of the inner-chain posterior density to construct the likelihood function for the outer chain. The construction of the likelihood is problem specific, being defined by the available information. Given that, here, we presume knowledge of the nominal values of $(\ln A, \ln E)$ and their marginal bounds, we construct the likelihood function accordingly to penalize deviations from these given values, as outlined below.

Denoting the nominal parameters by β_0, we define the likelihood function for the outer chain, capturing the consistency of the proposed data set with the given model parameters nominals and bounds, using the following normalized consistency weight function [41]:

$$p(I|\zeta) = w_\delta(z, \ln\sigma_d) = F_\delta(z)\, \frac{p(z, \ln\sigma_d\,|\,\beta_0)}{\max\limits_{\beta,\sigma}[p(z, \ln\sigma\,|\,\beta)]}. \tag{14}$$

The term $p(z, \ln\sigma_d\,|\,\beta_0)$ measures the likelihood of the $(z, \ln\sigma_d)$ data given the nominal model parameters. It is computed using the likelihood function of the inner chain [Eq. (13)] evaluated at $(\beta_0, \ln\sigma_d)$, since

$$p(z, \ln\sigma_d\,|\,\beta_0) = p(z\,|\,\beta_0, \ln\sigma_d)\, \pi(\ln\sigma_d\,|\,\beta_0) = p(z\,|\,\beta_0, \ln\sigma_d)\pi(\beta_0, \ln\sigma_d)/\pi(\beta_0), \tag{15}$$

where $p(z\,|\,\beta_0, \ln\sigma_d)$ is the inner chain likelihood function, and $\pi(\beta_0, \ln\sigma_d)$ the inner chain prior, both evaluated at the given $(\beta_0, \ln\sigma_d)$. Distinct from the baseline version of the algorithm in [41], the likelihood term $p(z, \ln\sigma_d\,|\,\beta_0)$ is normalized in Eq. (14) by its maximum value, evaluated from the inner chain samples, such that the ratio is in [0,1]. This makes the construction more robust to situations where z with a small σ_d is proposed in the data chain, resulting

in a large peak amplitude of $p(z,\ln\sigma|\beta)$, that could dominate the bounds check, in the absence of the normalization, to the point of ignoring it. The bounds check is provided by $F_\delta(z)$, as discussed in the following.

Given a positive number δ, $F_\delta(z)$ measures the mismatch between the given quantiles and those from the marginal inner chain posterior $p(\beta\,|\,z)$. It is defined in the present context, with the chosen quantile statistics, as

$$F_\delta(z) \propto \prod_{i=1}^{2} f_\delta\left([p_i(z), 1-p_i(z)-q_i(z), q_i(z)]\,\middle|\,[0.05, 0.90, 0.05]\right), \tag{16}$$

where

$$p_1(z) = \int_{\ln A<(\ln A)_5} p(\beta\,|\,z)\,d\beta, \qquad q_1(z) = \int_{\ln A>(\ln A)_{95}} p(\beta\,|\,z)\,d\beta, \tag{17}$$

$$p_2(z) = \int_{\ln E<(\ln E)_5} p(\beta\,|\,z)\,d\beta, \qquad q_2(z) = \int_{\ln E>(\ln E)_{95}} p(\beta\,|\,z)\,d\beta, \tag{18}$$

and the trinomial density with probabilities $p+r+q=1$ is noted $[p,r,q]$. Further, f_δ is defined as [41]:

$$f_\delta([p,r,q]\,|\,[0.05,0.90,0.05]) = \exp\left\{-\delta\left(p\ln\frac{p}{0.05} + r\ln\frac{r}{0.90} + q\ln\frac{q}{0.05}\right)\right\}, \tag{19}$$

being the Kullback-Leibler (KL) density [41] formed of the KL divergence from the density $[p,r,q]$ to $[0.05,0.90,0.05]$. This density estimates the average probability that samples from one density are consistent with another. It has a peak of 1 at the desired $p=q=0.05$, with a rate of decay toward zero that increases with increasing $\delta>0$. Thus, δ determines the width of the overall density $F_\delta(z)$, and therefore the tolerance for deviations from the desired values of the quantiles on $(\ln A, \ln E)$. With large δ, the bounds consistency check is tight, tending to strongly reject proposed data sets whose resulting quantiles differ slightly from the requirement. Effectively, δ also determines the relative weight given to the bounds check versus the nominal parameter value check in the likelihood function.

With the likelihood functions of both inner and outer chains defined, we now lay out the overall structure of the algorithm. The outer chain employs a single-site MCMC algorithm, Algorithm 1. The choice of a single-site algorithm is motivated by its relative ease of tuning, particularly since the chain is sampling a data manifold in a high dimensional space. The algorithm employs K_d steps. In each step, it explores each of the $2N+1$ directions sequentially,

Algorithm 1: Outer Chain Single Site MCMC algorithm

Input: $\zeta^0 = (z^0, \ln\sigma_d^0) \in \mathrm{R}^{2N+1}$, $p^0 = p(\zeta^0|I)$, $s^{(2N+1)\times 1}$
Output: MCMC chain samples
foreach $k = 1,\ldots,K_d$ **do**
  $\zeta^k \leftarrow \zeta^{k-1}$
  $p^k \leftarrow p^{k-1}$
  foreach $i = 1,\ldots,2N+1$ **do**
    $\delta\zeta_i \leftarrow \xi \sim N(0, s_i)$
    $\zeta_i^* \leftarrow \zeta_i^k + \delta\zeta_i$
    $p^* = p(\zeta^*|I)$
    $\alpha = \min\left(1, \frac{p^*}{p^k}\right)$
    if $u \sim U(0,1) < \alpha$ **then**
      $\zeta_i^k \leftarrow \zeta_i^*$
      $p^k \leftarrow p^*$
    end
  end
end

employing the conventional Metropolis-Hastings (MH) acceptance test to update the state in each dimension. The un-normalized posterior density evaluated for every candidate ζ^*, is

$$p(\zeta^*|I) = p(I|\zeta^*)\pi(\zeta^*), \tag{20}$$

employing the data likelihood function $p(I|\zeta)$ in Eq. (14) above, and the prior on ζ, $\pi(\zeta)$, further specified below. The data likelihood function evalution requires the above-described statistics on the parameter posterior, employing the inner MCMC chain on the parameter space. The inner chain is outlined in Algorithm 2. It employs K_p steps. Here we use the AMCMC construction referred to above, on the M dimensional parameter space ($M = 3$). In each step, a jump $\delta\lambda$ in M-dimensions is explored, followed by the usual MH acceptance test. The un-normalized posterior density evaluated for every candidate λ^*, is

$$p(\lambda^*|z) = p(z|\lambda^*)\pi(\lambda^*), \tag{21}$$

employing the parameter likelihood function $p(z|\lambda)$ in Eq. (13), and the prior on λ, $\pi(\lambda)$, specified below. The evaluation of $\delta\lambda$ follows [44, 45], and is outlined in Algorithm 3. The procedure requires an initial guess at a covariance matrix for the proposal distribution, C_{init}, which it uses up to step K_{start}, while accumulating data to estimate the actual local covariance matrix Θ of the posterior density. For chain steps in the range $[K_{\text{start}}, K_{\text{stop}}]$, and at chosen intervals q, an updated covariance matrix C for the mutlivariate normal proposal density is evaluated, as shown. The scaling by ρ follows [44, 45] and requires a specification of a scale factor κ, as shown. The tolerance ϵ is a guard against possible ill-conditioning of the Cholesky decomposition, used to arrive at a lower triangular matrix L, where $LL^T = C$. Finally, $\delta\lambda$ is evaluated by scaling an *i.i.d.* normal vector ξ by the matrix L.

