Combustion and Flame 208 (2019) 219–234

Contents lists available at ScienceDirect

Combustion and Flame

journal homepage: www.elsevier.com/locate/combustflame

Accounting for uncertainty in RCCE species selection

Esteban Cisneros-Garibay[a], Carlos Pantano[a,c], Jonathan B. Freund[a,b,*]

[a] Mechanical Science & Engineering, University of Illinois at Urbana–Champaign, Urbana, IL, USA
[b] Aerospace Engineering, University of Illinois at Urbana–Champaign, Urbana, IL, USA
[c] Department of Aerospace and Mechanical Engineering, University of Southern California, Los Angeles, CA, USA

ARTICLE INFO

Article history:
Received 11 March 2019
Revised 17 May 2019
Accepted 26 June 2019

Keywords:
Chemical model reduction
Rate-controlled constrained equilibrium
Uncertainty quantification
Bayesian model selection
Autoignition

ABSTRACT

A framework is presented to quantify, based on Bayesian evidence, the relative plausibility of species selection options in rate-controlled constrained equilibrium (RCCE) reduced chemical models, accounting for uncertainty in the kinetic parameters and experimental data used to refine them. This approach balances the joint goals of matching available data and avoiding overfitting, which is well-understood to limit extrapolative capacity for true prediction. The methodology is applied to homogeneous autoignition, where predictions are known to be particularly sensitive to chemical model details, specially at low temperatures. It is first introduced for hydrogen–air autoignition using an established mechanism, then demonstrated in two applications of methane–air autoignition using the larger GRI-1.2 mechanism. This larger mechanism significantly increases the computational cost of model selection (though not of the subsequent application in predictions), which is alleviated with a time-scale-guided pre-sorting strategy. Uses and extensions of this new formulation are discussed.

1. Introduction

Autoignition is an important challenge in combustion, and numerical simulations provide means for furthering our understanding of it, especially in complex flows where experiments are challenging or costly. Yet the inclusion of detailed combustion chemistry in reactive flow simulations is well-known to increase their computational expense. Their cost scales linearly with the number of transported species for explicit time integration, and quadratically if it is implicit [1]. This precludes the use of most sophisticated hydrocarbon mechanisms, which can have over a thousand species [2]. Often more restrictive than the cost per time step, detailed mechanisms, of course, also introduce computational stiffness, with chemical time scales often 10^2 to 10^7 times smaller than the flow time scales [3]. Methods for chemical model reduction, as we analyze here, attempt to alleviate this cost while retaining accuracy.

Typically, chemical model reduction methods target one of two nominal model levels [4]. High-level skeletal reduction methods extract suitable smaller chemical networks from the detailed mechanism by removing relatively unimportant species and reactions. Specific methodologies include directed relation graph (DRG) [4–6] and computational singular perturbation (CSP) [7] methods. Correspondingly, in low-level methods, the kinetic rate equations are filtered in some way to suppress or remove fast (and presumably less important) time scales. A reduced set of represented species is evolved in time, with unrepresented species reconstructed as needed from algebraic relations. These are generally known as manifold reduction methods because these algebraic relations define a manifold constraint. Specific methodologies include intrinsic low-dimensional manifolds (ILDM) [3], the quasi-steady state approximation (QSSA) [8], and rate-controlled constrained equilibrium (RCCE) [9]. A multi-stage reduction strategy can be formulated by combining methods from each level with the goal of maximizing efficiency [10,11].

Any reduction incurs approximation errors with respect to the detailed model. In particular, RCCE reduced-model predictions are sensitive to the specific choice of represented species [12]. Established species-selection approaches use calibration (training) data in simple canonical reactor configurations to guide selection of represented species [13–18]. An issue we address is that these species-selection methodologies all assume that kinetic mechanisms are trusted, although it is well-understood that they do contain structural and parametric uncertainties [19–21]. Including these is necessary to fully assess the accuracy of subsequent predictions. Furthermore, kinetic uncertainties are understood to affect chemical-model reduction, and have been incorporated in high-level model reduction strategies before. For example, Galassi et al. [22] extended a CSP-based skeletal reduction method to

* Corresponding author.
E-mail addresses: csnrsgr2@illinois.edu (E. Cisneros-Garibay), jbfreund@illinois.edu (J.B. Freund).

https://doi.org/10.1016/j.combustflame.2019.06.028
0010-2180/

the case of uncertain rate coefficients. Reactions were retained according to the probability that they contribute to the rate-controlling dynamics of selected target species, and it was shown that skeletal mechanisms are more likely to be deemed acceptable (according to a user-specified error tolerance) if the uncertainty in the detailed mechanism is taken into account. Xin et al. [23] analyzed the predictive uncertainties of skeletal mechanisms obtained through DRG methods and showed that strong coupling between discarded reactions can lead to non-monotonic predictive uncertainties with the number of retained reactions. Such high-level extended reduction strategies are independent from species selection and predictive uncertainty of RCCE reduction, although the resulting mechanisms can serve as a starting point for subsequent RCCE species selection.

Of course, parametric uncertainty and model reduction are not independent. Here we develop and demonstrate a framework that systematically accounts for kinetic uncertainties concurrently with species selection. This is applied specifically to RCCE reduction, and it ultimately quantifies predictive uncertainties for the reduced model. These different components are joined by Bayesian inference [24,25], a widely-used statistical framework that can quantify the relative merits of competing models. It also provides a means to bring new data into this assessment. Our framework is related to the high-level skeletal reduction method of Galagali and Marzouk [26], also based on Bayesian inference and in which reactions are included based both on quantitatively expressed prior knowledge and their plausibility in the light of new data. It is also related to the methodology of Hakim et al. [27], where Bayesian inference is used to calibrate and select an optimal error model for a globally-reduced chemistry. The fundamental novelty of our species selection framework is that reduction is linked with uncertainty quantification at the low level of the kinetic rate equations, which has the benefit that cost reduction and predictive uncertainty can be jointly assessed. The present method is designed in detail and demonstrated for the commonly used RCCE reduction, though the general approach for achieving overall predictive accuracy could be realized in conjunction with other reduction strategies.

The approach follows a common model reduction paradigm: calibration data is used to select species to represent and, in this case, calibrate uncertain rate parameters in a simple, yet representative, configuration. Because predicting autoignition in complex flows is particularly challenging and important, we demonstrate the approach for autoignition delay times in homogeneous reactive mixtures.

The RCCE approach we focus on is summarized in Section 2, including an introduction to its notation, which is important for subsequent analysis. References are provided to more detailed descriptions. The new framework for RCCE species selection is developed in Section 3. As we shall see, the framework balances fit-to-data against kinetic-parameter uncertainty to select the most plausible (determined quantitatively) set of represented species. This is demonstrated for the relatively simple case of hydrogen–air combustion in Section 4. To extend to more complex mechanisms, in Section 5 we integrate a time-scale-based methodology [18] to manage the large number of candidate sets of represented species. This enables the application of Bayesian species selection to methane–air autoignition in Section 6. Furthermore, the effects of kinetic uncertainties on model size are quantified in Section 7. Additional discussion of results, potential extensions and simplifications, and a summary of the principal features of our methodology are presented in Section 8.

2. Models and their reduction

This section introduces the basic mathematical notation that describes full and reduced models. All cases start with a combustion mechanism $\mathcal{M}$ in which N species in set $\mathcal{S}$, constrained by the law of mass action, participate in M elementary reactions. A model where all species obey differential conservation laws is a *full model*. The number of species in a mechanism for combustion of practical fuels can exceed hundreds or even thousands. To reduce the cost of representing all species, either the raw cost or due to numerical stiffness, we seek a *reduced model* where only a subset of represented species $\mathcal{R}$ evolve according to conservation laws while the rest obey algebraic equations founded on chemical equilibrium. These algebraic constraints define a low-dimensional manifold.

For specificity, the subsequent development will target ignition-time prediction. In an experiment, this is typically measured based on the increase of temperature or a combustion radical. For the hydrogen–air system of Sections 3 and 4, the time of maximum H radical molar fraction z_{H},

$$\tau_{\mathrm{ign}} = \arg\ \max_t z_{\mathrm{H}}(t), \tag{1}$$

coincides with other criteria [28]. For the methane–air system of Sections 5 through 7, H (or any particular species) will not necessarily be in the reduced set $\mathcal{R}$, so we simply use a threshold temperature T_{ign}:

$$\tau_{\mathrm{ign}} = \{t : T(t) = T_{\mathrm{ign}}\}. \tag{2}$$

2.1. Full models

For autoignition, a homogeneous mixture of ideal gases is assumed to evolve at constant pressure p and enthalpy h_0, while the chemical composition is characterized by the specific molar fractions $\boldsymbol{z} = \{z_i\}_{i=1}^N$ of the species. These evolve according to

$$\frac{d\boldsymbol{z}}{dt} = \boldsymbol{\Omega}(\boldsymbol{z}; \boldsymbol{\kappa}), \tag{3}$$

with initial condition $\boldsymbol{z}(0) = \boldsymbol{z}_0$. The chemical source term $\boldsymbol{\Omega}(\boldsymbol{z}; \boldsymbol{\kappa})$ is parameterized by the reaction rate coefficients $\boldsymbol{\kappa} = \{\kappa_i(T)\}_{i=1}^M$. Energy conservation equation imposes an algebraic constraint,

$$\boldsymbol{h}^T\boldsymbol{z} = h_0, \tag{4}$$

where $\boldsymbol{h} = \{h_i(T)\}_{i=1}^N$ are the species specific enthalpies. In addition, the specific molar fractions $\boldsymbol{z}$ satisfy the normalization condition

$$\boldsymbol{W}^T\boldsymbol{z} = 1, \tag{5}$$

where $\boldsymbol{W} = \{W_i\}_{i=1}^N$ are the species molecular weights. Detailed expressions for $\boldsymbol{\Omega}$ are readily available [29,30].

It is important to recognize that (3) has N_e invariants owing to the conservation of the N_e chemical elements. These are imposed by the constant element fractions $\boldsymbol{z}_e = \{z_{e,i}\}_{i=1}^{N_e}$, which are related to the species specific molar fractions by the $N \times N_e$ constant element composition matrix $\mathbf{E}$,

$$\boldsymbol{z}_e = \mathbf{E}^T\boldsymbol{z}. \tag{6}$$

2.2. RCCE reduced models

RCCE model reduction, including the underlying fundamental assumptions, its implementation, and its performance in reactive-flow simulations, is well-documented [31–34]. Of its various formulations, we follow most closely its original form introduced by Keck and Gillespie [9]. In RCCE, $K < N$ constraints $\mathcal{R}$ are evolved in time, each of which can be either (i) an individual (represented) species [15], or (ii) a linear combinations of species [35]. Due to ease of implementation and interpretation of results, we work exclusively with constraints of the first type, though the subsequent development is independent of this choice and can be easily applied to generalized linear constraints (as further discussed in

Section 8). The specific molar fractions $\boldsymbol{r}$ of the represented species $\mathcal{R}$ are related to $\boldsymbol{z}$ through

$$\boldsymbol{r} = \mathbf{R}^T \boldsymbol{z}, \tag{7}$$

where $\mathbf{R}$ is an $N \times K$ constant-valued matrix with columns corresponding to the unit vectors of the represented species. For example, if only the k^{th} and ℓ^{th} species are represented, $\mathbf{R}$ is zero except for $R_{k,1} = R_{\ell,2} = 1$. Combined, (3) and (7) form an evolution system for the reduced state,

$$\frac{d\boldsymbol{r}}{dt} = \mathbf{R}^T \boldsymbol{\Omega}(\boldsymbol{z}; \boldsymbol{\kappa}), \tag{8}$$

with initial condition $\boldsymbol{r}(0) = \boldsymbol{r}_0 = \mathbf{R}^T \boldsymbol{z}_0$. The immediate challenge is that evaluation of $\boldsymbol{\Omega}(\boldsymbol{z}; \boldsymbol{\kappa})$ requires an estimate of the full composition vector $\boldsymbol{z}$ based on the reduced representation $\boldsymbol{r}$ (along with the pressure p, enthalpy h_0, and element fractions $\boldsymbol{z}_e$). This reconstruction is defined by the constrained-equilibrium composition $\boldsymbol{z}_c = \{z_{c,i}\}_{i=1}^N$,

$$\boldsymbol{z} = \boldsymbol{z}_c(\boldsymbol{r}; p, h_0, \boldsymbol{z}_e), \tag{9}$$

which maximizes the entropy of the mixture, subject to the constraints (4) through (7),

$$z_{c,i} = \frac{1}{W} \exp\left(-\frac{g_i(T)}{\mathscr{R}T} + \sum_{n=1}^{N_e} E_{i,n}\, \gamma_n + \sum_{k=1}^{K} R_{i,k}\, \mu_k \right), \tag{10}$$

where $\mathscr{R}$ is the universal gas constant and $g_i(T) = h_i(T) - T s_i(T)$ for $i = 1, \ldots, N$, are the species standard Gibbs free energies, with $s_i(T)$ their corresponding specific entropies. The temperature T, the mean molecular weight W, the element potentials γ_n for $n = 1, \ldots, N_e$, and the constraint potentials μ_k for $k = 1, \ldots, K$, enforce (4) through (7) [9].

Existing methodologies for selection of $\mathcal{R}$ are typically based on how well candidate sets reproduce full-model data [15,17,18]. While effective, they disregard uncertainty in the rate coefficients and the data used to guide selection. This leads to the following two drawbacks: (i) that the reduction process is over-constrained; and (ii) that selected models lack predictive capacity, which is most assured only if all knowledge about the models, their parameters and the calibration data is incorporated. Including uncertainty in species selection is addressed in the following.

3. Species selection accounting for uncertainty

The proposed framework selects represented species $\mathcal{R}$, accounting for the uncertainty of the rate parameters in $\boldsymbol{\Omega}$ in (8). This affects what is deemed the best (or most justified) reduced model and is a key step in quantifying their predictive uncertainty. The approach is most easily introduced by example, for which we consider an isobaric, adiabatic, homogeneous hydrogen–air reactor. The combustible mixture is taken to include $N = 9$ species $\mathcal{S} = \{H_2, O_2, H, O, OH, HO_2, H_2O_2, H_2O, N_2\}$ and thus $N_e = 3$ elements $\mathcal{E} = \{H, O, N\}$, and reactions are modeled by the Saxena & Williams [36] mechanism.

