Learning and Optimization Under Epistemic Uncertainty with Bayesian Hybrid Models

Elvis A. Eugene[†,‡], Kyla D. Jones[†], Xian Gao, Jialu Wang, and Alexander W. Dowling[*]

Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

[†] contributed equally

[‡] current address: Pfizer Worldwide Research and Development, Groton, CT, 06340, USA

Keywords: Bayesian inference, optimization under uncertainty, grey-box modeling, digital twins

Abstract

Hybrid (i.e., grey-box) models are a powerful and flexible paradigm for predictive science and engineering. Grey-box models use data-driven constructs to incorporate unknown or computationally intractable phenomena into glass-box mechanistic models. The pioneering work of statisticians Kennedy and O'Hagan introduced a new paradigm to quantify epistemic (i.e., model-form) uncertainty. While popular in several engineering disciplines, prior work using Kennedy-O'Hagan hybrid models focuses on prediction with accurate uncertainty estimates. This work demonstrates computational strategies to deploy Bayesian hybrid models for optimization under uncertainty. Specifically, the predictive posterior distributions of Bayesian hybrid models provide a principled uncertainty set for stochastic programming, chance-constrained optimization,

[*]Corresponding Author: adowling@nd.edu, alex@dowlinglab.org

or robust optimization. Through two illustrative case studies, we demonstrate the efficacy of hybrid models, composed of a structurally inadequate glass-box model and Gaussian process, for decision-making using limited training data. From these case studies, we develop recommended best practices and explore the trade-offs between different hybrid model architectures.

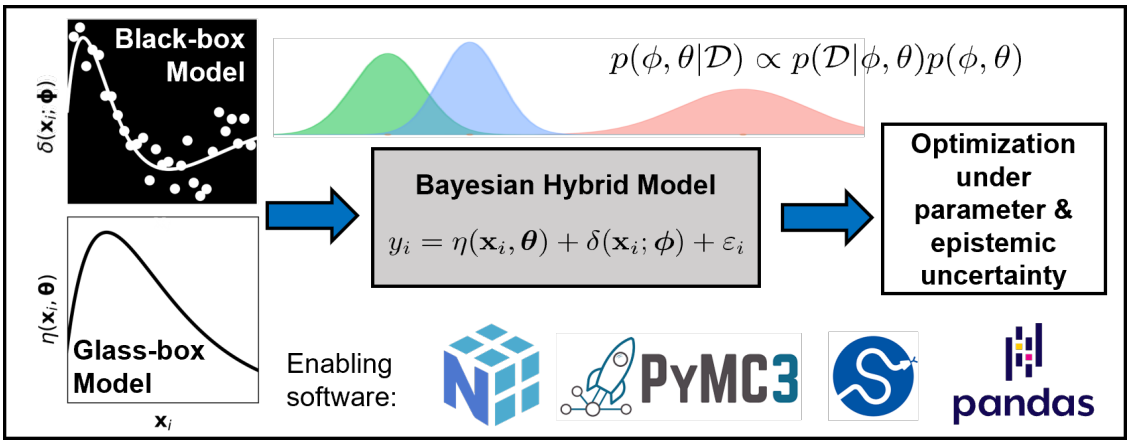


Figure 0: GRAPHICAL ABSTRACT

1 Introduction

Predictive models are fundamental to process systems engineering [1, 2] with ubiquitous applications in design, control, and decision-making applications. In practice, glass-box (i.e., first-principles) models are developed from foundational scientific theory. When accurately formulated and computationally tractable, these models can offer exceptional extrapolation capabilities [3]. When constructing these models, one must consider the trade-off in accurately capturing the underlying physics, the uncertainty of the parameter estimates, the computational burden, and the necessary data for calibration and validation. Furthermore, decreasing the bias between the model outputs and reality requires augmenting the mathematical model with additional terms, thereby increasing the total number of model parameters to be estimated [4]. For these reasons, glass-box models can be tedious to build and validate, especially when the underlying physics

spans multiple length and time scales. Alternatively, black-box (i.e., surrogate) models are easier to develop and, once trained, facilitate fast optimization. Due to their entirely data-driven nature, these models suffer in physical interpretability and can be unreliable for extrapolation outside the training set range [5]. At the intersection of these paradigms, grey-box (i.e., hybrid) models fuse glass-box and data-driven constructs. The resulting framework offers superior physics-informed extrapolation and prediction accuracy due to data-based automatic (re)learning [6] of unmodeled or simplified phenomena.