With the outer chain execution completed, we have K_d consistent data sets, which are pooled to arrive at the requisite pooled posterior. Employing logarithmic pooling [41], this is done by running the inner chain inference one more time, with the data set $Z = \{z^1, \ldots, z^{K_d}\}$, and with the likelihood function

$$p(Z\,|\,\lambda) = \left[\prod_{k=1}^{K_d} p(z^k\,|\,\lambda)\right]^{1/K_d}, \tag{22}$$

where $p(z^k|\lambda) \equiv p(z^k|\beta, \ln\sigma)$ is given above in Eq. (13).

The nested MCMC chain construction results in a computationally expensive algorithm. On the other hand, data chains can be run in parallel and the resulting chain statistics combined. This introduces significant computational savings.

In the following, we apply the above procedure to the chemical system at hand.

Algorithm 2: Inner Chain MCMC algorithm

Input: $\lambda^0 = (\beta_0, \ln\sigma_d) \in \mathrm{R}^M$; z; $p^0 = p(\lambda^0|z)$
Output: MCMC chain samples
foreach $k = 1, \ldots, K_p$ **do**
  $\delta\lambda \leftarrow$ **EvalStep**(λ^{k-1})
  $\lambda^* \leftarrow \lambda^{k-1} + \delta\lambda$
  $p^* = p(\lambda^*|z)$
  $\alpha = \min\left(1, \dfrac{p^*}{p^{k-1}}\right)$
  if $u \sim U(0,1) < \alpha$ **then**
    $\lambda^k \leftarrow \lambda^*$
    $p^k \leftarrow p^*$
  else
    $\lambda^k \leftarrow \lambda^{k-1}$
    $p^k \leftarrow p^{k-1}$
  end
end

Algorithm 3: Algorithm **EvalStep(λ^{k-1}).**

Input: $C_{\text{init}} \in \mathrm{R}^{M\times M}$; λ^{k-1}; M; $\rho = 2.4^2\kappa/M$; $\kappa = 1$; $\epsilon = 10^{-8}$; $K_{\text{stop}} = 10^8$; $K_{\text{start}} = 10^3$; $q = 10$
Output: $\delta\lambda$
if $k = 1$ **then**
  $\mu^k = \lambda^{k-1}$; $\Theta^k = 0^{M\times M}$; $C = C_{\text{init}}$
  Cholesky Decomposition: Eval lower triangular matrix L s.t. $LL^T = C$
else
  $r = \dfrac{k-2}{k-1}$; $s = \dfrac{1}{k-1}$
  if $k < K_{\text{stop}}$ **then**
    $\mu^k \leftarrow \left(\mu^{k-1}(k-1) + \lambda^{k-1}\right)/k$
    foreach $i = 1,\ldots,M;\ j = 1,\ldots,M$ **do**
      $\Theta^k_{ij} \leftarrow r\Theta^{k-1}_{ij} + s\left((k-1)\mu^{k-1}_i\mu^{k-1}_j - k\mu^k_i\mu^k_j + \lambda^{k-1}_i\lambda^{k-1}_j\right)$
    end
    if $k > K_{\text{start}}$ **and** $k \bmod q = 0$ **then**
      $C = \rho(\Theta^k + \epsilon\mathbf{1})$
      Cholesky Decomposition: Eval lower triangular matrix L s.t. $LL^T = C$
    end
  end
end
$\delta\lambda \leftarrow L\xi$, with $\xi^{M\times 1} \sim N(0,\mathbf{I})$

4. APPLICATION TO THE CHEMICAL SYSTEM

To begin with, given the computational cost involved in the double-chain structure of the algorithm, we employ a Polynomial Chaos (PC) [8, 19, 46] surrogate [17] for $\ln\tau$, with the ignition delay time τ as defined above, over the range of interest of the parameters $(\ln A, \ln E)$ and initial temperature T^o. This involves, most conveniently, and in order to enforce uniform accuracy of the surrogate over the input $(\ln A, \ln E, T^o)$ range, defining *i.i.d.* uniform distributions for these inputs on their chosen ranges, and propagating this specified uncertainty forward through the model to provide the corresponding PC representation of $\ln\tau(\ln A, \ln E, T^o)$. Note that the choice of these uniform distributions is simply a means for constructing a PC representation, and hence a polynomial response surface surrogate, for $\ln\tau$ over a range of variation of the selected inputs, and does not relate to "actual" uncertainties on inputs/outputs reflecting our knowledge. The input ranges are defined by estimating the expected range of exploration of each input $\ln A$, $\ln E$, and T^o, by the DFI procedure. In turn, this is informed by the above given initial temperature range and summary statistics on $(\ln A, \ln E)$ presumed reported, which also inform the priors for the inner chain. Of course if, in the course of using the surrogate model, the MCMC chains explore ranges of the parameters *outside* the range over which the surrogate is built, one has the option of rebuilding the surrogate over a larger input range or simply using the full model for the offending samples. Finally, we used a sixth order Legendre-Uniform PC [47] surrogate over the input 3D space, built using full tensor product quadrature [18, 48, 49]. The surrogate was found to have a relative rms error with respect to the actual model of 0.06%, as evaluated on a uniform test mesh distinct from the quadrature points. The local absolute and relative errors in $\ln\tau$ are plotted versus the mesh point index, for the 20^3 test mesh points in the 3D space ($\ln A$, $\ln E$, T^o), in Fig. 6.

Given that the number of data points is not known, we presume $N = 8$ data points for the data chain arbitrarily. This choice can be made probabilistically too. In particular, one can declare a distribution of possible values of N based on prior knowledge, e.g., $N \sim U[5, 50]$. Then, the DFI answer can be marginalized over this PDF. While this is computationally expensive, it is nonetheless straightforward. In the present context, while we did not marginalize over N, we did explore the effect of $N = 4, 8, 16$ on the resulting pooled posterior. These results, not shown, indicate that the pooled posterior $p(\ln A, \ln E, \ln\sigma|Z)$ exhibits smaller volume, i.e., lower uncertainty, with increasing N. However,

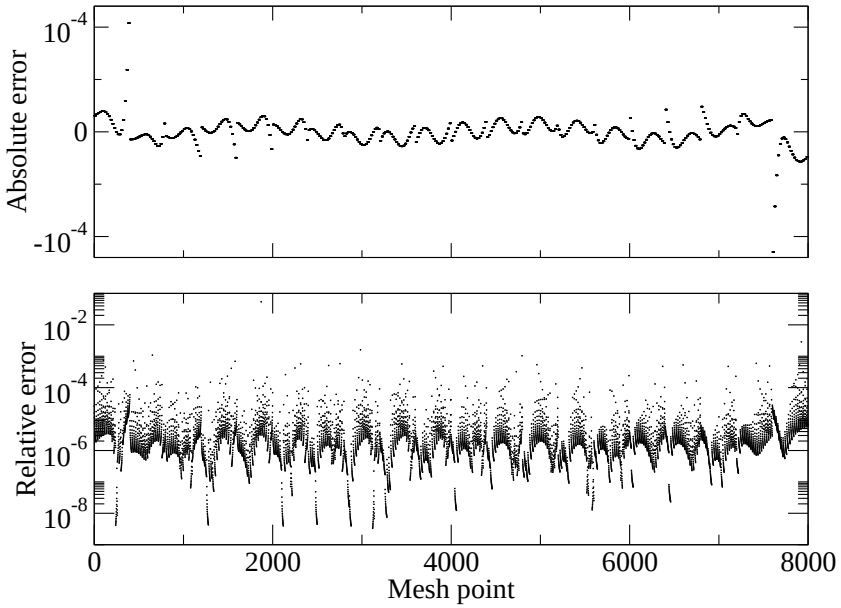

FIG. 6: Local absolute and relative surrogate errors for the 20^3 test mesh points.

this effect is manifested primarily through the nuisance parameter $\ln\sigma$. We find negligible difference in the marginal pooled posterior $p(\ln A, \ln E|Z)$ for the three N values, while the 1D marginal pooled posterior $p(\ln\sigma|Z)$ is narrower and peaked at lower $\ln\sigma$ values for higher N. The negligible impact on $p(\ln A, \ln E|Z)$ suggests that, in the present problem, the given information is sufficient to constrain the marginal pooled DFI posterior on $(\ln A, \ln E)$, irrespective of the presumed N.