3.1. Objective, illustrated for H_2-air autoignition

We predict autoignition time τ_{ign} (1) at stoichiometric, near-atmospheric conditions, and different initial temperatures, for which measured $\boldsymbol{\tau}^d_{\text{ign}}$ [37] are shown in Fig. 1. (A case for which no such data is available is considered in Section 6.) An RCCE reduced model with represented species $\mathcal{R}$ is designed to reproduce the data $\boldsymbol{\tau}^d_{\text{ign}}$. A priori, however, it is unclear how to best select $\mathcal{R}$. For this illustration, intuition and experience can be a guide, though this is not the case in general. We consider the two candidates sets $\mathcal{R}_a = \{H_2, O_2, H, HO_2\}$ and $\mathcal{R}_b = \{H_2, O_2, H, OH\}$,

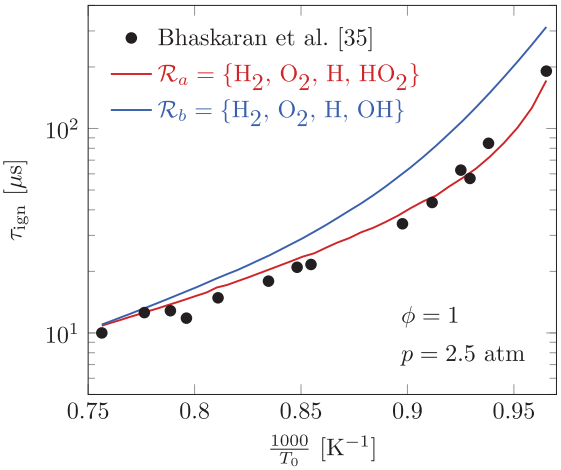


Fig. 1. Ignition-time predictions and measurements [37] for two RCCE species selection choices with standard kinetic parameter values.

which differ only in a single species. This is tantamount to selecting one of two radicals (OH or HO_2) to augment what we take to be the baseline set $\mathcal{R}_0 = \{H_2, O_2, H\}$. The predictions of the two reduced models are shown in Fig. 1 as a function of the initial temperature of the mixture T_0. This example makes the species selection challenge concrete: changing a single represented species leads to obviously different ignition times. Of the two candidates, $\mathcal{R}_a$ (which includes HO_2 rather than OH) matches the data better, and so without additional considerations $\mathcal{R}_a$ would be considered the best choice.

This conclusion, however, disregards any uncertainty in the mechanism, which we have tacitly assumed to be fully trusted. The effect of uncertainty on species selection will be quantified in the following. RCCE models will then be built upon mechanisms with uncertain kinetic parameters. In a simple model-fitting approach, we could choose values that lead to best matches of data. However, the ability to reproduce particular data does not guarantee that the RCCE models will extrapolate to different scenarios, which is the ultimate goal of prediction. Indeed, reproducing data can be achieved with any interpolating functions; designing chemistry models to reflect the underlying kinetics should grant them extrapolative capacity. A complicating factor is that their rate parameters, which come from auxiliary sources, such as shock tube experiments or *ab initio* theoretical calculations, bring in their own additional uncertainties. This information is also disregarded by simple model fitting, yet it is important for prediction and couples with species selection. The following subsections develop a means for using kinetic uncertainties in conjunction with reduced-model accuracy to inform the species selection that defines a particular RCCE model.

3.2. Quantifying relative confidence in candidate species

The immediate goal is to quantify the evidence supporting any particular set of represented species, $\mathcal{R}_a$ or $\mathcal{R}_b$ in this example, as the best for any available data and kinetic-parameter uncertainties. Probabilities, interpreted as degrees-of-belief, are well-understood to quantify this [24] and applying such quantification to physical systems is becoming increasingly common [38–40]. Following the standard Bayesian perspective [24,25], the degree-of-belief for each candidate $\mathcal{R}$, before comparing with data $\boldsymbol{\tau}^d_{\text{ign}}$, is denoted as $P(\mathcal{R} \mid I)$, representing our *prior* confidence in $\mathcal{R}$, occasionally referred to as prior model probability [41]. It is conditioned only on the auxiliary information I that frames the range of scenarios. For chemical kinetics, I thus includes physical reasoning,

measurement error estimates from rate experiments, and experience with, say, related mechanisms or quantities of interest (e.g., flame speeds or extinction limits). Therefore, $P(\mathcal{R}\,|\,I)$ serves to bridge intuition and qualitative (possibly even subjective) knowledge with quantitative and reportable input [24]. For example, established intuitive guidance for RCCE species selection [42]—to represent species that evolve slowly, occur in significant concentrations, or are thought to participate in rate-controlling pathways—can be built into the framework through $P(\mathcal{R}\,|\,I)$, with candidate species that satisfy these requirements being assigned a relatively high prior probability. The extensive reasoning behind prior development is well-established [24,25] and not discussed extensively here since it is not central to our goals. Rather, we simply recognize that it should encompass all available information aside from the measured $\boldsymbol{\tau}^d_{\text{ign}}$, and for our illustration provide uniform priors to all candidates.

The next step is to refine the relative confidence in particular candidate species given data τ^d_{ign} and kinetic-parameter uncertainty. Given $\boldsymbol{\tau}^d_{\text{ign}}$, the updated relative confidence in candidate species is $P(\mathcal{R}\,|\,\boldsymbol{\tau}^d_{\text{ign}},\, I)$, which is a *posterior* degree-of-belief, or plausibility, regarding candidate $\mathcal{R}$. It is also referred to as posterior model probability [41]. The candidate with the highest $P(\mathcal{R}\,|\,\boldsymbol{\tau}^d_{\text{ign}},\, I)$ is the most supported. Note that the posterior is also conditioned on I. Therefore, our conclusions will only be as general as I allows them to be, effectively constraining what best means. In this illustration, I includes our selected starting-point represented species $\mathcal{R}_0$, upon which candidate representations are constructed. The selected species in this case are the best to augment $\mathcal{R}_0$.

3.3. Supporting evidence for candidate species

3.3.1. Bayesian evidence

As expressed, the posterior $P(\mathcal{R}\,|\,\boldsymbol{\tau}^d_{\text{ign}},\, I)$ (for $\mathcal{R}=\mathcal{R}_a$ or $\mathcal{R}_b$ in this example) might seem abstract, but it can be evaluated via Bayes' theorem [24,25], which relates it to the prior $P(\mathcal{R}\,|\,I)$:

$$P(\mathcal{R}\,|\,\boldsymbol{\tau}^d_{\text{ign}},\, I) = \frac{P(\boldsymbol{\tau}^d_{\text{ign}}\,|\,\mathcal{R},\, I)P(\mathcal{R}\,|\,I)}{P(\boldsymbol{\tau}^d_{\text{ign}}\,|\,I)}. \tag{11}$$

The numerator of (11) is our principal concern. The $P(\boldsymbol{\tau}^d_{\text{ign}}\,|\,\mathcal{R},\, I)$ multiplying the prior $P(\mathcal{R}\,|\,I)$ modifies our degree-of-belief supporting $\mathcal{R}$ based on how well a particular $\mathcal{R}$ reproduces $\boldsymbol{\tau}^d_{\text{ign}}$. As such, it is the *evidence* (also refered to as model evidence [27]), and unlike $P(\mathcal{R}\,|\,\boldsymbol{\tau}^d_{\text{ign}},\, I)$ it can be directly evaluated. A complication is that it also incorporates (in I) the uncertainty in the kinetic parameters $\boldsymbol{\kappa}$, which must be included in the evaluation. These too have relative probabilities, based on a corresponding degree-of-belief $P(\boldsymbol{\kappa}\,|\,I)$ about their true values. Even simple $\pm$ error specifications can be expressed in corresponding specific forms for $P(\boldsymbol{\kappa}\,|\,I)$, as will be considered in Section 3.3.3. For now, integrating over all $\boldsymbol{\kappa}$ yields an expression for the evidence in (11) that includes the kinetic uncertainty $P(\boldsymbol{\kappa}\,|\,I)$:

$$P(\boldsymbol{\tau}^d_{\text{ign}}\,|\,\mathcal{R},\, I) = \int P(\boldsymbol{\tau}^d_{\text{ign}}\,|\,\boldsymbol{\kappa},\, \mathcal{R},\, I)P(\boldsymbol{\kappa}\,|\,I)\,\mathrm{d}\boldsymbol{\kappa}. \tag{12}$$

The first term in the integrand is the *likelihood*, so-called because it quantifies how likely it is that a reduced model with particular $\mathcal{R}$ and $\boldsymbol{\kappa}$ matches $\boldsymbol{\tau}^d_{\text{ign}}$. Specific formulas for the likelihood will also be considered subsequently in Section 3.3.2.

Since the denominator in (11) does not depend on $\mathcal{R}$, it is not explicitly needed to compare our relative confidence in candidate species. Instead, we can infer its value from the normalization condition

$$\sum_{\mathcal{R}\in\mathbb{A}} P(\mathcal{R}\,|\,\boldsymbol{\tau}^d_{\text{ign}},\, I) = 1, \tag{13}$$

where $\mathbb{A}$ is the set of all candidate sets of represented species. For the current hydrogen autoignition demonstration it is simply $\mathbb{A} = \{\mathcal{R}_a, \mathcal{R}_b\}$. This normalization shows that candidate models can only be compared; if only one $\mathcal{R}$ is available, it is necessarily the best with $P(\mathcal{R}\,|\,\boldsymbol{\tau}^d_{\text{ign}},\, I) = 1$.

3.3.2. Errors and likelihood

The likelihood function is a key component of the formulation since it quantifies the mismatch ε_{ign} between a reduced-model prediction $\tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{R})$ and the corresponding data τ^d_{ign},

$$\tau^d_{\text{ign}} = \tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{R}) + \varepsilon_{\text{ign}}. \tag{14}$$

We assume that the ignition data is also accompanied by an estimate of its uncertainty σ_{ign}. Taking measurement uncertainty as the only source of mismatch between data and model predictions, and in the absence of any additional information, we take ε_{ign} to be normally distributed with zero mean and standard deviation σ_{ign}, which maximizes statistical entropy in a way that reflects maximal ignorance about other aspects of the mismatch [25]. This assumption leads to the likelihood

$$P(\tau^d_{\text{ign}}\,|\,\boldsymbol{\kappa},\, \mathcal{R},\, I) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{ign}}}} \exp\left\{-\frac{1}{2\sigma^2_{\text{ign}}}\left[\tau^d_{\text{ign}} - \tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{R})\right]^2\right\}. \tag{15}$$

To generalize (15) to N_{ign} measurements (e.g., the ignition times in Fig. 1) it is common practice to assume, for simplicity, that the experimental errors $\varepsilon_{\text{ign},i}$, $i = 1, \ldots, N_{\text{ign}}$, are independent and identically distributed with standard deviations $\sigma_{\text{ign},i}$ [38], which yields

$$P(\boldsymbol{\tau}^d_{\text{ign}}\,|\,\boldsymbol{\kappa},\, \mathcal{R},\, I) = \prod_{i=1}^{N_{\text{ign}}} \frac{1}{\sqrt{2\pi\sigma^2_{\text{ign},i}}} \times \exp\left\{-\frac{1}{2\sigma^2_{\text{ign},i}}\left[\tau^d_{\text{ign},i} - \tau_{\text{ign},i}(\boldsymbol{\kappa}, \mathcal{R})\right]^2\right\}. \tag{16}$$

This additive form of the error model (14) is a choice made based on objectives, and any particular conclusions are tied to it. Here it is motivated by its simplicity for illustration; the goals of any particular application might suggest different forms. An alternative form that has been used in combustion chemistry uncertainty quantification is multiplicative [38],

$$\tau^d_{\text{ign}} = e^{\varepsilon_{\text{ign}}}\tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{R}) \tag{17}$$

which is useful when the data and quantity of interest span a wide range or must have the same sign. For N_{ign} measurements, and following the same assumptions as for the additive error model (14), leads to the likelihood

$$P(\boldsymbol{\tau}^d_{\text{ign}}\,|\,\boldsymbol{\kappa},\, \mathcal{R},\, I) = \prod_{i=1}^{N_{\text{ign}}} \frac{1}{\sqrt{2\pi\sigma^2_{\text{ign},i}}} \times \exp\left\{-\frac{1}{2\sigma^2_{\text{ign},i}}\log\left[\frac{\tau^d_{\text{ign},i}}{\tau_{\text{ign},i}(\boldsymbol{\kappa}, \mathcal{R})}\right]^2\right\}, \tag{18}$$

which is used in Section 5. We do not delve into more sophisticated error modeling, which would depend on the specifics of corresponding experiments, or model inadequacies that lead to model errors as a source of mismatch. These efforts are challenging and require considerable modeling [43].

3.3.3. Parametric uncertainty

A specific form of $P(\boldsymbol{\kappa}\,|\,I)$ is also needed to evaluate (12), which in turn requires specification of what exactly $\boldsymbol{\kappa}$ represents. Reactions are accompanied by parameterized forms (e.g., Arrhenius, third-body, or falloff) for their rate coefficients $\boldsymbol{k} = \{k_i\}_{i=1}^{M}$. For the nominal rate coefficient value k_i^0, taken to be its recommended parameterization, a normalized random variable κ_i is used to represent its uncertainty relative to its true (unknown) value k_i,

$$\kappa_i = \frac{\log\left(k_i/k_i^0\right)}{\log f_i}, \tag{19}$$

where $f_i > 1$ is the corresponding multiplicative uncertainty factor. These factors are often inferred in estimates of k_i uncertainty [44–46]. As formulated, the most plausible values of the rate coefficient are taken to lie between k_i^0/f_i and $f_i k_i^0$. Without additional information, taking a provided uncertainty range in k_i to be a standard deviation yields a Gaussian distribution,

$$P(\boldsymbol{\kappa}\,|\,I) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\kappa_i^2}{2}\right), \tag{20}$$

based, as for the likelihood, on maximizing statistical entropy [24,25].

A more thorough treatment of rate uncertainties would involve formulating states of knowledge of the individual Arrhenius and any falloff or other parameters. This could be done using existing frameworks that infer the correlation structure for the Arrhenius parameters from the recommended nominal values and provided uncertainty factors [46,47]. However, (20) is simpler for introducing the method, and if deemed necessary, generalizing to these more involved approaches should be straightforward.

3.4. Calibration and overfitting

In addition to informing species selection ($\mathcal{R}_a$ versus $\mathcal{R}_b$ in this example), data such as τ_{ign}^d, when available, can be used to infer the relative likelihood of $\boldsymbol{\kappa}$ values. This can be utilized to improve the accuracy of any $\mathcal{R}$ and refine bounds on its predictive uncertainty by refining $P(\boldsymbol{\kappa}\,|\,I)$, not just selecting a best-fit value as for the simplest model fitting. This *calibration* is also achieved through the application of Bayes' theorem [24,25]:

$$P(\boldsymbol{\kappa}\,|\,\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R},\,I) = \frac{P(\boldsymbol{\tau}_{\text{ign}}^d\,|\,\boldsymbol{\kappa},\,\mathcal{R},\,I)P(\boldsymbol{\kappa}\,|\,I)}{P(\boldsymbol{\tau}_{\text{ign}}^d\,|\,\mathcal{R},\,I)}. \tag{21}$$

The updated state of knowledge $P(\boldsymbol{\kappa}\,|\,\mathcal{R},\,I)$ is then the $\boldsymbol{\kappa}$ *posterior*, which now depends on $\mathcal{R}$, and is thus linked to that model.