Regardless of the modeling paradigm, all models are prone to uncertainty [7]. Left unaccounted for, uncertainty can bias decision-making due to over- or under-confident predictions. Non-systematic or aleatoric uncertainty arises from uncontrollable phenomena such as experimental variability and is generally quantifiable by repeating the number of experiments conducted. Epistemic or model-form uncertainty, however, induces systematic bias between a predictive model and the observed phenomena. Epistemic uncertainty can arise when simplifying a mechanistic model or when the underlying phenomena are (partially) misspecified. Epistemic uncertainty is harder to evaluate as it seeks to quantify an abstraction, i.e., unknown unknowns.

This paper proposes a Bayesian hybrid modeling framework for decision-making under aleatoric and epistemic uncertainty. Building upon the pioneering work of Kennedy and O'Hagan [8], we integrate mechanistic or physics-informed glass-box models with data-driven Gaussian process (GP) discrepancy functions into a grey-box hybrid model. Literature to date using Kennedy and O'Hagan paradigm for epistemic UQ focuses on prediction. In contrast, the novel contribution of this work is the extension of Bayesian hybrid models to optimization under both aleatoric and epistemic uncertainty. Specifically, the joint posterior prediction distribution of the Bayesian hybrid model defines the uncertainty sets for a stochastic program. Computational experiments in two case studies demonstrate the performance of Bayesian hybrid models for small data sets.

Through the case studies, we establish best practices regarding the choice of model architecture and estimation methods.

The remainder of the paper is organized as follows. Section 2 reviews related literature. Section 3 describes the general modeling framework and computational implementation. Section 4 describes a ballistics case study, which shows the superiority of Bayesian hybrid models compared to two alternatives. Section 5 develops further computational simplifications in a reaction kinetics case study using the best practices from this first case study. Finally, Section 6 summarizes the key findings and identifies future research directions.

2 Literature Review

2.1 Machine learning and hybrid modeling in chemical engineering

Machine learning (ML) in chemical engineering was explored in the 90s by researchers such as Ydstie [9], Kramer [10], and Bakshi and Stephanopoulos [11], however progress slowed due to computational challenges in training deep networks [12]. ML resurged as a popular technology in the 21st century with many applications across domains, prompting its revival in the chemical engineering community. Jackson, Webb, and Pablo [13], Haghighatlari and Hachmann [14], Lee, Shin, and Realff [12], and Ning and You [15] discuss recent ML advances in domains such as molecular modeling and simulation, soft materials design, process systems engineering, and optimization under uncertainty. Of particular interest, Haghighatlari and Hachmann [14] identify the crucial need to develop ML techniques trained on small or sparse data for applications where data generation is the bottleneck. Lee, Shin, and Realff [12] pose the challenge of balancing exploitation or improving the objective versus exploration or reducing model uncertainty for future ML applications. These challenges are complementary since understanding the model uncertainty

can enable active learning for improved model building with small datasets.

Prompting the advancement of ML, the advent of the big data environment catalyzed the development of hybrid modeling paradigms [4], and was further motivated by problems such as process control [16]. Pioneered by Psichogios and Ungar [17], a combination of first-principles and artificial neural networks (ANN) were devised to improve the model for a fedbatch bioreactor whose dynamic behavior is tough to model despite the use of complex kinetic expressions due to the presence of unmodeled interactions between living cells. Their use of hybrid models evaded overfitting issues neural networks faced and demonstrated a significantly lower data requirement. The works by Thompson and Kramer [18] underscored Psichogios and Ungar's findings and proposed ANN-hybrid models for sparse and noisy data to maximize the value of domain specific knowledge. These early successes lead to new grey-box modeling [6, 19, 20] in diverse applications such as design of reactors [21, 22, 23, 24] and distillation columns [25], polymerization [26], crystallization [27], hydraulic fracturing [28], fluid catalytic cracking [29], model predictive control [30], separations [5], smart manufacturing and digital twins [31], and many more.