For the outer chain, we presume independent uniform priors on $T^o \in [1000, 1300]$, and $\ln\sigma_d \in [-7, 3]$ and a strictly positive improper uniform prior on $\tau > 0$. The prior bounds on the ignition temperature are based on the given information on the range of the experimental conditions. The lower bound on the ignition time is of course physically motivated. As for those on $\ln\sigma_d$, they are subjectively chosen, and are in the present context inconsequential as the chain never reaches them. They reflect the degree of belief on the maximal and minimal possible noise levels in the data. The likelihood function, which, as stated above, involves the density $w_\delta(z, \ln\sigma_d)$ in Eq. (14), employs $\delta = 100$. The chain starting point is chosen at $\ln\sigma_d^0 = -2$, $T_i^o = T_{\min}^o + ih_1$, and $\tau_i = \tau^m(T_i^o, \beta_0)(1 + \sigma_d^0\epsilon_i)$ where $h_1 = (T_{\max}^o - T_{\min}^o)/(N+1)$, for $i = 1, \ldots, N$. The proposal distribution in each dimension is a zero-mean Gaussian with a specified variance.

As for the inner chain, with the state vector $(\beta, \ln\sigma)$, we use *i.i.d.* uniform priors for all variables. We conservatively set the lower and upper bounds on each uniform prior at $\beta_- = \beta_0 - 5(\beta_0 - \beta_5)$, and $\beta_+ = \beta_0 + 5(\beta_{95} - \beta_0)$. Further, consistent with the outer chain, we set $\ln\sigma \sim U[-7, 3]$. The inner chain starting point is at $(\beta_0, \ln\sigma_d)$, and we employ $C_{\text{init}} = \text{diag}[10^{-3}, 2.8 \times 10^{-6}, 4 \times 10^{-4}]$.

A sample of resulting data from a 5000-long outer/data chain is plotted in Fig. 7, where each step in the chain is defined by the elements of the 17-dimensional data vector defined above. We note that there is no expectation of "good mixing" here, as the data chain simply provides a means of generating a random walk on a manifold in the data space that is defined/constrained implicitly by the given information [41]. The key requirement is to provide sufficient coverage of the data space, which we will examine further below employing data posterior statistics. The corresponding sampled data are well clustered around the nominal model, and the original data, as can be seen in Fig. 8. Clearly, only data sets clustered within the neighborhood defined by the missing data are found by the algorithm to be consistent with the given information, and are therefore accepted.

Beyond this short data chain sample, a sufficiently large number of data samples is needed to ensure convergence of the algorithm. However, this can be done in parallel, with multiple data chains. Accordingly, we ran 50 data chains

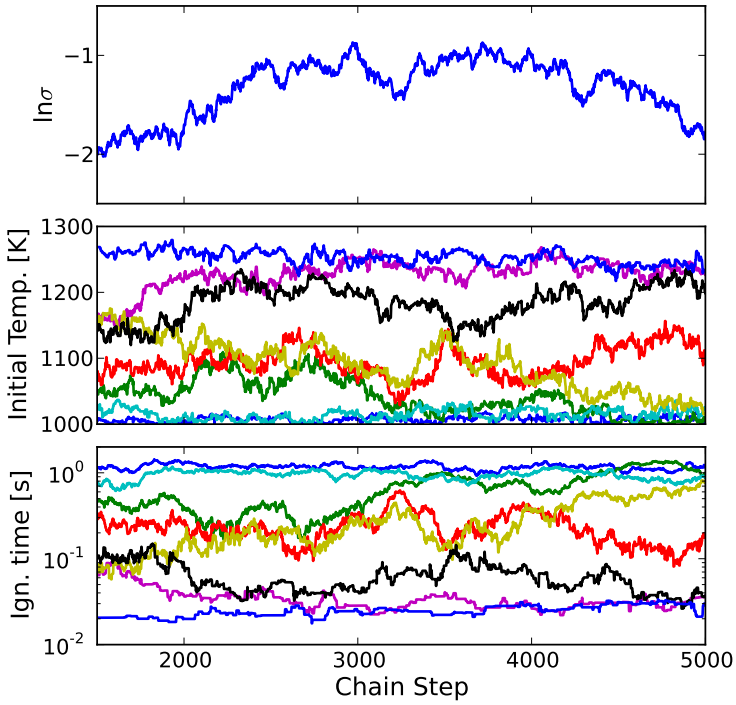

FIG. 7: A segment of the DFI data chain, showing the MCMC sampled values of the data vector. In the bottom two frames, each line corresponds to the sampled value of each data point, (τ_i, T_i^o), $i = 1, \ldots, N$, with $N = 8$.

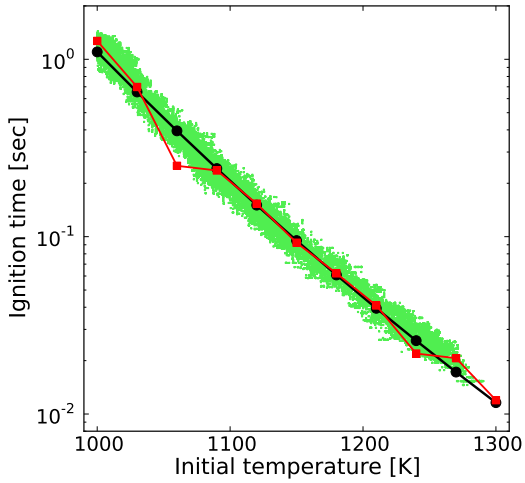

FIG. 8: A plot of the DFI data chain samples from Fig. 7, superposed on the original model predictions and data.

in parallel, each 5000-long, and used unions of the data sets, to arrive at the pooled posteriors. Since there is no expectation of good mixing, there is no associated definition of a burn-in length for any given chain. We do define such a burn-in period in a statistical sense, however. A scatter plot of the posterior probability values on the sampled data sets from the 50 chains is shown in Fig. 9. The plot exhibits a small upward average trend in the first 500 steps before "typical" behavior is observed in a statistical sense. Conservatively, we define a burn-in period of 1500 steps, thereby neglecting the first 1500 samples from each data chain. Again, we note, as shown in Fig. 10, that the 50-chain