A risk of any calibration is overfitting, meaning that the model parameters become too strongly tied to the available data. This is undesirable since it diminishes confidence in the extrapolative capacity of the model, which is the ultimate goal of any model. Overfitting is countered in Bayesian model selection, guarding against the selection of overfitted models. This constraint quantifies the simplifying principle of Occam's razor: of all candidate models with equal accuracy, the most robust—least slavishly linked to data—is favored.

That the evidence (12) counters overfitting can be seen by decomposing its logarithm $F(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R}) \equiv \log P(\boldsymbol{\tau}_{\text{ign}}^d\,|\,\mathcal{R},\,I)$ [48],

$$F(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R}) = \underbrace{\int \log\left[P(\boldsymbol{\tau}_{\text{ign}}^d\,|\,\boldsymbol{\kappa},\,\mathcal{R},\,I)\right] P(\boldsymbol{\kappa}\,|\,\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R},\,I)\,\mathrm{d}\boldsymbol{\kappa}}_{U(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})}$$
$$- \underbrace{\int \log\left[\frac{P(\boldsymbol{\kappa}\,|\,\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R},\,I)}{P(\boldsymbol{\kappa}\,|\,\mathcal{R},\,I)}\right] P(\boldsymbol{\kappa}\,|\,\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R},\,I)\,\mathrm{d}\boldsymbol{\kappa}}_{S(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})}. \tag{22}$$

The first term, $U(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})$, rewards the fit-to-data achieved using $\mathcal{R}$, averaged over the posterior PDF of the uncertain kinetic parameters (which in turn depends on their prior PDF through (21)). The second term, $S(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})$, is the information entropy change between the prior and posterior PDFs [25], which measures the amount of information gained about the uncertain parameters from the data. The larger this term, the more adjustment of model parameters is required to achieve the corresponding data-match $U(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})$, and the more it risks being enslaved to that particular data. Because $S(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})$ penalizes overfitting, it is referred to as an Occam factor: it distinguishes between models that achieve the same fit-to-data $U(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})$ by favoring those that rely less on calibration. Occam factors have a physical interpretation. Physical understanding—less assumptions—restricts the prior range of plausible model-parameter values, and new data governed by this same physics will be less likely to further restrict the parameters [24]. For (22), this would mean that $P(\boldsymbol{\kappa}\,|\,\tau_{\text{ign}}^d,\,\mathcal{R},\,I) \approx P(\boldsymbol{\kappa}\,|\,I)$, so $S(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R}) \approx 0$. In contrast, a model with little physical basis would not have corresponding restriction and thus would have broader priors. These would allow the calibration procedure to find $\boldsymbol{\kappa}$ values that lead to close match with the data, resulting in relatively high $U(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})$. However, new data would then be highly informative about the uncertain parameters, yielding large Occam penalty $S(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})$, countering gains in $U(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R})$.

To simply illustrate how the Occam factor in (22) functions, we assume (crudely) that the reduced-model ignition-time predictions depend linearly in the uncertain parameters,

$$\boldsymbol{\tau}_{\text{ign}} \approx \boldsymbol{\tau}_{\text{ign}}^0 + \mathbf{L}^T\boldsymbol{\kappa}, \tag{23}$$

where the intercept $\boldsymbol{\tau}_{\text{ign}}^0$ and the coefficients $\mathbf{L}$, if specific values were needed, could be obtained by Taylor expansions or least-squares fitting. This linearization is tantamount to assuming a Gaussian posterior distribution (21) with covariance matrix

$$\boldsymbol{\Xi}_* = \left[\mathbf{L}\boldsymbol{\Xi}_{\text{ign}}\mathbf{L}^T + \mathbf{I}\right]^{-1}, \tag{24}$$

where $\boldsymbol{\Xi}_{\text{ign}} = \text{diag}\left(\sigma_{\text{ign},i}^{-2}\right)$ for $i = 1, \ldots, N_{\text{ign}}$, and $\mathbf{I}$ is the identity matrix. The mean $\boldsymbol{\kappa}^*$ is

$$\boldsymbol{\kappa}^* = \boldsymbol{\Xi}_*\mathbf{L}\boldsymbol{\Xi}_{\text{ign}}\left(\boldsymbol{\tau}_{\text{ign}}^d - \boldsymbol{\tau}_{\text{ign}}^0\right). \tag{25}$$

Applying this linearization to S in (22) leads to

$$S(\boldsymbol{\tau}_{\text{ign}}^d,\,\mathcal{R}) = \frac{1}{2}\|\boldsymbol{\kappa}^*\|^2 + \frac{1}{2}\log\left|\boldsymbol{\Xi}_*^{-1}\right|, \tag{26}$$

which shows two distinct overfitting features penalized. The $\|\boldsymbol{\kappa}^*\|^2$ term is the penalty for the calibration shifting the mean $\boldsymbol{\kappa}^*$ away from the $\boldsymbol{\kappa} = \mathbf{0}$ of the prior PDF (20). The further $\boldsymbol{\kappa}^*$ moves from the nominal values, the more it is affected by the data and the higher the risk of overfitting. The $\log\left|\boldsymbol{\Xi}_*^{-1}\right|$ term quantifies the change in volume from prior to posterior PDF. It thus penalizes the reduction of parametric uncertainty by the data. The larger this term, the more the data restricts parameter values. This also risks overfitting because if very specific values relative to the prior, are required to match the data, then it is less likely to extrapolate. A physically-grounded model should change little if calibrated with additional data from a corresponding physical system. It should be emphasized that uncertainty minimization by data is not of itself detrimental: it is only important when comparing models of similar prior expectation. The model that can fit the data without restricting its parameters $\boldsymbol{\kappa}$ is more probable to be accurately extrapolative.

3.5. Prediction with uncertainty

With $\boldsymbol{\kappa}$ calibrated, we can evaluate the model to make predictions. Because $\boldsymbol{\kappa}$ is a random variable, the model output $\tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{R})$

will be a distribution, so an expectation is a convenient way to express a particular prediction,

$$\langle \tau_{\text{ign}} \rangle = \int \tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{R}) P(\boldsymbol{\kappa} \,|\, \boldsymbol{\tau}^d_{\text{ign}}, \mathcal{R}, I)\, \mathrm{d}\boldsymbol{\kappa}. \tag{27}$$

For true prediction, we have no data to assess the accuracy of (27), so we rely on the predictive uncertainty, here measured as a standard deviation from the expected value:

$$\hat{\sigma}_{\text{ign}} = \left\{ \int \left[\tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{R}) - \langle \tau_{\text{ign}} \rangle \right]^2 P(\boldsymbol{\kappa} \,|\, \boldsymbol{\tau}^d_{\text{ign}}, \mathcal{R}, I)\, \mathrm{d}\boldsymbol{\kappa} \right\}^{\frac{1}{2}}. \tag{28}$$

The smaller $\hat{\sigma}_{\text{ign}}$, the more we trust the prediction $\langle \tau_{\text{ign}} \rangle$. A signature of overfitting can be recognized in (28): growth of $\hat{\sigma}_{\text{ign}}$ away from the calibration range indicates lost confidence in the extrapolative predictive accuracy of $\mathcal{R}$.

In some cases, the predictive uncertainty (28) can be decomposed into contributions due to measurement uncertainty and a remainder (model) error. This is possible in general when (i) the prediction (27) falls well within measurement uncertainty, and (ii) $\hat{\sigma}_{\text{ign}} > \sigma_{\text{ign}}$. As we shall see, it is possible to apply for the H_2-air demonstration, and there it does not alter any conclusions.

4. Demonstration: H_2-Air Ignition

4.1. Species priors and supporting evidence

We demonstrate our framework evaluating reduced model $\mathcal{R}_a$ versus $\mathcal{R}_b$ for the relatively simple case of H_2-air autoignition, given $\boldsymbol{\tau}^d_{\text{ign}}$ from Fig. 1. Assuming no prior information regarding the suitability of $\mathcal{R}_a$ versus $\mathcal{R}_b$, we specify equal probabilities

$$P(\mathcal{R}_a \,|\, I) = P(\mathcal{R}_b \,|\, I) = \frac{1}{2}, \tag{29}$$

so their relative plausibility simply becomes the ratio of their evidence, or Bayes' factor [49],

$$B \equiv \frac{P(\boldsymbol{\tau}^d_{\text{ign}} \,|\, \mathcal{R}_b, I)}{P(\boldsymbol{\tau}^d_{\text{ign}} \,|\, \mathcal{R}_a, I)} = \frac{P(\mathcal{R}_b \,|\, \boldsymbol{\tau}^d_{\text{ign}}, I)}{P(\mathcal{R}_b \,|\, \boldsymbol{\tau}^d_{\text{ign}}, I)}. \tag{30}$$

For comparing actual results, we use $\log_{10}$ to make interpretation more intuitive:

$$\Delta F \equiv \log_{10} B = F_b - F_a, \tag{31}$$

where $F \equiv \log_{10} P(\boldsymbol{\tau}^d_{\text{ign}} \,|\, \mathcal{R}, I)$. As such, $\mathcal{R}_b$ is more plausible if $\Delta F > 0$ and $\mathcal{R}_a$ more plausible if $\Delta F < 0$.

4.2. Kinetic and measurement uncertainties

To illustrate how parametric uncertainty affects conclusions, the uncertainty factors f_i in (19) are adjusted artificially, and for simplicity all are assigned the same value $f_i = f$. Typical reported kinetic parameter uncertainties have $f \approx 2$ [50], so we take $f = 1.1$ to indicate relatively trusted rate coefficients and $f = 3.6$ to indicate significant uncertainty.

For this demonstration, we assume an additive error model (14), which leads to the additive likelihood function (16). To demonstrate the role of the uncertainty in the data, we take $\sigma_{\text{ign},i} = \sigma'_{\text{ign}} \tau_{\text{ign},i}$ for $i = 1, \ldots, N_{\text{ign}}$ in (16), with $\sigma'_{\text{ign}} = 0.05$ indicating relatively trusted data and $\sigma'_{\text{ign}} = 0.25$ indicating less reliable data. These are based on typical shock-tube measurements, which would have $\sigma'_{\text{ign}} \approx 0.1$ [51].

4.3. Numerical evaluation

Evaluation of the evidence (12) requires high-dimensional integration over $\boldsymbol{\kappa}$. Even the 24 uncertain rate coefficients $\boldsymbol{\kappa}$ in the relatively small San Diego hydrogen mechanism [36] preclude the use of tensor-product quadrature rules. A viable alternative is a Monte Carlo sampling of the likelihood (16) over $P(\boldsymbol{\kappa} \,|\, I)$ (20),

$$P(\boldsymbol{\tau}^d_{\text{ign}} \,|\, \mathcal{R}) \approx \tilde{P}(\boldsymbol{\tau}^d_{\text{ign}} \,|\, \mathcal{R} I) = \frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} P(\boldsymbol{\tau}^d_{\text{ign}} \,|\, \boldsymbol{\kappa}^{(i)}, \mathcal{R}, I), \tag{32}$$

where $\{\boldsymbol{\kappa}^{(i)}\}_{i=1}^{N_{\text{MC}}}$ are the samples of $P(\boldsymbol{\kappa} \,|\, I)$. We quantify the convergence of

$$\tilde{F} \equiv \log \tilde{P}(\boldsymbol{\tau}^d_{\text{ign}} \,|\, \mathcal{R}, I) \tag{33}$$

by evaluating (32) M_{MC} times with different random seeds, which provides an estimate of the sampling error [52]

$$\tilde{\sigma}_{\text{MC}} = \left[\frac{1}{M_{\text{MC}} - 1} \sum_{j=1}^{M_{\text{MC}}} \left(\tilde{F}^{(j)} - \langle \tilde{F} \rangle \right)^2 \right]^{\frac{1}{2}}, \tag{34}$$

where $\tilde{F}^{(j)}$ is the jth evaluation of $\tilde{F}$, and

$$\langle \tilde{F} \rangle = \frac{1}{M_{\text{MC}}} \sum_{j=1}^{M_{\text{MC}}} \tilde{F}^{(j)}. \tag{35}$$

Figure 2 shows $\tilde{\sigma}_{\text{MC}}$ relative to $\langle \tilde{F} \rangle$ with $M_{\text{MC}} = 15$ for $5 \times 10^3 \le N_{\text{MC}} \le 294 \times 10^3$ for two representative pairs of kinetic f and measurement σ_{ign} uncertainties. As expected, sampling errors decay as $N_{\text{MC}}^{-1/2}$. Sampling errors are less than 10% for $N_{\text{MC}} \ge 73 \times 10^3$ for all cases, which is sufficient to establish N_{MC} independence to the needed fidelity.

Though sufficient for demonstrating our species selection framework, Monte Carlo integration (32) is not without drawbacks. Substantial misalignment between the likelihood (16) and the prior of the uncertain parameters (20) can require a large number of samples N_{MC} to achieve convergence. Furthermore, it does not easily generalize to the case arbitrary priors (such as, for example, obtained from a previous Bayesian calibration). Methods that seek to overcome these limitations include nested sampling [53], power posteriors [54], annealed importance sampling [55], harmonic mean estimators [56], transitional Markov chain Monte Carlo [57], the Chib–Jeliazkov method [58], and multi-level sampling [59].

Predictions (27) and predictive uncertainties (28) also require high-dimensional integration over $\boldsymbol{\kappa}$, but with the added challenge that they involve computing the posterior distribution of the uncertain parameters (21). For this, multi-level sampling is used, as implemented in the QUESO library [60]. Results in the next section are insensitive to the number of samples per level, which was confirmed by comparing results for 10^4, 2×10^4, and 4×10^4 samples.

Finally, the reduced-model rate equations (8) are evaluated using the adaptive error-controlled CVode package [61], which adaptively determines the time step Δt to maintain the local truncation error under user-specified absolute E_{abs} and relative E_{rel} tolerances. Results were confirmed insensitive to error tolerances, and are presented here for the specific choices of $\mathrm{E}_{\text{rel}} = 10^{-8}$ (dimensionless) and $\mathrm{E}_{\text{abs}} = 10^{-13}$ (in kmol/kg) tolerances. The constrained-equilibrium composition (10) is obtained using CEQ [62,63].