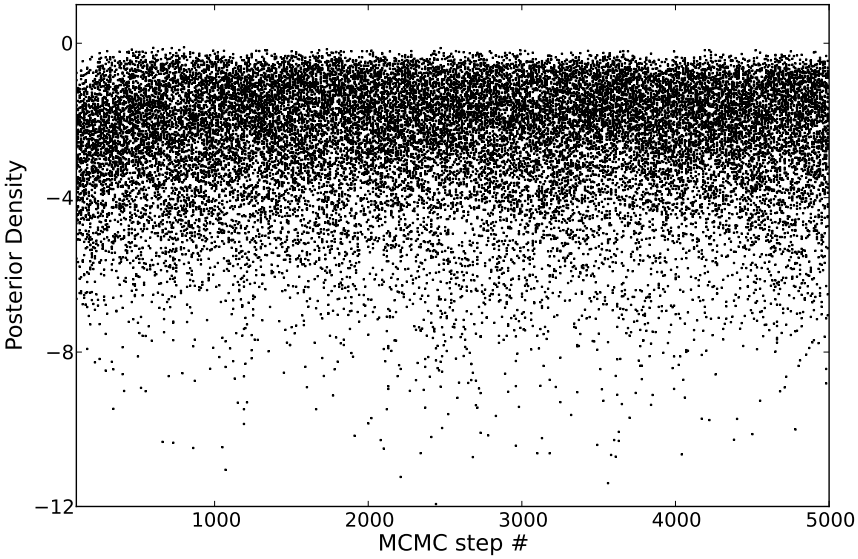

FIG. 9: Plot of the data posterior density samples from all 50 chains superposed. Every fifth chain step is plotted for convenience.

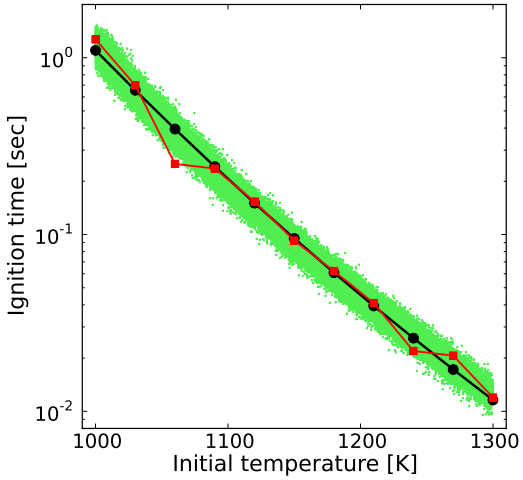

FIG. 10: Scatter plot of 50 chains, with the first 1500 burn-in states removed from each chain. Only every 25th chain state is plotted for convenience.

data are clustered around the nominal model. The data shown in the figure exclude 1500 samples from the beginning of each chain.

Further, we illustrate in Fig. 11 the means as well as the 5% and 95% quantiles extracted from the inner/parameter posteriors evaluated at each data chain step, for each of the 50 chains, shown for the last 500 steps of each chain. The full length of the chains exhibits a similar statistical distribution of the means and quantiles; this is illustrated here with this short segment for clarity. The figure shows the general agreement of the summary statistics extracted from each data chain step parameter posterior with those of the reference posterior. Following [41], the scatter of the 5/95% quantiles can be reduced by increasing δ, while that of the nominal can be controlled by raising $p(z, \ln\sigma_d|\beta_0)/\max_{\beta,\sigma}[p(z, \ln\sigma|\beta)]$ to a suitable power greater than unity, in Eq. (14).

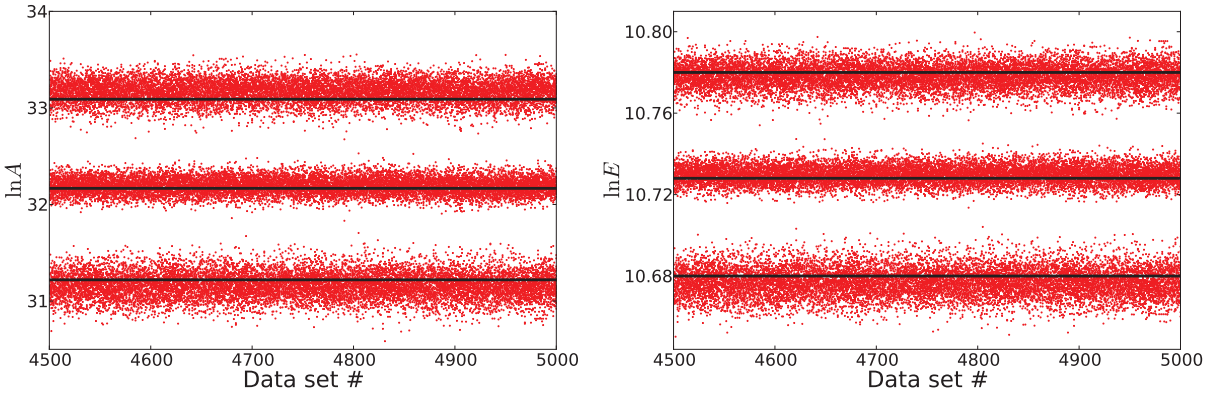

FIG. 11: Plot of the means, as well as the 5% and 95% quantiles from the parameter posteriors from each data chain step, for all 50 chains, and for the last 500 steps from each chain. The left frame shows those for $\ln A$, while the right frame shows those for $\ln E$. The solid lines indicate the corresponding values from the reference posterior.

The pooled parameter posteriors show similarly good agreement with the reference posterior. In fact, individual marginal $(\ln A, \ln E)$ posteriors, from each data chain step, exhibit a structure very much in alignment with the reference density, albeit with a scatter in their summary statistics as seen in Fig. 11. We illustrate in Fig. 12 the 2D

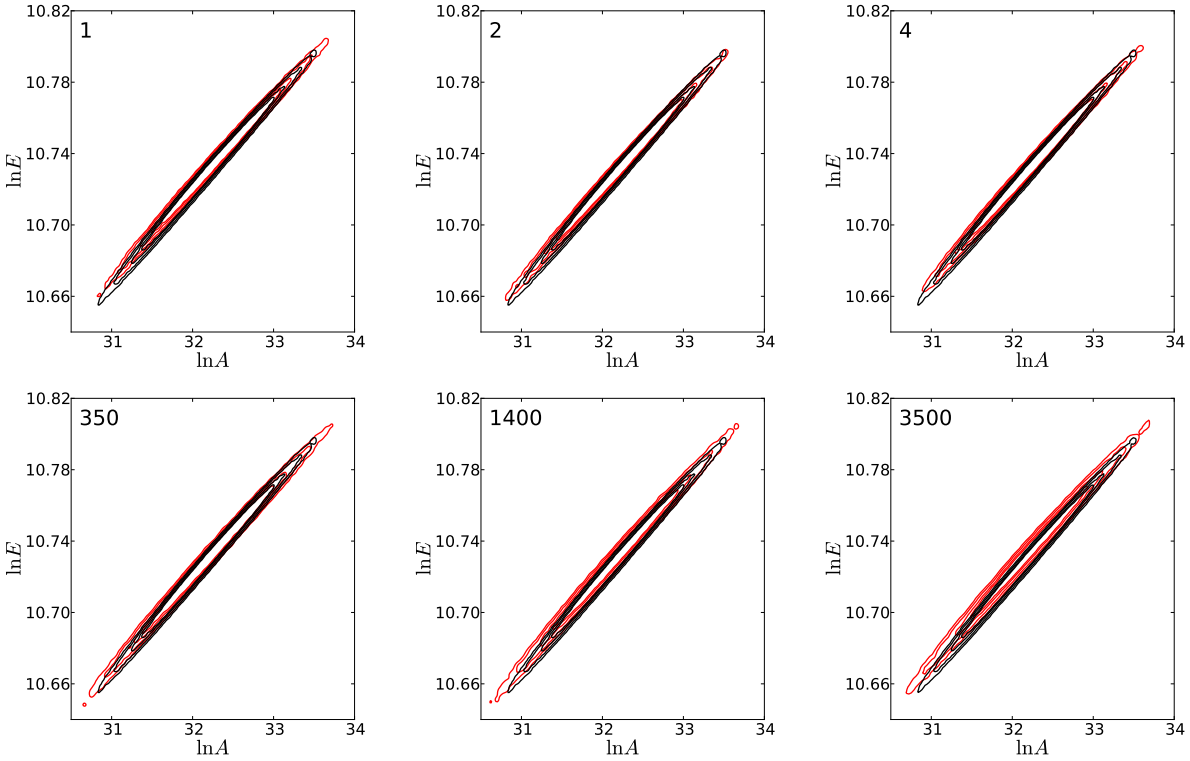

FIG. 12: Plot of the pooled 2D marginal $(\ln A, \ln E)$ posteriors computed from a range of number of data steps, from 1 to 3500, as indicated in each frame, from a single data chain. The pooled posteriors (red) are superposed over the reference posterior (black). Posterior densities are estimated from posterior samples using KDE.

marginalized pooled distributions for an increasing data-chain segment length, from 1 to 3500 steps, each superposed on top of the reference posterior. There is little change in the marginal pooled density with increasing data size, beyond a small change in slope of the major axis of the PDF ellipse. The data from all 3500×50 retained data sets in all 50 data chains lead to a similar marginalized pooled density, shown in Fig. 13. Clearly, in this case, the constraints provided by the given information are sufficiently strong, such that very little change in the 2D marginal parameter posterior is evident beyond a single data set. In other words, the step-by-step variation of the posterior, evident in Fig. 11, exhibits sufficient stationarity, such that the essential global structure of the 2D pooled posterior density is relatively stable, even with a single data set. Nonetheless, there is a discernible slight change in the slope of the distribution in going from 1 to 50 chains, evident in comparing Figs. 12 and 13.

Similar robustness can be seen in the 1D marginals, albeit now with discernible scatter, which can be more easily seen in line plots than in contour plots. Figure 14 shows the 1D marginals for a range of data sample set sizes from a single data chain. The plots show small differences among the various shown posteriors, and between each of them and the reference density. Similarly, the relatively small scatter of the 1D marginal posteriors from each of the 50 chains, each involving 3500 pooled posteriors, is shown in Fig. 15. Further, the 1D marginal pooled distributions resulting from pooling successively added numbers of chains, up to the full set of 50 chains, are shown in Fig. 16. There is little evident change in the marginal pooled posterior among the different lengths of aggregated chains. Moreover, as with the shorter length pooled posteriors, there are only minor differences here between any of the marginal pooled posteriors and the reference.

Note, nonetheless, that the 2D marginal pooled posterior on $(\ln A, \ln E)$ does not converge to the 2D marginal reference posterior (Fig. 13). This is not surprising, as we only have partial/summary information on the latter. What is remarkable, in fact, is the extent to which the pooled posteriors, even with small data volumes, do approximate the missing posterior. This is a testament to the nearly complete information available in the summary statistics, and constrained by the other given information, in the present context.

In order to more quantitatively examine the self-convergence of the pooled posterior with increasing data volume, we examine the Kullback-Leibler (KL) divergence between pooled posteriors corresponding to successively larger data volumes, as shown in Fig. 17. Specifically, defining the pooled posterior with k chains as $p_k(\ln A, \ln E, \ln\sigma)$, and the KL divergence between p_k and $p_{k/2}$ as D_k, given by

$$D_k = D_{KL}(p_{k/2}||p_k) \equiv \int p_{k/2} \ln \frac{p_{k/2}}{p_k} \, \mathrm{d}\ln A \, \mathrm{d}\ln E \, \mathrm{d}\ln\sigma, \tag{23}$$

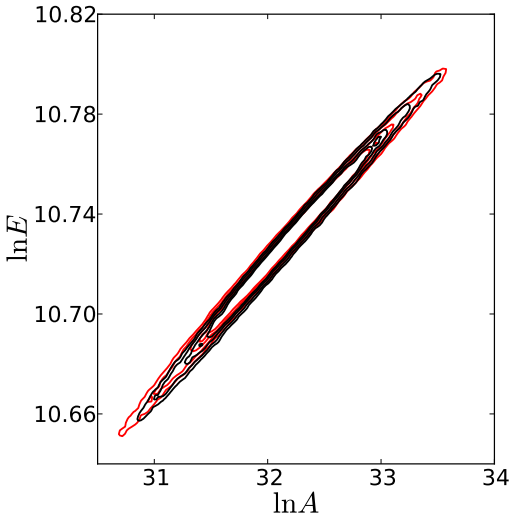

FIG. 13: Two-dimensional marginal parameter posteriors on $(\ln A, \ln E)$ resulting from pooling 50 chains of data (red), compared with the reference known posterior (black).