4.4. Species selection results

The combined effects of kinetic f and data σ'_{ign} uncertainties on selecting $\mathcal{R}_a$ versus $\mathcal{R}_b$ are shown in Fig. 3, with corresponding reduced-model predictions shown in Fig. 4 for particular cases. For small rate parameter uncertainties ($f \lesssim 1.35$), $\mathcal{R}_a = \{H_2, O_2, H, HO_2\}$ is more justified regardless of experimental error σ'_{ign}. Figure 4a shows that for $f = 1.1$ and $\sigma'_{\text{ign}} = 0.1$ all $\mathcal{R}_a$ predictions (27) fall well within $2\sigma_{\text{ign}}$ of τ^d_{ign}. In contrast, $\mathcal{R}_b = \{H_2, O_2, H, OH\}$ does not match the data so closely, particularly

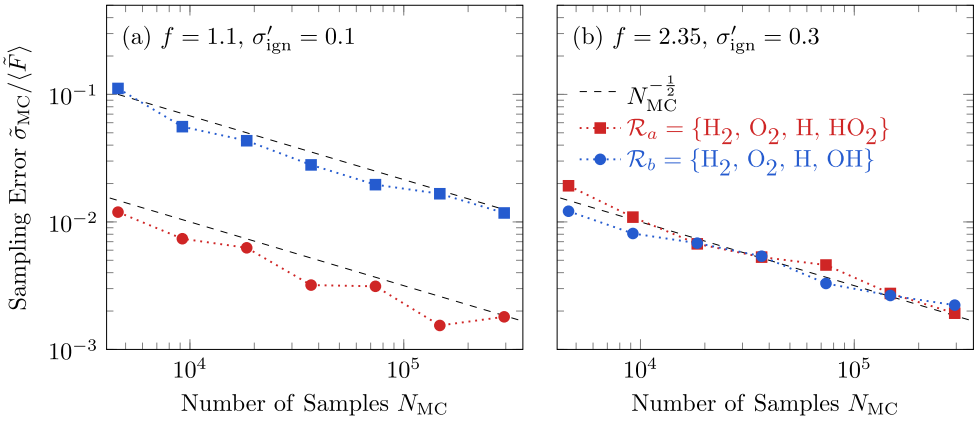


Fig. 2. Relative sampling error (34) in the Monte Carlo approximation of the evidence (32) for two representative pairs of kinetic f and measurement σ'_{ign} uncertainties.

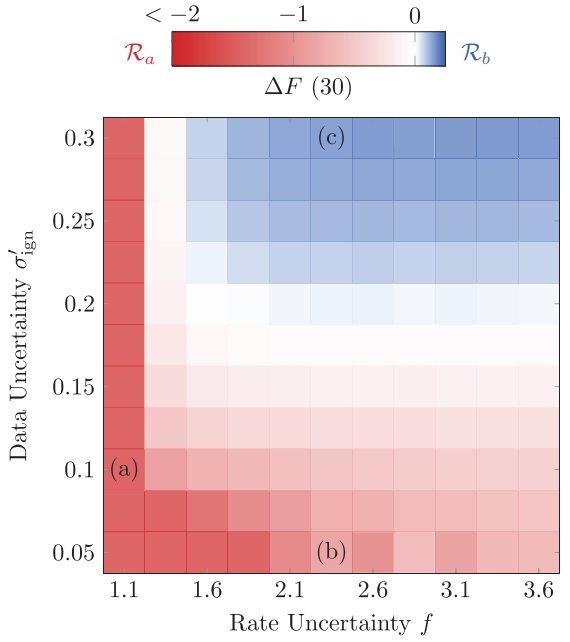


Fig. 3. Species selection evidence (31) for an illustrative range of rate uncertainty factors f in (19) and data uncertainty σ'_{ign} in (16). Colors correspond to ΔF for the cell-centered $(f, \sigma'_{\text{ign}})$ values. Labels indicate cases shown in Fig. 4.

missing the lowest-temperature ignition measurement. This is the case despite of the $\boldsymbol{\kappa}$ re-calibration procedure that aids agreement: Fig. 4b shows much better agreement than nominal-value prediction Fig. 1.

For larger kinetic-parameter uncertainty ($f \gtrsim 1.35$) the $\mathcal{R}_a$ versus $\mathcal{R}_b$ conclusion depends on the measurement error. For $\sigma'_{\text{ign}} \lesssim 0.175$, $\mathcal{R}_a$ remains the better candidate, though $\mathcal{R}_b$ is also close to the data, as seen in Fig. 4b. In this case, the broader rate uncertainties, f, allow for the re-calibration procedure to utilize $\boldsymbol{\kappa}$ that better match the data. Indeed, the fit-to-data becomes comparable to that of $\mathcal{R}_a$, though this better fit is offset by the Occam factor in (22), so $\mathcal{R}_a$ remains more credible. Also indicative of this, predictions with $\mathcal{R}_b$ display a signature feature of overfitting: they closely match the data within the calibration range, yet their predictive uncertainty $\hat{\sigma}_{\text{ign}}$ (28) grows near the end of this range (Fig. 4b inset). This is more significant near the lowest-temperature ignition condition, where it exceeds the measurement uncertainty. As explained in Section 3.4, this suggests that $\mathcal{R}_b$ for this f is overly dependent on the particular data, which reduces confidence in its overall extrapolative capacity. Overall, $\mathcal{R}_a$ is favored because it relies less on calibration to achieve its good match, and thus it is a better candidate for extrapolative predictions.

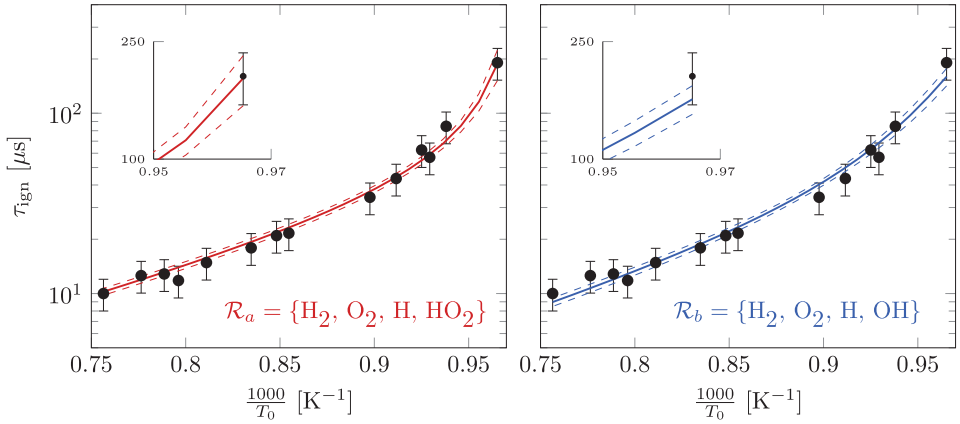
When the data is less reliable, with $\sigma'_{\text{ign}} \gtrsim 0.175$, $\mathcal{R}_a$ is no longer universally favored. For $0.175 \lesssim \sigma'_{\text{ign}} \lesssim 0.2$, the candidate models are comparably credible. We note that in such cases, *model averaging* [64], where reduced-model predictions are combined according to their corresponding plausibilities, can be used to provide the best actual prediction. For $\sigma'_{\text{ign}} \gtrsim 0.2$, $\mathcal{R}_b$ is the more credible. Rate uncertainties are sufficiently large for re-calibration of $\mathcal{R}_b$ to yield predictions that fall within the assumed data uncertainty, though predictions (27) are not very close to every data point, as seen for $\mathcal{R}_b$ in Fig. 4c. In contrast, predictions for $\mathcal{R}_a$ provide near-perfect match to the data, though this closeness is not highly rewarded given the measurement uncertainty. In this case, it is $\mathcal{R}_a$ that displays evidence of overfitting predictions, with especially high predictive uncertainty at low temperature.

5. Extension to larger mechanisms

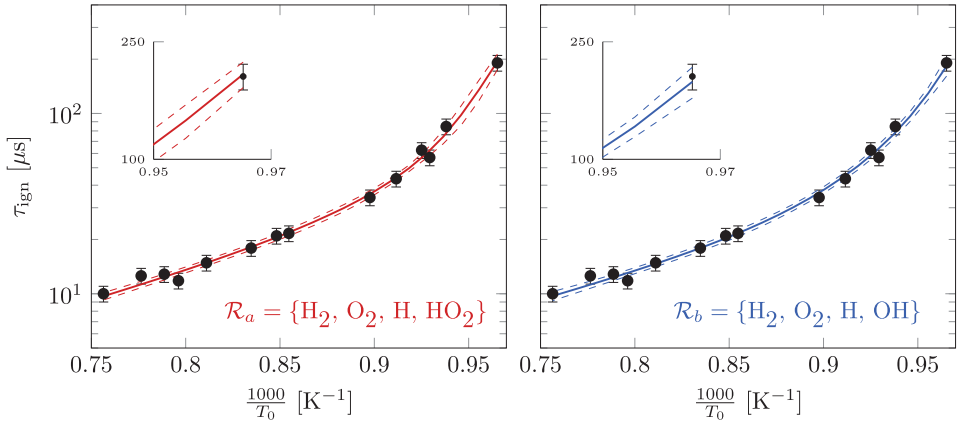
5.1. Species selection challenge

Extending our species selection framework to mechanisms with more species follows the same formulation, though the exploding number of reduced-model options introduces an additional challenge. With only 9 species, 24 reactions, and (for demonstration) only two choices for species selection, the H_2-air autoignition evidence was evaluated in full for all options. For a large mechanism with a broader pool of options, the cost becomes prohibitive. Here, we develop methods to provide similar species selection and illustrate them for methane-air autoignition, modeled with the GRI-1.2 mechanism [65,66], which has 30 reactive species and 177 reactions. Throughout, we chose $T_{\text{ign}} = 2100$ K in (2) to define the ignition time.

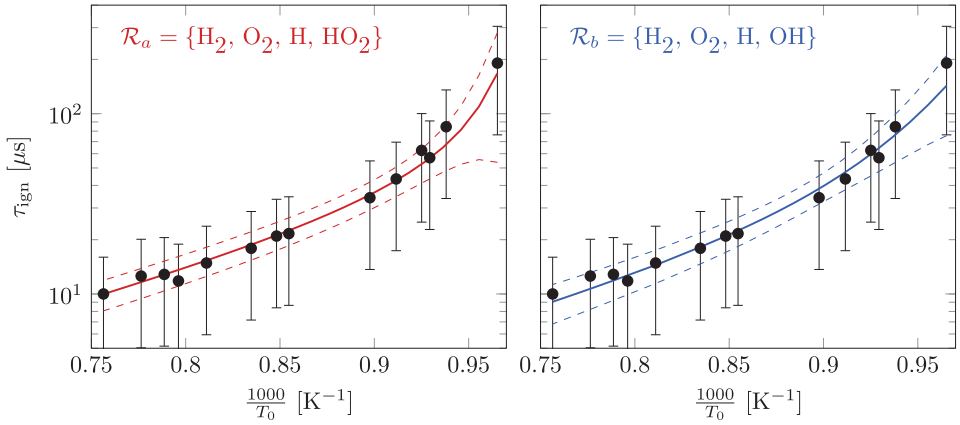
The raw number of candidate sets of represented species is immense. To represent K species, there are $\binom{N}{K}$ candidates, so for GRI-1.2 and only $K = 7$, there are over 2 million candidate sets. Increasing computational power will alleviate this, but will not soon provide a full species selection solution. To overcome this challenge, a CSP time-scale ranking of candidate species is used to narrow the pool, such that options anticipated to be unimportant are discarded outright [17,18]. There is no longer proof that the best possible model is found, but the procedure finds the best amongst many options that would be considered viable based on simple time-scale analysis, and it is possible to show convergence for the



(a) Small measurement $\sigma'_{ign} = 0.1$ and rate $f = 1.1$ uncertainty: $P(\mathcal{R}_a \,|\, \boldsymbol{\tau}^d_{ign}, I) > P(\mathcal{R}_b \,|\, \boldsymbol{\tau}^d_{ign}, I)$.



(b) Small measurement $\sigma'_{ign} = 0.05$ but large rate $f = 2.35$ uncertainty: $P(\mathcal{R}_a \,|\, \boldsymbol{\tau}^d_{ign}, I) > P(\mathcal{R}_b \,|\, \boldsymbol{\tau}^d_{ign}, I)$.



(c) Large measurement $\sigma'_{ign} = 0.3$ and rate $f = 2.35$ uncertainty: $P(\mathcal{R}_a \,|\, \boldsymbol{\tau}^d_{ign}, I) < P(\mathcal{R}_b \,|\, \boldsymbol{\tau}^d_{ign}, I)$.

Fig. 4. Reduced-model predictions (27) for selected values of f and σ'_{ign} as labeled in Fig 3: measurements •, with error bars representing $2\sigma_{ign}$; ——— reduced-model predictions $\langle \tau_{ign} \rangle$ (27); and — — — predictive uncertainty $\pm 2\hat{\sigma}_{ign}$ (28).

number of candidates considered. The resulting species selection problem, while still challenging, is tractable.