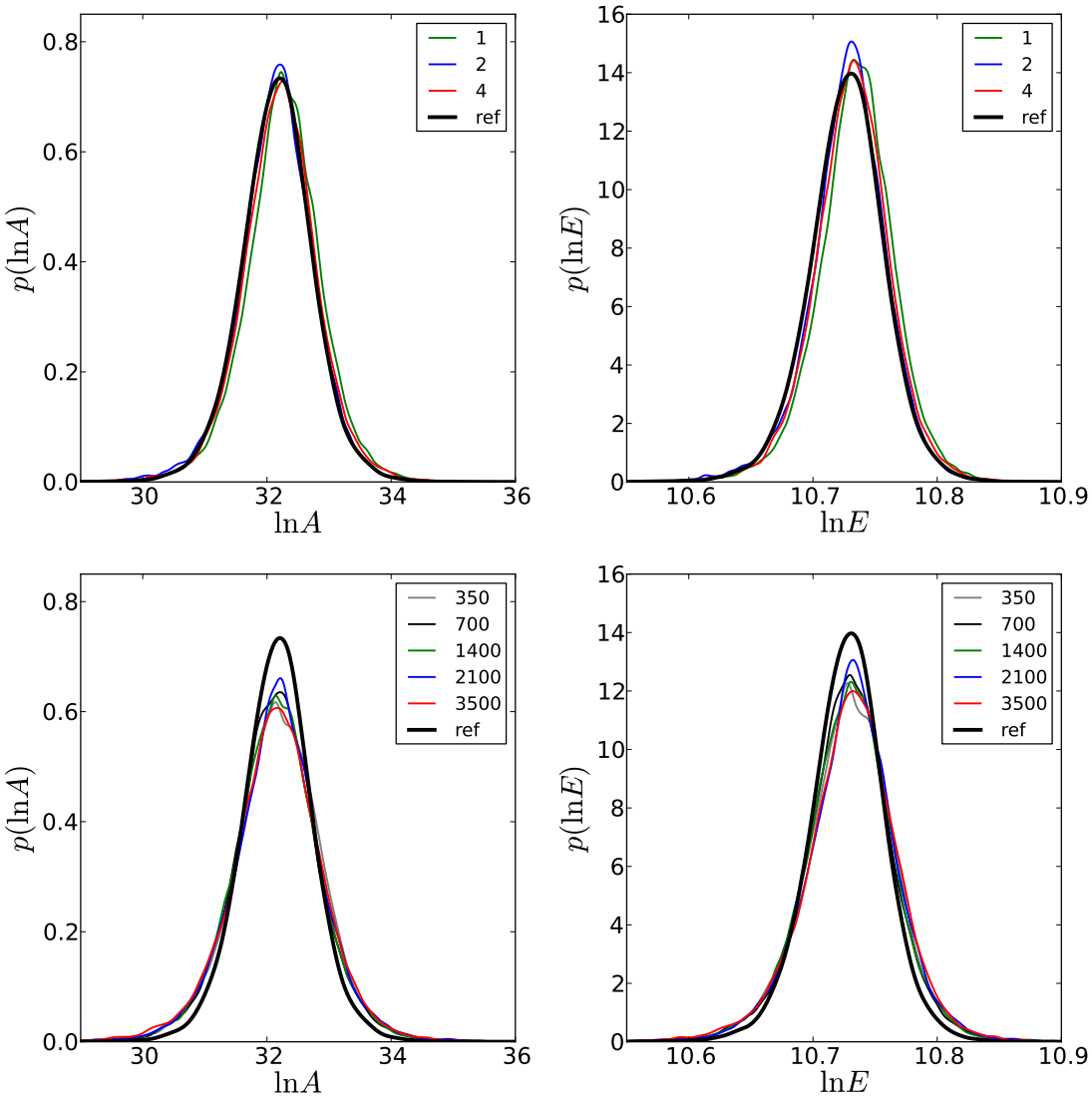

FIG. 14: Plot showing the pooled 1D marginals for $\ln A$ (left) and $\ln E$ (right) for a range of data sample set sizes from a single data chain.

we show in the figure the evolution of D_k for $k = 2, 4, 8, 16, 32$. Given the $\binom{50}{k}$ ways of combining k chains out of 50 without regard to order, the plot shows the scatter of D_k for 1024 randomly sampled permutations of the 50 chains[2] for each k, as well as the associated mean value. This is done in order to examine statistical convergence trends. The results clearly indicate the convergence of the mean with increased number of chains. Moreover, the slope of the mean line indicates roughly inverse linear dependence on data set size, i.e., $\overline{D}_k \propto k^{\gamma}$ with $\gamma \approx -1$. The empirically observed γ value shows some dependence on the particular KL divergence choice, namely as above or as $D_k^* = D_{KL}(p_k||p_{k/2})$, which is not surprising given the finite number of chains and the asymmetry of the KL divergence. The observed values, with

[2]The random samples are generated and used as follows. We sample 1024 random integer vectors $(c_1, c_2, \ldots, c_{50})$ using the "entry-by-entry brute force" method [50]. Then, for each sample, we evaluate p_k from the first k chains $c_1, \ldots, c_k$, for each k.

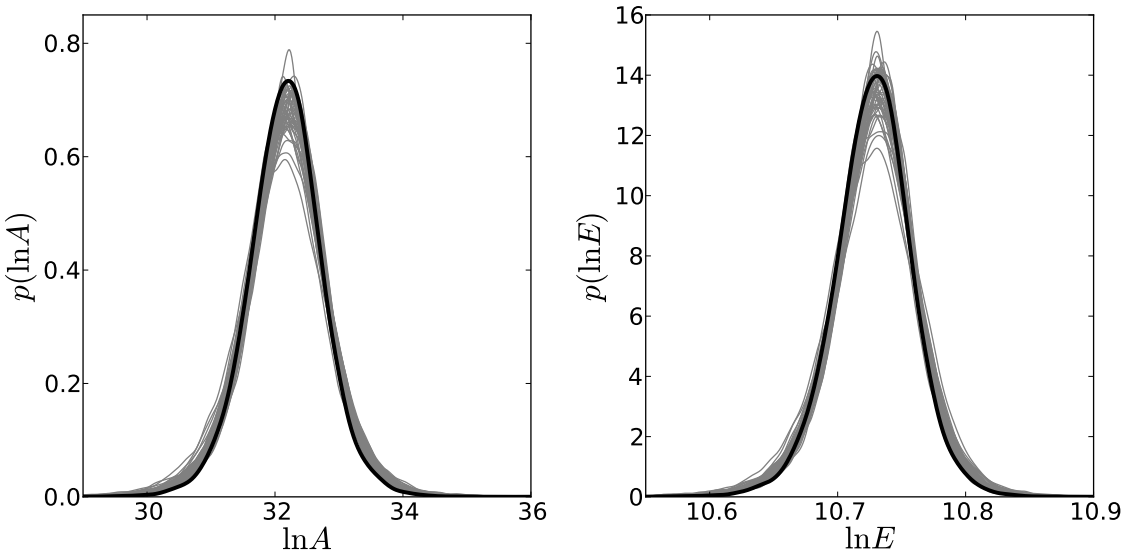

FIG. 15: Plot showing the scatter of the pooled 1D marginal $\ln A$ (left) and $\ln E$ (right) posteriors for each of the 50 chains (gray), based on 3500 data sets from each data chain, superposed on the reference posteriors (black).

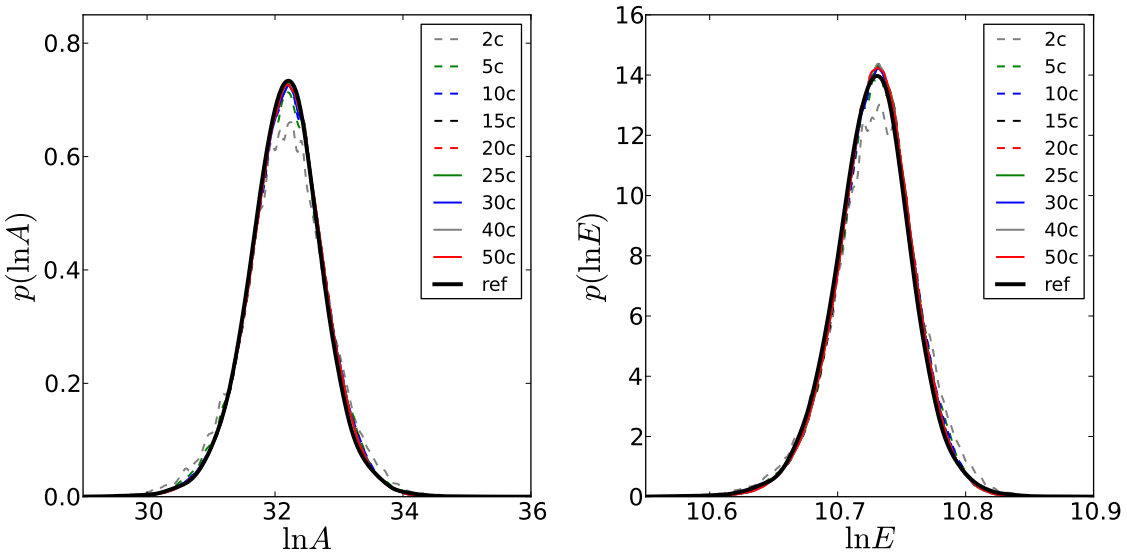

FIG. 16: One-dimensional marginal parameter posteriors on $\ln A$ (left) and $\ln E$ (right) resulting from pooling different volumes of data, compared with the reference known posterior. We show the posteriors resulting from pooling of 2 chains (2c), 5 chains (5c), etc. There is very little change in the pooled marginal posterior past the first five chains.

$$2^{\gamma} = \frac{\overline{D}_k - \overline{D}_{k/2}}{\overline{D}_{k/2} - \overline{D}_{k/4}} \tag{24}$$

and γ^* defined similarly in terms of $\overline{D}^*$, are shown in Table 1. The exponent is close to -1 in both cases. Note that these values were found to change by only $\pm$ a few percent in going from 512 to 1024 samples, and are therefore of corresponding accuracy.