5.2. CSP constraint pointers

CSP [67,68] provides a way to separate the full model, constrained by N_e elements, into N_r rapid modes that quickly exhaust and collapse subsequent evolution onto a low-dimensional manifold, and $N - N_r - N_e$ slow modes that evolve the system on this manifold. The rate-of-change vector (3) is decomposed accordingly as

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}_r + \boldsymbol{\Omega}_s = \sum_{r=1}^{N_r} \varphi^r \boldsymbol{a}_r + \sum_{s=N_r+1}^{N-N_e} \varphi^s \boldsymbol{a}_s, \tag{36}$$

where $\boldsymbol{a}_r$ and $\boldsymbol{a}_s$ are the basis vectors of the fast and slow subspaces, respectively, approximated by the eigenvectors of the Jacobian matrix $\partial_{\boldsymbol{z}}\boldsymbol{\Omega}$. The amplitude of the mth mode is

$$\varphi^m = \boldsymbol{b}^m \cdot \boldsymbol{\Omega}, \tag{37}$$

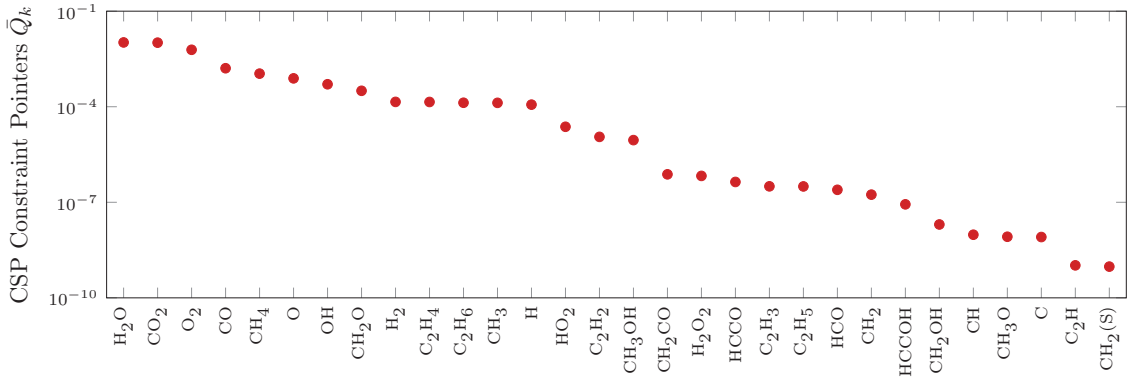


Fig. 5. CH_4–air autoignition CSP constraint pointers $\bar{Q}_k$ (42) for $\phi = 1$, $p = 1$ atm, and $N_{\text{ign}} = 11$ initial temperatures uniform in the range 1500 K $\leq T_0 \leq$ 1700 K.

with $\boldsymbol{b}^m \cdot \boldsymbol{a}_m = 1$. The condition that the fast modes are unimportant for evolution on the manifold leads to N_r nonlinear constraints

$$\varphi^r = \boldsymbol{b}^r \cdot \boldsymbol{\Omega} = 0 \qquad \text{for} \quad r = 1, \ldots, N_r, \tag{38}$$

which specify the manifold [7]. In practice, the number of exhausted modes N_r is selected to satisfy a tolerance threshold

$$\left| \sum_{r=1}^{N_r} \tau_r a_r^j \varphi^r \right| < \text{tol}_{rel}|z_j| + \text{tol}_{abs}, \qquad \text{for} \quad j = 1, \ldots, N, \tag{39}$$

where $\tau_r = 1/|\lambda_r|$ is the rth fastest time scale, based on the rth eigenvalue λ_r of the Jacobian matrix $\partial_{\boldsymbol{z}}\boldsymbol{\Omega}$. The tolerances tol_{rel} and tol_{abs} are relative and absolute tolerances. While tol_{rel} is dimensionless, tol_{abs} matches the units of $\boldsymbol{z}$ (kmol/kg). A new version of the exhaustion criterion (39) provides a less conservative (larger) estimate of N_r [69], which could allow for a sharper distinction between "slow" and "fast" species. However, (39) is sufficient for sorting species prior to Bayesian selection.

The contribution of the slow modes to the kth species is determined by the projection

$$\mathscr{Q}_k = \left| \sum_{s=N_r+1}^{N-N_e} a_s^k b_k^s \right|. \tag{40}$$

The resulting *constraint pointer* $\mathscr{Q}_k$ is large if this species evolves slowly. Standard practice is to weigh $\mathscr{Q}_k$ by the mass fractions y_k to avoid selecting unimportant trace species that would be designated slow by this measure [18], and since conditions evolve, this is usually determined over a relevant time period $t \in [0, t_{\text{final}}]$,

$$Q_k = \frac{1}{t_{\text{final}}} \int_0^{t_{\text{final}}} \mathscr{Q}_k(t) y_k(t)\, \mathrm{d}t, \tag{41}$$

where Q_k then serves as the global constraint pointer for the k^{th} species. Here, t_{final} is the time for which $N_r = N - N_e$ satisfies (39), indicating that all modes have exhausted.

A complexity is that the standard Q_k pointers depend on the particular initial condition (T_0, p, ϕ), though calibration data for species selection will in general cover a range of conditions. Fortunately, this dependence is often weak, and our species selection procedure is insensitive to minor changes in the ordering. To account for the range of conditions, the pointers are simply averaged over the N_{ign} calibration scenarios,

$$\bar{Q}_k = \frac{1}{N_{\text{ign}}} \sum_{i=1}^{N_{\text{ign}}} Q_k^{(i)}, \tag{42}$$

where $Q_k^{(i)}$ is the global pointer for the kth species based on the ith ignition scenario. This is similar in spirit to the time-averaged mass-fraction-weighing in (41). A similar challenge is that the pointers also depend on uncertain kinetic parameters $\boldsymbol{\kappa}$, so the pointers are also random variables. However, in none of our examples is the uncertainty in any rate parameter so large to substantially upset the CSP species ordering. The range of time scales for CSP ordering far exceeds rate uncertainties. Thus, the CSP pointers (41), as used for ordering species for Bayesian selection, do not retain their uncertainty. Instead, they are simply evaluated at the nominal values of rate coefficients.

Nominal mean CSP constraint pointers $\bar{Q}_k$ (42) are shown in Fig. 5, for $\text{tol}_{\text{abs}} = 10^{-13}$ and $\text{tol}_{\text{rel}} = 10^{-5}$ in (39). Species ordering changes only in minor ways—usually switching the order of neighbors with nearly identical $\bar{Q}_k$—for any tested $\text{tol}_{abs} \in \left[10^{-10}, 10^{-13}\right]$ kmol/kg and $\text{tol}_{rel} \in \left[10^{-3}, 10^{-5}\right]$.

5.3. CSP-guided Bayesian species selection

To initialize species selection, we pick the fuel and oxidized $\mathcal{R}_0 = \{CH_4, O_2\}$ as obviously important species to be the initial set. The algorithm then proceeds in stages, each adding one represented species in two steps:

1. *Candidate selection sets*: At stage m, with represented $\mathcal{R}_m$, the unrepresented species $\mathcal{U}_m$ are sorted in descending order with respect to their CSP constraint pointer, the first being the slowest. We form a list of n_c candidate species sets $\mathbb{A}_m$ by augmenting $\mathcal{R}_m$ with each of the first n_c species in $\mathcal{U}_m$,
$$\mathbb{A}_m = \bigcup_{p=1}^{n_c} \{\mathcal{R}_m \cup \{s_p\}\}, \quad s_p \in \mathcal{U}_m. \tag{43}$$
The search depth n_c is thus the number of options considered in each stage. For the example in Fig. 5 and $n_c = 4$, the first candidate species to augment $\mathcal{R}_0$ are H_2O, CO, CO_2, and O. The cost of the search is thus proportional to n_c; multiple n_c values are consider in the following demonstrations.
2. *Species selection*: For every candidate in $\mathbb{A}_m$, we compute the posterior probability (11). This requires model priors $P(\mathcal{R}_m^p \,|\, I)$, though without any reason here to favor any of the candidates, their priors are uniform, so the posterior probabilities depend only on the evidence (12). The current set of represented species $\mathcal{R}_m$ is then augmented with the species with largest evidence, denoted by s^*:
$$\mathcal{R}_{m+1} = \mathcal{R}_m \cup \{s^*\}. \tag{44}$$

As a search algorithm, the approach is designated as greedy since it makes locally optimal choices [70]; in this sense, it is similar to the algorithm proposed by Hiremath et al. [15] for RCCE species selection for trusted rates. The key difference, in addition

to limiting the number of candidate species using constraint pointers, is that it accounts for multiple uncertain factors using Bayesian evidence. It is demonstrated in the following section.

6. Demonstration: CH_4 - Air Autoignition

We continue to use the GRI-1.2 mechanism to model reactions, and assume all of its rate coefficients are uncertain and normally distributed about their nominal values, per (19) and (20). Concurrently, we illustrate a case in which no experimental measurements are available to inform species selection, and present a procedure to generate calibration data using the full model in Section 2.1. In the H_2-air demonstration of Section 4, an additive likelihood (16) was used for simplicity. Here, a multiplicative likelihood (18) is used instead because, as we shall see, it provides a clearer comparison between the generated calibration data and reduced-model predictions $\tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{R})$.

6.1. Calibration data

The model reduction procedure—the main goal of this paper—is the same irrespective of the source of the data; the approach only depends on the goals and available information for any particular application. If the goal is predicting ignition, say in a turbulent flow, and there is adiabatic reactor data available for similar conditions, then there is an opportunity to tailor the reduced model based on that data within this framework. For the H_2-air autoignition example, we used experimental data τ_{ign}^d to select species and re-calibrate uncertain rate coefficients. If the important conditions for an application are less clear, or if auxiliary data is unavailable, direct generation of calibration data is a viable alternative. In such cases, the full model described in Section 2.1, which despite the uncertainty in the rates is expected to be of higher fidelity than any reduction because it resolves all time scales, can be used to generate data to calibrate reduced models. Since there is no experimental uncertainty, in this case σ_{ign} stems from the underlying kinetic uncertainty f_i of the mechanism, for which we take established values [20].

Even for model-generated data, the first step is selecting conditions representative of the prediction scenario. For illustration, we again take $\phi = 1$, $p = 1$ atm, and $1500\,\text{K} \le T_0 \le 1700\,\text{K}$. Propagation of kinetic uncertainties, given by (19) and (20), through the full model (3) yields ignition time predictions (2) with uncertainty. Ignition times are expected to be approximately exponentially sensitive to at least some kinetic parameters $\boldsymbol{\kappa}$. Therefore, it is not surprising that full-model predictions $\tau_{\text{ign}}^{\text{full}} \equiv \tau_{\text{ign}}(\boldsymbol{\kappa}, \mathcal{S})$, for full set of species $\mathcal{S}$, span several orders of magnitude, as shown in Fig. 6. Using the logarithm of ignition time thus make a more natural metric. Each data point generated with the full model is taken to be the expectation

$$\mu_{\text{ign}}^g = \int \log\left(\frac{\tau_{\text{ign}}^{\text{full}}}{\tau_{\text{ref}}}\right) P(\boldsymbol{\kappa} \mid I)\, \mathrm{d}\boldsymbol{\kappa}, \tag{45}$$

where $\tau_{\text{ref}} = 1\,\mu\text{s}$. The uncertainty in $\tau_{\text{ign}}^{\text{full}}$ is the corresponding standard deviation,

$$\sigma_{\text{ign}}^g = \left\{ \int \left[\mu_{\text{ign}}^g - \log\left(\frac{\tau_{\text{ign}}^{\text{full}}}{\tau_{\text{ref}}}\right)\right]^2 P(\boldsymbol{\kappa} \mid I)\, \mathrm{d}\boldsymbol{\kappa} \right\}^{\frac{1}{2}}. \tag{46}$$

As for the evaluation of evidence integrals (12), the 177 uncertain rate coefficients in the GRI-1.2 mechanism preclude tensor-product quadrature for (45) and (46), so Monte Carlo integration is used. In this case, it entails averaging full-model evaluations over

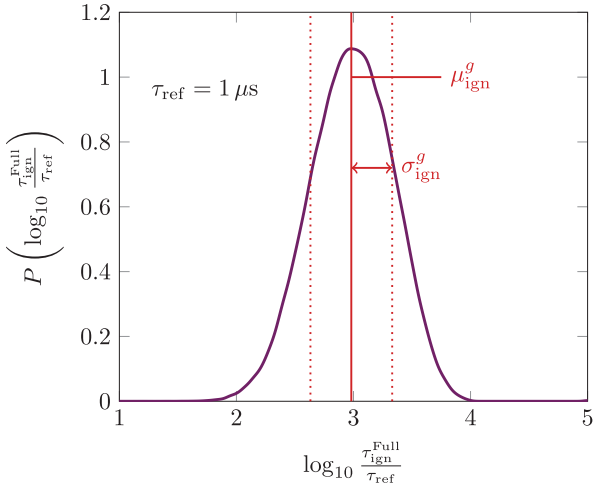


Fig. 6. Ignition-time distribution, based on the parametric uncertainty of the GRI-1.2 mechanism [20], for $\phi = 1$, $p = 1$ atm, and $T_0 = 1500$ (K).

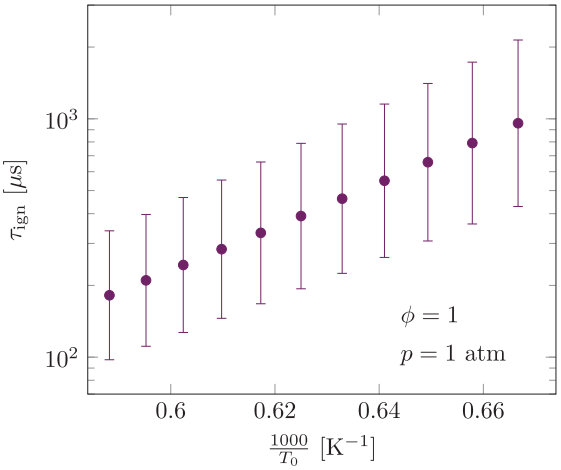


Fig. 7. Full-model ignition times for CH_4-air for $\phi = 1$, $p = 1$ atm, and initial temperatures $1500\,\text{K} \le T_0 \le 1700\,\text{K}$. The error bars represent full-model data uncertainty (46).

N_{MC} samples of the uncertain rate parameters $\boldsymbol{\kappa}$ from their distribution (20),

$$\mu_{\text{ign}}^g \approx \frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} \log\left[\frac{\tau_{\text{ign}}\left(\boldsymbol{\kappa}^{(i)}, \mathcal{S}\right)}{\tau_{\text{ref}}}\right], \tag{47}$$

$$\sigma_{\text{ign}}^g \approx \left\{ \frac{1}{N_{\text{MC}} - 1} \sum_{i=1}^{N_{\text{MC}}} \left[\mu_{\text{ign}}^g - \log\left(\frac{\tau_{\text{ign}}\left(\boldsymbol{\kappa}^{(i)}, \mathcal{S}\right)}{\tau_{\text{ref}}}\right)\right]^2 \right\}^{\frac{1}{2}}, \tag{48}$$

where $\boldsymbol{\kappa}^{(i)}$ is the ith sample from $P(\boldsymbol{\kappa} \mid I)$.

The calibration data, shown in Fig. 7, is generated for $N_{\text{ign}} = 11$ uniform initial temperatures in the range $1500\,\text{K} \le T_0 \le 1700\,\text{K}$. Following the same procedure for establishing convergence of Monte Carlo integration in Section 4.3 with $M_{\text{MC}} = 15$, the sampling error of μ_{ign}^g and σ_{ign}^g is less than 1% for $N_{\text{MC}} = 73 \times 10^3$. The full model was evaluated using CVode [61]; results are presented here for relative $\text{E}_{\text{rel}} = 10^{-8}$ and absolute $\text{E}_{\text{abs}} = 10^{-13}$ kmol/kg tolerances, and were insensitive to these.

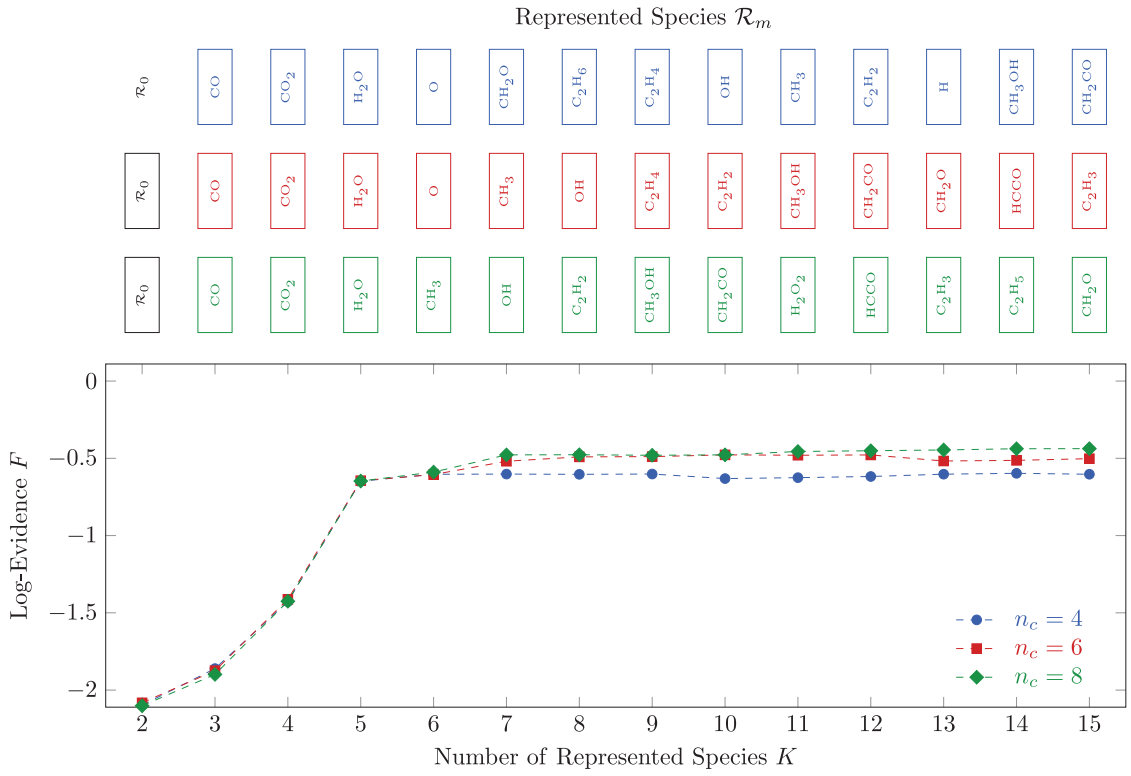


Fig. 8. Selected species and their evidence (12) for $\mathcal{R}_0 = \{CH_4, O_2\}$ and different n_c.

6.2. Numerical evaluation

The evidence integrals (12) for CH_4-air autoignition are evaluated using Monte Carlo sampling, as described in Section 4.3. Results are presented for $N_{MC} = 73 \times 10^3$ with an estimated sampling error of less than 1% for $M_{MC} = 5$. As for the H_2-air example, predictions (27) and predictive uncertainties (28) are computed using multi-level sampling [59]. The key difference is that, to speed up convergence, the dimension of the uncertain-parameter space is reduced through global sensitivity analysis [71,72], and the Markov chain is constrained to a small subset of active parameters (identified a priori), with all others held at their nominal values. Results are insensitive to the number of samples per level for 5×10^3, 10^4, and 2×10^4 samples.

6.3. Results

Selected species for K up to 15 and their supporting evidence are shown in Fig. 8 for the initial set of represented species $\mathcal{R}_0 = \{CH_4, O_2\}$ and how this is augmented for search-depths $n_c = 4$, 6, and 8 in (43). Note that all these n_c values yield the same next three species: CO, then CO_2, and finally H_2O. For $K \geq 5$, species selection depends on n_c, though the evidence is only weakly sensitive to these differences. The overall cost is proportional to n_c, but the algorithm searches deeper into the species ranking (Fig. 5). This search depth affects, for example, how early CH_3 is selected, which leads to modestly more plausible sets of represented species, at least once it is complemented with O or OH. This is not unexpected, given the understood controlling role that the methyl radical CH_3 has on methane ignition [19], though its additional supporting evidence in the reduced model is marginal.

In all the cases in Fig. 8, the evidence grows rapidly up to $K = 5$ and is nearly constant thereafter. The eventual peak K only weakly depends on n_c: $K = 6$ for $n_c = 4$; $K = 8$ for $n_c = 6$; and $K = 7$ for $n_c = 8$. We denote this reduced-model size by K^*, and the corresponding set of represented species by $\mathcal{R}^*$. Subsequent addition of species to $\mathcal{R}^*$ is not significantly supported by the data, and all $\mathcal{R}^*$ reproduce full-model ignition time predictions (45) within their predictive uncertainty (46), as seen in Fig. 9. This balance of fit-to-data and kinetic uncertainty is accomplished without heavy reliance on the re-calibration of the uncertain rate coefficients, which would be penalized by the Occam factors (22).

Both nominal and re-calibrated reduced-model predictions (27) are compared to full-model data for a representative scenario in Fig. 10. With calibration, all reduced models with evidence-based selection of species can reproduce the full-model data well within its uncertainty, though those with $K = 2$ to 4 rely heavily on the re-calibration described in Section 3.4 to compensate for their poor nominal fit-to-data. This dependence on data is penalized by the Occam factors in (22), which suppresses their plausibility in Fig. 8. This illustrates how the framework guards against over-reduction through the exploitation of the uncertainty of the parameters. A more focused study on the effects of kinetic uncertainties on RCCE model size follows in Section 7.

The ultimate purpose of model reduction is to reduce computational cost. This is quantified in Fig. 11, based on the time of one evaluation for a representative calibration scenario and nominal kinetic parameters. In general, the larger n_c, the more expensive the resulting reduced model, since larger n_c enables faster species, as quantified by the CSP pointers, to be selected. Indeed, for our CH_4-air example, despite the marginal added evidence, species selected with $n_c = 8$ substantially increase the computational cost for $K \geq 7$, which coincides with the leveling-off of the evidence (Fig. 8). This is due to the rejection of key slowly-evolving species O in favor of radicals that evolve on CSP fast time scales. The species selection framework provides the relative evidence supporting a model,

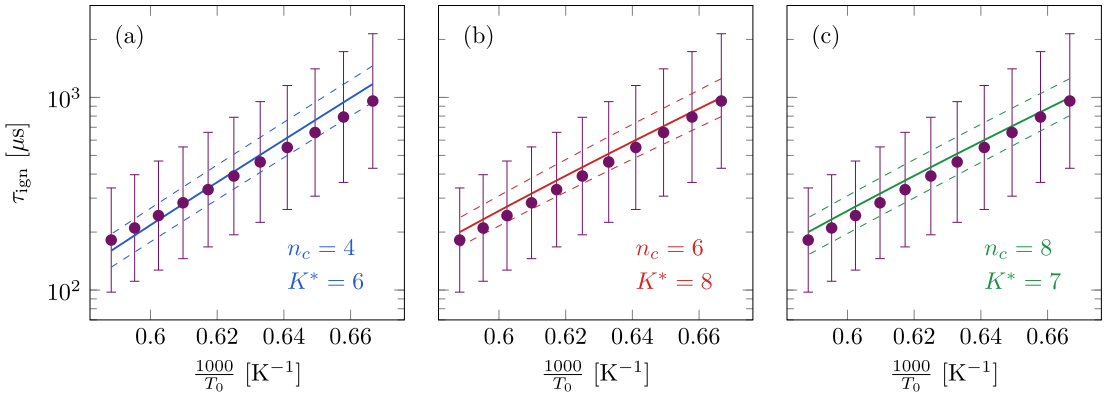


Fig. 9. Reduced-model predictions (27) using selected species $\mathcal{R}^*$ for (a) $n_c = 4$, (b) 6, and (c) 8 in Fig. 8: _____ reduced-model predictions $\langle\tau_{\text{ign}}\rangle$ (27); – – – predictive uncertainty $\pm\hat{\sigma}_{\text{ign}}$ (28); and full-model data •, with error bars representing its corresponding uncertainty (46).

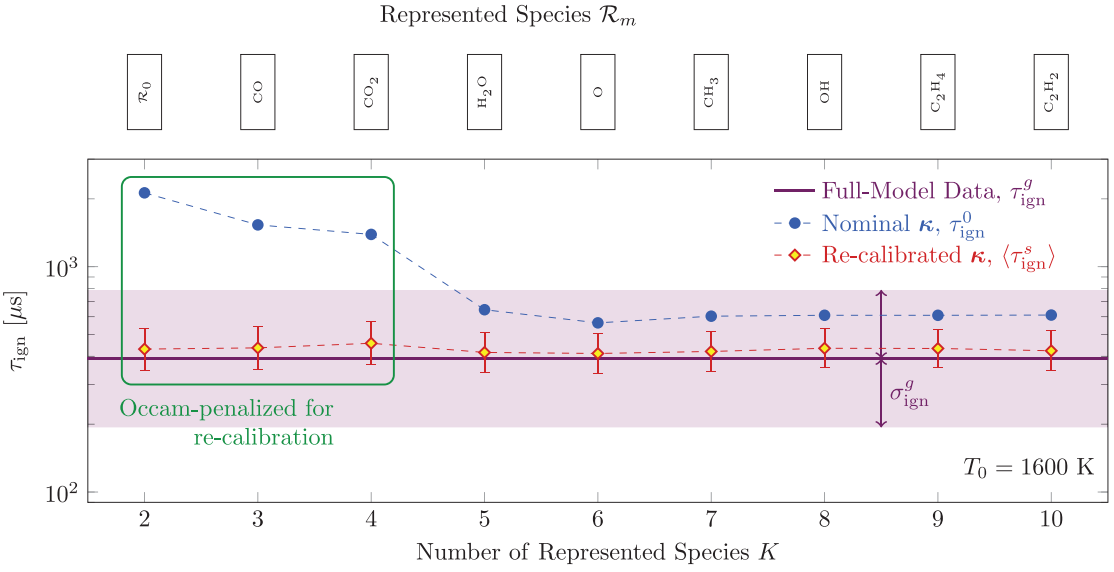


Fig. 10. Nominal • and re-calibrated ◇ (27) reduced-model predictions for $T_0 = 1600$ K and $n_c = 6$. The solid line _____ and shaded region represent the full-model data τ^g_{ign} (45) and ignition-time standard deviation uncertainty $\pm\sigma^g_{\text{ign}}$ (46). The error bars represent reduced-model predictive uncertainty $\pm\hat{\sigma}_{\text{ign}}$ (28).

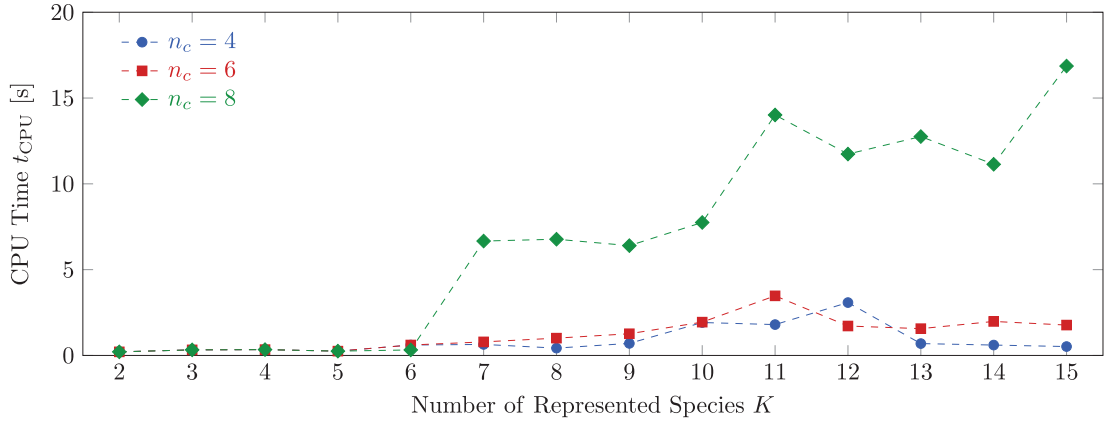


Fig. 11. Computation time of a reduced-model evaluation for ignition time for $\phi = 1$, $p = 1$ atm and $T_0 = 1600$ K using nominal kinetic parameters. Represented species match those in Fig. 8. For this example, CVode [61] was used with dimensionless $E_{\text{rel}} = 10^{-8}$ and $E_{\text{abs}} = 10^{-13}$ kmol/kg.

which can then be weighed against the cost of including additional species, especially if they are faster.

The plausible sets of represented species $\mathcal{R}_m$ in Fig. 8 include full and partial products CO_2, H_2O and CO, and key radicals CH_3, O and OH, along with the reactants. These are commonly selected in established species selection methodologies under the assumption that the GRI-1.2 is fully trusted [15,18,73]. Here, this was asserted quantitatively without disregarding rate-parameter uncertainty based on the evidence computation, such that the benefits of more costly models can be weighed with confidence. This

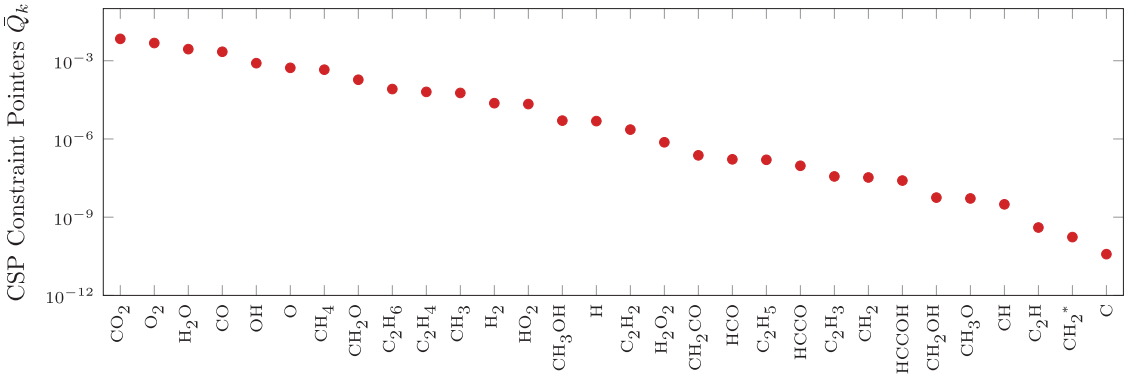


Fig. 12. CH_4–air autoignition CSP constraint pointers $\bar{Q}_k$ (42) for the range of initial conditions in Table 1.