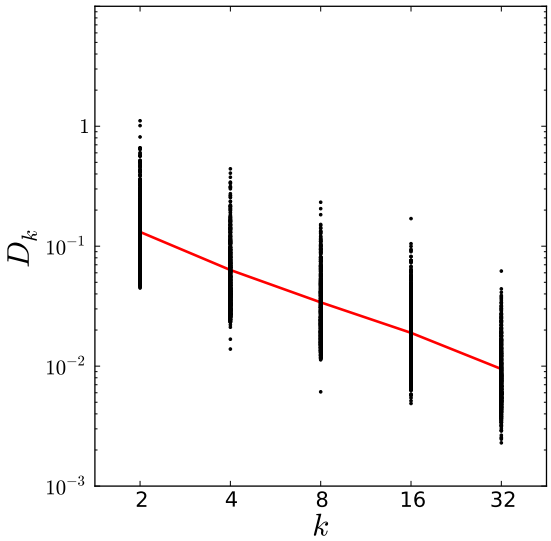

FIG. 17: Plot showing the statistical convergence of the algorithm in terms of the Kulback-Leibler divergence (KLdiv) between successive pooled posteriors. Also shown is the scatter in each of the KLdiv values. The red line indicates the evolution of the mean with increased number of chains.

TABLE 1: Observed values of the decay rate of the KL divergence between pooled posteriors with increasing data volume. The left column is the number of chains, while the second and third columns are the observed decay rates for the two alternate KL divergence definitions

k	γ	γ^*
4	–1.56	–1.05
8	–1.09	–0.90
16	–0.95	–0.84
32	–1.12	–1.00

5. CONCLUSIONS

We have explored the use of our recently developed DFI procedure for the estimation of input uncertainties for a chemical model for methane-air ignition. The procedure provides a means for estimation of a joint posterior on model parameters that is consistent with available information, in the absence of data that had been previously used to arrive at some/all of this information. This joint posterior provides a more accurate/consistent characterization of parametric uncertainty than the commonly used alternative employing independent parameters.

We set up the problem by first creating a hypothetical noisy data set, and using it to estimate a posterior distribution on model parameters with Bayesian inference. We subsequently discard the data and the posterior while retaining summary statistics on the latter. Then, we use DFI to find a posterior density that is consistent with these summary statistics and other known information.

The method was applied in the context of methane-air ignition with a simple global single-step irreversible Arrhenius kinetic model. Ignition time, over a range of input temperature, was the model observable of interest. Model parameters of interest were the Arrhenius rate expression pre-exponential and activation energy.

We detailed the specification of the nested pair of MCMC chains involved in the procedure with the requisite likelihood functions for the problem at hand. We examined the performance of the algorithm for a range of data volumes, highlighting the characteristics of the data chain samples and the consistency of the accepted parametric

posteriors. We used a number of parallel data chains in order to accelerate the scheme. We examined the resulting pooled posteriors in terms of their self-similarity and in comparison to the reference posterior, with increasing numbers of data chains. We also measured the convergence rate of the KL divergence between pooled DFI posteriors obtained with increasing data volumes.

In this chemistry problem, likely due to the strong nonlinearity of exothermal ignition, we find very small variation of the marginal pooled DFI posterior after a very short number of data samples. Similarly, we find negligible impact on the marginal pooled DFI posterior on $(\ln A, \ln E)$ with a $2\times$ up/down variation in the presumed number of data points. This overall robustness suggests that acceptable data sets are strongly constrained by the given information. In other words the available summaries are essentially sufficient statistics. We note that, while the accepted posteriors at each data chain step exhibit some variability in their means and marginal quantiles, their overall structure and slope in the 2D parameter plane was found to be quite robust. Finally, we find that the mean KL divergence between pooled posteriors decays nearly linearly with the number of data samples on which they are based.

Moreover, the marginal pooled DFI posterior is found to be very close to the missing reference posterior. While this is generally desirable, it is not necessarily expected, unless sufficient information based on the missing data is in fact available. In a context where less information is available, e.g., say, conditional/marginal bounds are reported on only one of the two parameters, or nothing is known about the nature of the instrument noise structure, it is expected that the pooled posterior may well differ significantly from the reference posterior. This is acceptable in principle, as the object of the algorithm is *not* to discover the missing posterior (a potentially impossible feat in the absence of the original data), but rather to find a posterior that is consistent with the given information.

We will explore in future work the consequences of more limited information in this context. We will also explore the utility of the methodology in the estimation of joint posteriors among parameters in more realistic chemical models involving multiple elementary reversible reaction steps. In this setting, parameters of reaction rate expressions for multiple reaction steps can be correlated by the experimental fitting procedures, where previously measured parameters of one reaction are used in the fitting employed to estimate parameters of the rate expression of one or more other reactions. Accordingly, a dependence tree, to be built based on examination of experimental procedures, is expected to influence the correlation structure of the joint posterior on the full set of uncertain model parameters.

ACKNOWLEDGMENTS

This work was supported by the US Department of Energy (DOE), Office of Basic Energy Sciences (BES) Division of Chemical Sciences, Geosciences, and Biosciences; by the DOE Office of Advanced Scientific Computing Research (ASCR), under the Scientific Discovery through Advanced Computing (SciDAC) program; and by the DOE ASCR Applied Mathematics program via the 2009 American Recovery and Reinvestment Act. Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the US Department of Energy's National Nuclear Security Administration under contract DE-AC04-94-AL85000.

REFERENCES

1. Kee, R., Grcar, J., Smooke, M., and Miller, J., A Fortran Program for Modeling Steady Laminar One-Dimensional Premixed Flames, Sandia Report SAND85-8240, Sandia National Labs., Livermore, CA, September 1993.
2. Droege, A. and Tully, F., Hydrogen-atom abstraction from alkanes by OH. 3. propane, *J. Phys. Chem.*, 90(9):1949–1954, 1986.
3. Sivaramakrishnan, R., Srinivasan, N., Su, M.-C., and Michael, J., High temperature rate constants for OH + alkanes, *Proc. Comb. Inst.*, 32(1):107–114, 2009.
4. Curran, H., Gaffuri, P., Pitz, W., and Westbrook, C., A comprehensive modeling study of n-heptane oxidation, *Combust. Flame*, 114:149–177, 1998.
5. Curran, H., Gaffuri, P., Pitz, W., and Westbrook, C., A comprehensive modeling study of iso-octane oxidation, *Combust. Flame*, 129:253–280, 2002.