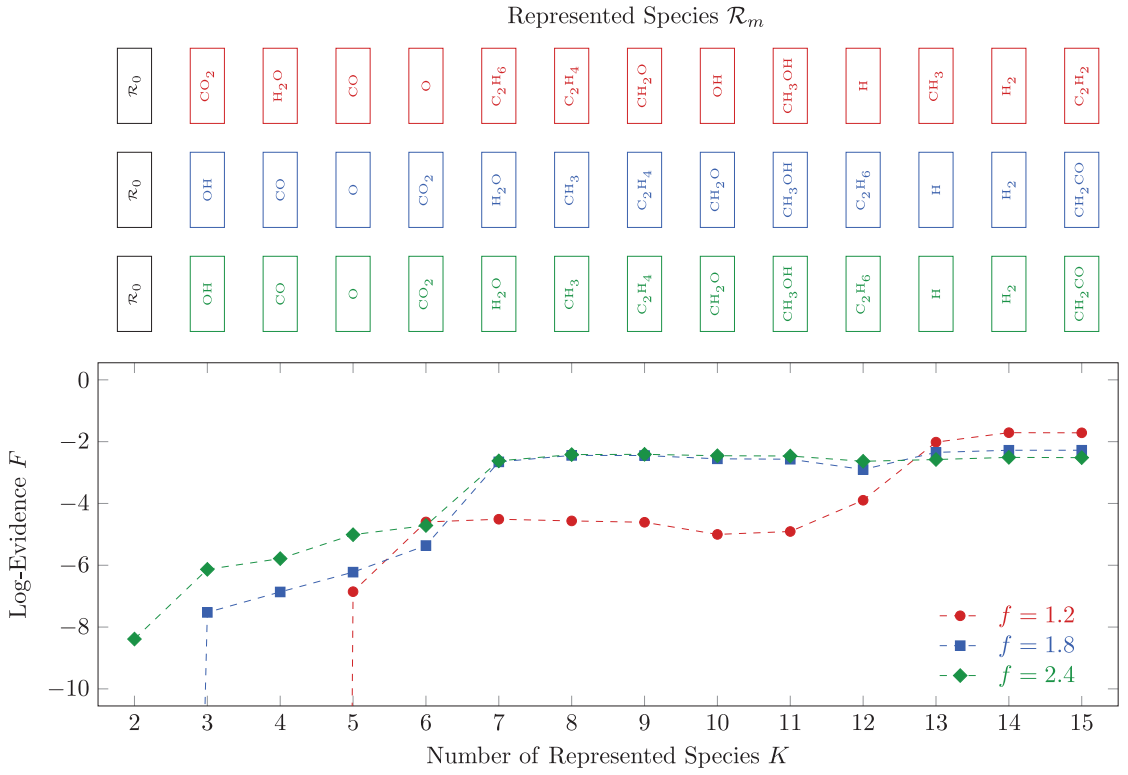


Fig. 13. Selected species and their evidence for $\mathcal{R}_0 = \{CH_4, O_2\}$, $n_c = 4$, and different global uncertainty factors f in (19).

is one of the key contributions of this framework. The other is that it can be applied to less understood mechanisms, where rate kinetics might be less certain and key pathways unclear.

7. Kinetic uncertainty and model size

As for the H_2-air autoignition example of Section 4, additional experimental data can be used to target a particular regime of application, or guide the determination of reduced model size (and cost). This is considered in this section.

7.1. Data & uncertainty

The ignition-time measurements of Zhukov et al. [74] in Table 1 are used for this demonstration with the additive error model (14). The provided errors estimates for these data are no larger than 10%, though to facilitate convergence and focus on the impact of kinetic uncertainties, we simply take $\sigma_{\text{ign},i} = \sigma'_{\text{ign}} \tau^d_{\text{ign},i}$ in (16), with $\sigma'_{\text{ign}} = 0.1$. As in Section 4, all reactions in the GRI-1.2 are ascribed a common uncertainty factor $f_i = f$ in (19), which we will vary to illustrate its impact on selected model size.

Table 1
Ignition-time measurements [74] in lean ($\phi = 0.5$) CH_4-Air mixtures used for rate coefficient re-calibration and species selection.

T_0 [K]	p [atm]	τ^d_{ign} [μs]
1536	3.42	295
1537	3.21	262
1584	3.44	189
1585	3.86	195
1644	3.76	100
1653	3.16	101

7.2. CSP species ordering

Following the procedure in Section 5.2, the CSP constraints pointers are shown in Fig. 12, with results that are independent of the CSP tolerances $\text{tol}_{\text{abs}} = 10^{-13}$ and $\text{tol}_{\text{rel}} = 10^{-5}$ in (39) For this illustration, we use $n_c = 4$.

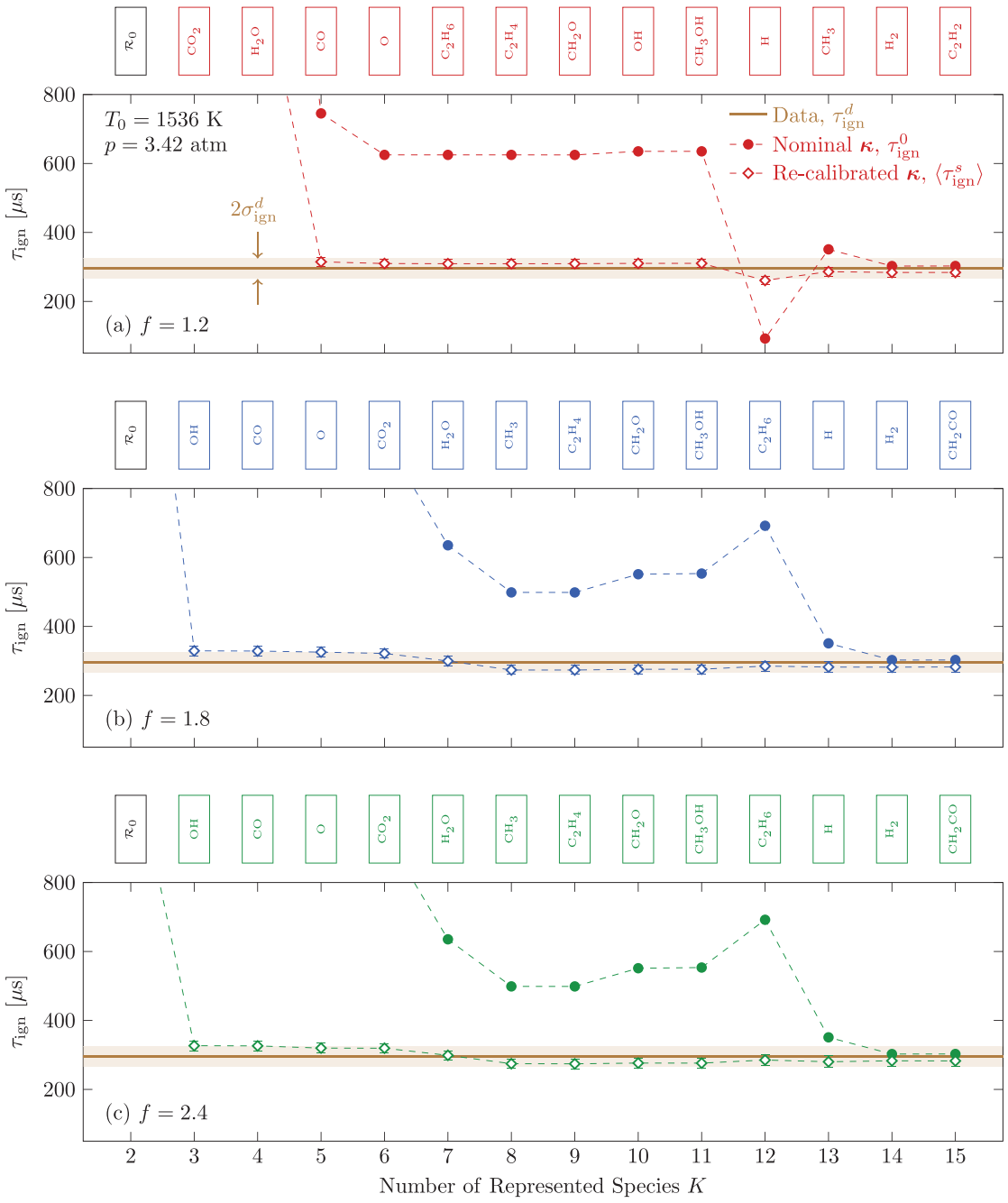


Fig. 14. Nominal • and re-calibrated ◇ (27) reduced-model predictions for the first calibration scenario in Table 1 ($T_0 = 1536$ K and $p = 3.42$ atm) using species selected in Fig. 13 for parametric uncertainty factors (a) $f = 1.2$, (b) 1.8, and (c) 2.4. The solid line ——— and shaded region represent the measurement and its σ^d_{ign} uncertainty. The error bars on re-calibrated predictions (mostly smaller than the symbols) represent predictive uncertainty $\hat{\sigma}_{\text{ign}}$ (28).

7.3. Results

Selected species up to $K = 15$ and their evidence for different values of uncertainty factor f are shown in Fig. 13. Species ordering differs substantially between $f = 1.2$ and $f = 1.8$, but is identical for $f = 1.8$ and $f = 2.4$. The evidence also differs most between $f = 1.2$ and $f = 1.8$, being insensitive to f for $f > 1.8$. In general, selected species are consistent with the analogous $n_c = 4$ case for full-model data of Section 6.3.

For small kinetic-parameter uncertainty $f = 1.2$, addition of species is supported up until $K^* = 13$. However, for $f > 1.2$, the plausibility of models with $7 \le K \le 12$ is comparable to that of $K \ge 13$ for $f = 1.2$. It also saturates at $K^* = 7$. For $f = 2.4$, species addition is supported only up to $K^* = 8$. In this case, the greater kinetic uncertainty allows for a less restrictive re-calibration of $\boldsymbol{\kappa}$, and thus weaker Occam factors in (22). Predictions, shown in Fig. 14, confirm this: they match the data closely for $K \ge 5$, but reduced models with $5 \le K \le 12$ rely heavily on re-calibration. Thus, they are Occam-penalized, though less so for broader kinetic-parameter uncertainties. These results quantitatively support the notion that fewer species need to be represented when the kinetic mechanism is less trusted. For the general RCCE user, the predictive accuracy of models selected with our framework is at least that for simply accepted rates, and

likely significantly better in some cases as we see in some of our demonstrations.

8. Summary & additional discussion

In summary, we have introduced a Bayesian framework for RCCE species selection under kinetic uncertainty. Its principal strength is that it guards against overfitting and, consequently, lost confidence in the predictive (extrapolative) capacity of the reduced model. Selected species balance fit-to-data against Occam factors. We demonstrated these basic properties on a simple H_2-air autoignition problem, where we showed how kinetic and measurement uncertainties affect the plausibility of our choices. We then extended our framework to more complex mechanisms by limiting the number of candidate species through CSP pointers, and demonstrated this on CH_4-air autoignition. Specifically, we considered two cases that differ in the source of calibration data: (i) one where relevant data is missing, so it is generated from the full model; and (ii) one where experimental data is used to tailor species selection to relevant scenarios and assess the role of kinetic-parameter uncertainties in model size. The results are mostly consistent with what might be expected for methane, though obtained with firm quantitative basis. The real strength of the method would be realized for still more complex and less-understood chemical models.

The key component of the approach is that quantitative evidence (12), accounting for different sources of input uncertainty, guides reduced-model selection while guarding against overfitting through the Occam factors (22). This way the selected species and reduced models are expected to be accurate away from the calibration scenarios. Possible extensions of this framework might entail (i) the inclusion other combustion quantities of interest and broader ranges of calibration scenarios, (ii) accounting in more detail for the uncertainty of the Arrhenius parameters, (iii) the inclusion other uncertainties such as thermodynamic properties, and (iv) the use of linear combinations of species as candidate constraints. This last extension is straightforward, as our framework can easily select from among these type of constraints. In our H_2-air autoignition demonstration, for example, one may also assess the plausibility of a third candidate reduced model, say $\mathcal{R}_c = \{H_2, O_2, H, O + OH + H_2O\}$ and compare it against $P(\mathcal{R}_a \,|\, \boldsymbol{\tau}^d_{\text{ign}}, I)$ and $P(\mathcal{R}_b \,|\, \boldsymbol{\tau}^d_{\text{ign}}, I)$, so long as the constraints $\mathcal{R}$ are linearly independent.

It should be recognized that this species selection procedure involves significant up-front computation. To estimate evidence integrals for the methane application of Section 7, over 60 million ignition-time predictions were computed. Although this is obviously far more costly than just making a best-fit, it can be justified for the information it provides. This is especially true if the reduced model is to be used in a flow-coupled simulation, the computational cost of which can easily dwarf that of species selection. Advancing computational resources will also, of course, continue to make this up-front cost less and less of a concern, allow for the evaluation of more possible choices, and enable the incorporation of more parametric uncertainties and data. However, the evaluation of evidence integrals (12) for larger data sets, different quantities of interest and more uncertain parameters can still result in a computationally taxing inference problem. In this sense, the framework could benefit from the integration with the method of uncertainty minimization (MUM) [45], an efficient simplification of the more general Bayesian methodology. In MUM, the combustion models would be approximated with surrogates, and the posterior PDF of the parameters (21) would be approximated as Gaussian, with mean $\boldsymbol{\kappa}^*$ and covariance matrix $\boldsymbol{\Xi}_*$. The log-evidence (22) would then be available analytically in terms of the approximate posterior statistics,

$$U(\mathcal{R},\, \boldsymbol{\eta}^d) \approx -\frac{1}{2}\sum_{i=1}^{N_\eta}\left[\frac{\eta_i^d - \eta_i(\boldsymbol{\kappa}^*, \mathcal{R})}{\sigma_{\eta,i}}\right]^2, \tag{49}$$

$$S(\mathcal{R},\, \boldsymbol{\eta}^d) \approx \frac{1}{2}\|\boldsymbol{\kappa}^*\|^2 + \frac{1}{2}\log\left|\boldsymbol{\Xi}_*^{-1}\right|, \tag{50}$$

where $\boldsymbol{\eta}^d$ denotes calibration data for a combustion quantity of interest. This form for (50) is identical to the linearization (26) we analyzed as a model Occam factor.

The ultimate cost of reduced-model evaluation was not directly included, though it is also useful for informing decisions in face of cost concerns. The cost of RCCE depends on the time scales and number of represented species, so quantifying the supporting evidence of selecting any particular K species can guide subsequent decisions based on cost. Perhaps most importantly, it also asserts when any cost of increasing K is unjustified. The final methane example of Section 7 incorporated additional experimental data with this as its primary purpose.

Acknowledgments

This material is based in part upon work supported by the Department of Energy, National Nuclear Security Administration, under Award Number DE-NA0002374. ECG also acknowledges partial fellowship support from the National Council of Science and Technology of Mexico (CONACyT).