6. Green, W., Allen, J., Ashcraft, R., Beran, G., Class, C., Gao, C., Goldsmith, C., Harper, M., Jalan, A., Magoon, G., Matheu, D., Merchant, S., Mo, J., Petway, S., Raman, S., Sharma, S., Song, J., Geem, K. V., Wen, J., West, R., Wong, A., Wong, H.-W., Yelvington, P., and Yu, J., Rmg—reaction mechanism generator v3.3, 12 April 2014, http://rmg.sourceforge.net.
7. Moréac, G., Blurock, E., and Mauss, F., Automatic generation of a detailed mechanism for the oxidation of *n*-decane, *Combust. Sci. Technol.*, 178:2025–2038, 2006.
8. Ghanem, R. and Spanos, P., *Stochastic Finite Elements: A Spectral Approach*, Springer Verlag, New York, 1991.
9. Ghanem, R., Hybrid stochastic finite elements and generalized Monte Carlo simulation, *ASME J. Appl. Mech.*, 65:1004–1009, 1998.
10. Le Maître, O., Knio, O., Najm, H., and Ghanem, R., A stochastic projection method for fluid flow I. Basic formulation, *J. Comput. Phys.*, 173:481–511, 2001.
11. Tartakovsky, D., Lichtner, P., and Pawar, R., PDF methods for reactive transport in porous media, *Acta Universitatis Carolinae Geologica*, 46(2-3):113–116, 2002.
12. Xiu, D. and Karniadakis, G., Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos, *Comput. Methods Appl. Mech. Eng.*, 191:4927–4948, 2002.
13. Le Maître, O., Ghanem, R., Knio, O., and Najm, H., Uncertainty propagation using Wiener-Haar expansions, *J. Comput. Phys.*, 197(1):28–57, 2004.
14. Soize, C. and Ghanem, R., Physical systems with random uncertainties: Chaos representations with arbitrary probability measure, *SIAM J. Sci. Comput.*, 26(2):395–410, 2004.
15. Wan, X. and Karniadakis, G. E., An adaptive multi-element generalized polynomial chaos method for stochastic differential equations, *J. Comput. Phys.*, 209:617–642, 2005.
16. Ganapathysubramanian, B. and Zabaras, N., Sparse grid collocation schemes for stochastic natural convection problems, *J. Comput. Phys.*, 225(1):652–685, 2007.
17. Marzouk, Y. M., Najm, H. N., and Rahn, L. A., Stochastic spectral methods for efficient Bayesian solution of inverse probelms, *J. Comput. Phys.*, 224(2):560–586, 2007.
18. Nobile, F., Tempone, R., and Webster, C., A sparse grid stochastic collocation method for partial differential equations with random input data, *SIAM J. Num. Anal.*, 46(5):2309–2345, 2008.
19. Najm, H., Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics, *Ann. Rev. Fluid Mech.*, 41(1):35–52, 2009.
20. Tartakovsky, D. and Broyda, S., PDF equations for advective-reactive transport in heterogeneous porous media with uncertain properties, *J. Contam. Hydrol.*, 120-121:129–140, 2011.
21. Warnatz, J., Resolution of gas phase and surface combustion chemistry into elementary reactions, in *Twenty-Fourth Symposium (International) on Combustion*, Vol. 24, The Combustion Institute, Sydney, Australia, pp. 553–579, 5–10 July, 1992.
22. Walters, R. and Huyse, L., Uncertainty analysis for fluid mechanics with applications, Tech. Rep., ICASE Report No. 2002-1; NASA/CR-2002-211449, February 2002.
23. Putko, M., Newman, P., Green, L., and Taylor III, A., Approach for input uncertainty propagation and robust design in CFD using sensitivity derivatives, *ASME J. Fluids Eng.*, 124:60–69, 2002.
24. Sivia, D., *Data Analysis: A Bayesian Tutorial*, Oxford Science, Oxford, 1996.
25. Jaynes, E., *Probability Theory: The Logic of Science*, G. L. Bretthorst (ed.), Cambridge University Press, Cambridge, UK, 2003.
26. Phenix, B., Dinaro, J., Tatang, M., Tester, J., Howard, J., and McRae, G., Incorporation of parametric uncertainty into complex kinetic mechanisms: Application to hydrogen oxidation in supercritical water, *Combust. Flame*, 112:132–146, 1998.
27. Reagan, M., Najm, H., Ghanem, R., and Knio, O., Uncertainty quantification in reacting flow simulations through non-intrusive spectral projection, *Combust. Flame*, 132:545–555, 2003.
28. Reagan, M., Najm, H., Debusschere, B., Le Maître, O., Knio, O., and Ghanem, R., Spectral stochastic uncertainty quantification in chemical systems, *Combust. Theory Model.*, 8:607–632, 2004.
29. Reagan, M., Najm, H., Pébay, P., Knio, O., and Ghanem, R., Quantifying uncertainty in chemical systems modeling, *Int. J. Chem. Kin.*, 37(6):368–382, 2005.

30. Turányi, T., Zalotai, L., Dóbé, S., and Bérces, T., Effect of uncertainty of kinetic and thermodynamic data on methane flame simulation results, *Phys. Chem. Chem. Phys.*, 4:2568–2578, 2002.
31. Skodje, R., Tomlin, A., Klippenstein, S., Harding, L., and Davis, M., Theoretical validation of chemical kinetic mechanisms: Combustion of methanol, *J. Phys. Chem. A*, 114:8286–8301, 2010.
32. Davis, M., Skodje, R., and Tomlin, A., Global sensitivity analysis of chemical-kinetic reaction mechanisms: Construction and deconstruction of the probability density function, *J. Phys. Chem. A*, 115:1556–1578, 2011.
33. Najm, H., Debusschere, B., Marzouk, Y., Widmer, S., and Le Maître, O., Uncertainty quantification in chemical systems, *Int. J. Numer. Methods Eng.*, 80:789–814, 2009.
34. Baulch, D., Cobos, C., Cox, R., Esser, C., Frank, P., Just, T., Kerr, J., Pilling, M., Troe, J., Walker, R., and Warnatz, J., Evaluated kinetic data for combustion modelling, *J. Phys. Chem. Ref. Data*, 21:411–429, 1992.
35. Baulch, D., Cobos, C., Cox, R., Frank, P., Hayman, G., Just, T., Kerr, J., Murrels, T., Pilling, M., Troe, J., Walker, R., and Warnatz, J., Evaluated kinetic data for combustion modelling. Supplement I, *J. Phys. Chem. Ref. Data*, 23:847–1033, 1994.
36. Rubin, D. and Little, R., *Statistical Analysis with Missing Data*, Wiley, New York, 2002.
37. Allison, P., *Missing Data*, Sage Publications Inc., Thousand Oaks, CA, 2001.
38. Soares, P. and Paulino, C., Incomplete categorial data analysis: A Bayesian perspective, *J. Stat. Comput. Simul.*, 69:157–170, 2001.
39. Gelman, A., King, G., and Liu, C., Not asked and not answered: Multiple imputation for multiple surveys, *J. Am. Stat. Assoc.*, 93:846–857, 1998.
40. Schafer, J., *Analysis of Incomplete Multivariate Data*, Chapman and Hall, New York, 1997.
41. Berry, R., Najm, H., Debusschere, B., Adalsteinsson, H., and Marzouk, Y., Data-free inference of the joint distribution of uncertain model parameters, *J. Comput. Phys.*, 231:2180–2198, 2012.
42. Smith, G., Golden, D., Frenklach, M., Moriarty, N., Eiteneer, B., Goldenberg, M., Bowman, C., Hanson, R., Song, S., Gardiner Jr., W., Lissianski, V., and Zhiwei, Q., GRI mechanism for methane/air, version 3.0, 7/30/99, 12 April 2014, www.me.berkeley.edu/gri_mech/.
43. Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, 1996.
44. Haario, H., Saksman, E., and Tamminen, J., An adaptive Metropolis algorithm, *Bernoulli*, 7:223–242, 2001.
45. Atchade, Y. and Rosenthal, J., On adaptive Markov chain Monte Carlo algorithms, *Bernoulli*, 11:815–828, 2005.
46. Wiener, N., The homogeneous chaos, *Am. J. Math.*, 60:897–936, 1938.
47. Xiu, D. and Karniadakis, G., The Wiener-Askey polynomial chaos for stochastic differential equations, *SIAM J. Sci. Comput.*, 24(2):619–644, 2002.
48. Le Maître, O., Reagan, M., Najm, H., Ghanem, R., and Knio, O., A stochastic projection method for fluid flow II. Random process, *J. Comput. Phys.*, 181:9–44, 2002.
49. Nobile, F., Tempone, R., and Webster, C., An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data, *SIAM J. Num. Anal.*, 46(5):2411–2442, 2008.
50. Random permutation, 12 April 2014, http://en.wikipedia.org/wiki/Random_permutation.