References

[1] W. Sun, X. Gou, H.A. El-Asrag, Z. Chen, Y. Ju, Multi-timescale and correlated dynamic adaptive chemistry modeling of ignition and flame propagation using a real jet fuel surrogate model, Combust. Flame 162 (4) (2015) 1530–1539.
[2] T. Lu, C.K. Law, Toward accommodating realistic fuel chemistry in large-scale computations, Prog. Energy Combust. Sci. 35 (2) (2009) 192–215.
[3] U. Maas, S.B. Pope, Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space, Combust. Flame 88 (3) (1992) 239–264.
[4] P. Pepiot-Desjardins, H. Pitsch, An efficient error-propagation-based reduction method for large chemical kinetic mechanisms, Combust. Flame 154 (1–2) (2008) 67–81.
[5] T. Lu, C.K. Law, A directed relation graph method for mechanism reduction, Proc. Combust. Inst. 30 (1) (2005) 1333–1341.
[6] X.L. Zheng, T.F. Lu, C.K. Law, Experimental counterflow ignition temperatures and reaction mechanisms of 1, 3-butadiene, Proc. Combust. Inst. 31 (1) (2007) 367–375.
[7] M. Valorani, F. Creta, D.A. Goussis, J.C. Lee, H.N. Najm, An automatic procedure for the simplification of chemical kinetic mechanisms based on CSP, Combust. Flame 146 (1–2) (2006) 29–51.
[8] N. Peters, B. Rogg, Reduced kinetic mechanisms for applications in combustion systems, Lecture Notes in Physics, 15, Springer-Verlag, Berlin, 1993.
[9] J.C. Keck, D. Gillespie, Rate-controlled partial-equilibrium method for treating reacting gas mixtures, Combust. Flame 17 (2) (1971) 237–241.
[10] P. Pepiot-Desjardins, H. Pitsch, An automatic chemical lumping method for the reduction of large chemical kinetic mechanisms, Combust. Theory Model. 12 (6) (2008) 1089–1108.
[11] T. Lu, C.K. Law, Strategies for mechanism reduction for large hydrocarbons: n-heptane, Combust. Flame 154 (1–2) (2008) 153–163.
[12] Q. Tang, S.B. Pope, A more accurate projection in the rate-controlled constrained-equilibrium method for dimension reduction of combustion chemistry, Combust. Theory Model. 8 (2) (2004) 255–279.
[13] V. Yousefian, A rate-controlled constrained-equilibrium thermochemistry algorithm for complex reacting systems, Combust. Flame 115 (1–2) (1998) 66–80.
[14] S. Rigopoulos, T. Løvås, A LOI–RCCE methodology for reducing chemical kinetics, with application to laminar premixed flames, Proc. Combust. Inst. 32 (1) (2009) 569–576.
[15] V. Hiremath, Z. Ren, S.B. Pope, A greedy algorithm for species selection in dimension reduction of combustion chemistry, Combust. Theory Model. 14 (5) (2010) 619–652.
[16] G.P. Beretta, M. Janbozorgi, H. Metghalchi, Degree of disequilibrium analysis for automatic selection of kinetic constraints in the rate-controlled constrained-equilibrium method, Combust. Flame 168 (2016) 342–364.
[17] Z. Ren, Z. Lu, Y. Gao, T. Lu, L. Hou, A kinetics-based method for constraint selection in rate-controlled constrained equilibrium, Combust. Theory Model. 21 (2) (2017) 159–182.
[18] P. Koniavitis, S. Rigopoulos, W.P. Jones, A methodology for derivation of RCCE-reduced mechanisms via CSP, Combust. Flame 183 (2017) 126–143.

[19] C.K. Westbrook, F.L. Dryer, Chemical kinetic modeling of hydrocarbon combustion, Prog. Energy Combust. Sci. 10 (1) (1984) 1–57.
[20] D.A. Sheen, X. You, H. Wang, T. Løvås, Spectral uncertainty quantification, propagation and optimization of a detailed kinetic model for ethylene combustion, Proc. Combust. Inst. 32 (1) (2009) 535–542.
[21] H. Wang, D.A. Sheen, Combustion kinetic model uncertainty quantification, propagation and minimization, Prog. Energy Combust. Sci. 47 (2015) 1–31.
[22] R.M. Galassi, M. Valorani, H.N. Najm, C. Safta, M. Khalil, P.P. Ciottoli, Chemical model reduction under uncertainty, Combust. Flame 179 (2017) 242–252.
[23] Y. Xin, D.A. Sheen, H. Wang, C.K. Law, Skeletal reaction model generation, uncertainty quantification and minimization: combustion of butane, Combust. Flame 161 (12) (2014) 3031–3039.
[24] E.T. Jaynes, Probability theory: the logic of science, Cambridge University Press, 2003.
[25] D. Sivia, J. Skilling, Data analysis: a Bayesian tutorial, Oxford University Press, 2006.
[26] N. Galagali, Y.M. Marzouk, Bayesian inference of chemical kinetic models from proposed reactions, Chem. Eng. Sci. 123 (2015) 170–190.
[27] L. Hakim, G. Lacaze, M. Khalil, K. Sargsyan, H. Najm, J. Oefelein, Probabilistic parameter estimation in a 2-step chemical kinetics model for n-dodecane jet autoignition, Combust. Theory Model. 22 (2018) 1–21.
[28] A.L. Sánchez, F.A. Williams, Recent advances in understanding of flammability characteristics of hydrogen, Prog. Energy Combust. Sci. 41 (2014) 1–55.
[29] I. Glassman, R.A. Yetter, N.G. Glumac, Combustion, Academic Press, 2014.
[30] F.A. Williams, Combustion theory, CRC Press, 2018.
[31] J.C. Keck, Rate-controlled constrained-equilibrium theory of chemical reactions in complex systems, Prog. Energy Combust. Sci. 16 (2) (1990) 125–154.
[32] Z. Ren, G.M. Goldin, V. Hiremath, S.B. Pope, Simulations of a turbulent non-premixed flame using combined dimension reduction and tabulation for combustion chemistry, Fuel 105 (2013) 636–644.
[33] J. Kim, S.B. Pope, Effects of combined dimension reduction and tabulation on the simulations of a turbulent premixed flame using a large-eddy simulation/probability density function method, Combust. Theory Model. 18 (3) (2014) 388–413.
[34] S. Lapointe, B. Bobbitt, G. Blanquart, Impact of chemistry models on flame–vortex interaction, Proc. Combust. Inst. 35 (1) (2015) 1033–1040.
[35] D. Hamiroune, P. Bishnu, M. Metghalchi, J.C. Keck, Rate-controlled constrained-equilibrium method using constraint potentials, Combust. Theory Model. 2 (1) (1998) 81–94.
[36] P. Saxena, F.A. Williams, Testing a small detailed chemical-kinetic mechanism for the combustion of hydrogen and carbon monoxide, Combust. Flame 145 (1–2) (2006) 316–323.
[37] K.A. Bhaskaran, M.C. Gupta, T. Just, Shock tube study of the effect of unsymmetric dimethyl hydrazine on the ignition characteristics of hydrogen-air mixtures, Combust. Flame 21 (1) (1973) 45–48.
[38] K. Braman, T.A. Oliver, V. Raman, Bayesian analysis of syngas chemistry models, Combust. Theory Model. 17 (5) (2013) 858–887.
[39] S.H. Cheung, T.A. Oliver, E.E. Prudencio, S. Prudhomme, R.D. Moser, Bayesian uncertainty analysis with applications to turbulence modeling, Reliab. Eng. Syst. Saf. 96 (9) (2011) 1137–1149.
[40] J.B. Freund, R.H. Ewoldt, Quantitative rheological model selection: good fits versus credible models using Bayesian inference, J. Rheol. 59 (3) (2015) 667–701.
[41] D. Heckerman, A.F.M. Smith, M.W. Eds, A.E. Raftery, M.A. Newton, P.N. Krivitsky, J.M. Satagopan, Estimating the integrated likelihood via posterior simulation using the harmonic mean identity, in: J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, M. West (Eds.), Bayesian statistics 8, Oxford University Press (2007), pp. 1–45.
[42] Z. Ren, G.M. Goldin, V. Hiremath, S.B. Pope, Reduced description of reactive flows with tabulation of chemistry, Combust. Theory Model. 15 (6) (2011) 827–848.
[43] K. Sargsyan, H.N. Najm, R. Ghanem, On the statistical calibration of physical models, Int. J. Chem. Kinet. 47 (4) (2015) 246–276.
[44] D.L. Baulch, C. Cobos, R.A. Cox, C. Esser, P. Frank, T. Just, J.A. Kerr, M.J. Pilling, J. Troe, R.W. Walker, J. Warnatz, Evaluated kinetic data for combustion modelling, J. Phys. Chem. Ref. Data 21 (3) (1992) 411–734.
[45] D.A. Sheen, H. Wang, The method of uncertainty quantification and minimization using polynomial chaos expansions, Combust. Flame 158 (12) (2011) 2358–2374.
[46] T. Nagy, T. Turányi, Uncertainty of Arrhenius parameters, Int. J. Chem. Kinet. 43 (7) (2011) 359–378.
[47] M. Khalil, H.N. Najm, Probabilistic inference of reaction rate parameters from summary statistics, Combust. Theory Model. 22 (2018) 1–31.
[48] M. Muto, J.L. Beck, Bayesian updating and model class selection for hysteretic structural models using stochastic simulation, J. Vib. Control 14 (1–2) (2008) 7–34.
[49] R.E. Kass, A.E. Raftery, Bayes factors, J. Am. Stat. Assoc. 90 (430) (1995) 773–795.
[50] A.A. Konnov, Remaining uncertainties in the kinetic mechanism of hydrogen combustion, Combust. Flame 152 (4) (2008) 507–528.
[51] A. Kéromnès, W.K. Metcalfe, K.A. Heufer, N. Donohoe, A.K. Das, C. Sung, J. Herzler, C. Naumann, P. Griebel, O. Mathieu, M. Krejci, E. Petersen, W. Pitz, H. Curran, An experimental and detailed chemical kinetic modeling study of hydrogen and syngas mixture oxidation at elevated pressures, Combust. Flame 160 (6) (2013) 995–1011.
[52] D.P. Kroese, T. Taimre, Z.I. Botev, Handbook of Monte Carlo methods, 706, John Wiley & Sons, 2011.
[53] J. Skilling, Others, Nested sampling for general Bayesian computation, Bayesian Anal. 1 (4) (2006) 833–859.
[54] N. Friel, A.N. Pettitt, Marginal likelihood estimation via power posteriors, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 70 (3) (2008) 589–607.
[55] R.M. Neal, Annealed importance sampling, Stat. Comput. 11 (2) (2001) 125–139.
[56] M.A. Newton, A.E. Raftery, Approximate Bayesian inference with the weighted likelihood bootstrap, J. R. Stat. Soc. Ser. B 56 (1) (1994) 3–26.
[57] J. Ching, Y.-C. Chen, Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging, J. Eng. Mech. 133 (7) (2007) 816–832.
[58] S. Chib, I. Jeliazkov, Marginal likelihood from the Metropolis–Hastings output, J. Am. Stat. Assoc. 96 (453) (2001) 270–281.
[59] E. Prudencio, S.H. Cheung, Parallel adaptive multilevel sampling algorithms for the Bayesian analysis of mathematical models, Int. J. Uncertain. Quantif. 2 (3) (2012) 215–237.
[60] E.E. Prudencio, K.W. Schulz, The Parallel C++ Statistical Library 'QUESO': Quantification of Uncertainty for Estimation, Simulation and Optimization, in: M. Alexander, et al. (Eds.), Euro-Par 2011: Parallel Processing Workshops. Euro–Par 2011. Lecture Notes in Computer Science, 7155, Springer, Berlin, Heidelberg, 2012.
[61] A.C. Hindmarsh, P.N. Brown, K.E. Grant, S.L. Lee, R. Serban, D.E. Shumaker, C.S. Woodward, SUNDIALS: suite of nonlinear and differential/algebraic equation solvers, ACM Trans. Math. Softw. 31 (3) (2005) 363–396.
[62] S.B. Pope, CEQ: a Fortran library to compute equilibrium compositions using Gibbs function continuation, http://eccentric.mae.cornell.edu/pope/CEQ, 2003.
[63] S.B. Pope, The computation of constrained and unconstrained equilibrium compositions of ideal gas mixtures using Gibbs function continuation, Cornell University, 2003.
[64] J.A. Hoeting, D. Madigan, A.E. Raftery, C.T. Volinsky, Bayesian model averaging: a tutorial, Stat. Sci. 14 (1999) 382–401.
[65] M. Frenklach, H. Wang, C.L. Yu, M. Goldenberg, C.T. Bowman, R.K. Hanson, D.F. Davidson, E.J. Chang, G.P. Smith, D.M. Golden, W.C. Gardiner, V. Lissianski, GRI-Mech--an optimized detailed chemical reaction mechanism for methane combustion, GRI Top. Rep. No. GRI-95/0058, Nov. 1, 1995, Gas Research Institute, 1995.
[66] M. Frenklach, H. Wang, C.L. Yu, M. Goldenberg, C.T. Bowman, R.K. Hanson, D.F. Davidson, E.J. Chang, G.P. Smith, D.M. Golden, W.C. Gardiner, V. Lissianski, http://www.me.berkeley.edu/gri_mech/,
[67] S.H. Lam, D.A. Goussis, Understanding complex chemical kinetics with computational singular perturbation, Symp. Int. Combust., 22, Elsevier (1989), pp. 931–941.
[68] S.H. Lam, D.A. Goussis, The CSP method for simplifying kinetics, Int. J. Chem. Kinet. 26 (4) (1994) 461–486.
[69] M. Valorani, P.P. Ciottoli, R.M. Galassi, S. Paolucci, T. Grenga, E. Martelli, Enhancements of the G-Scheme framework, Flow, Turbul. Combust. 101 (4) (2018) 1023–1033.
[70] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to algorithms, The MIT Press, McGraw-Hill Book Company, 2001.
[71] P.G. Constantine, Active subspaces: emerging ideas for dimension reduction in parameter studies, 2, SIAM, 2015.
[72] P.G. Constantine, P. Diaz, Global sensitivity metrics from active subspaces, Reliab. Eng. Syst. Saf. 162 (2017) 1–13.
[73] V. Hiremath, S.B. Pope, A study of the rate-controlled constrained-equilibrium dimension reduction method and its different implementations, Combust. Theory Model. 17 (2) (2013) 260–293.
[74] V.P. Zhukov, V.A. Sechenov, A.Y. Starikovskii, Spontaneous ignition of methane–air mixtures in a wide range of pressures, Combust. Explos. Shock Waves 39 (5) (2003) 487–495.