More recently, methodological advances to hybrid modeling leverage system derivatives and mechanistic parameters of ordinary differential equations with neural networks in the emerging paradigm of neural differential equations [32]. Motivated by the fouling of an electrodialysis membrane, De Jaegher et al. [33] pioneered fundamental contributions to this line of inquiry by developing neural differential equations to accurately predict decreasing fluxes across the membrane despite simplifying assumptions in the underlying force balance equations. Across these studies, ANN-based ML dominates the choice for the surrogate model and fails to consider epistemic uncertainty. Investigations into explicitly accounting for model uncertainty with ANNs involve the development of several surrogate models for strategies such as bootstrapping, leading to an increased computational workload [34] impractical for online applications and iterative

model development.

Reaction networks are often too complex to formulate as mathematical models to embed in optimization problems or are (partially) unknown, making reactive systems well-suited for hybrid models [35]. Under sequential hybrid model architectures, parameters of kinetic models can be estimated with black-box models, which are used to generate outputs from a the glass-box component that describes the known physics. Alternatively, the black-box model can provide a data-driven correction term to the white-box model. Regarding the former approach, Saraceno et al. [36] developed a coupled neural network with a logic condition model to predict the kinetic parameters for the fermentation production of ethanol. Later, Azarpour et al. [37, 38] applied a neural network model to calculate the reaction rates for a terephthalic acid three-phase reaction system to solve mass balance equations. Recently, Bui et al. [39] used partial least squares and Kalman filtering to estimate and update the activity of a catalyst bed to predict the catalyst lifetime of industrial-scale PFRs with real plant data.

2.2 UQ and Kennedy-O'Hagan in chemical engineering

In their 2001 seminal work, Kennedy and O'Hagan [8] proposed a statistical framework for the calibration of models under both epistemic (model-form) and aleatoric uncertainty. They were the first to use Gaussian process (GP) discrepancy functions to quantify the systematic bias between the model predictions and truth. In contrast, classical (non)linear regression theory often assumes that the model structure is correct and the errors are independent and identically distributed (i.i.d.) normal [40]. When these assumptions fail, estimates of the model parameters are biased and unreliable. Moreover, input-dependent errors can lead to predictions that have large random deviations from the observations. Kennedy and O'Hagan [8] overcame this limitation by quantifying model bias using correlations between control variables via the GP discrepancy

function.

Over the past two decades, Kennedy and O'Hagan (KOH) models for uncertainty quantification (UQ) have been successful in various applications such as water quality management [41, 42], thermal engineering [43, 44], fluid dynamics [45], energy storage [46], structural dynamics [47], surrogate-based optimization [48], and carbon capture [49]. Kalyanaraman et al. [50] demonstrated how KOH models can be used to overcome model discrepancy and predict the breakthrough in a rapid thermal swing adsorption process for CO_2 capture. However, the complexity of estimating multiple hyperparameters of the discrepancy function was noted as a computational barrier. Such barriers often impede iterative model development. Mebane et al. [51] used quantum chemical calculations and experimental data to identify parameter uncertainty and model discrepancy for CO_2 adsorption with mesoporous silica-supported amines.

Extensions to KOH models by Bhat et al. [52] include the development of dynamic discrepancy models based on BSS-ANOVA GPs [53]. This work enabled uncertainty propagation in multiscale systems but highlighted the computational complexity of the process and suggested the investigation of adaptive techniques to economize on the Markov chain Monte Carlo (MCMC) calibration of the Bayesian models. Additional experiments using dynamic discrepancy models with BSS-ANOVA GPs were motivated by Li et al. [54], who propagated uncertainty from the bench to process scale for CO_2 capture using reaction-diffusion kinetics on solid adsorbents in a bubbling fluidized-bed. Most recently, Ostace et al. [55] incorporated a discrepancy function with a Langmuir adsorption model and used the Bayesian information criterion to guard against overfitting. Their stochastic Langmuir model demonstrated robust predictions over variability in operating conditions due to the inclusion of model-form and parameter uncertainty. These contributions demonstrate the potential to upscale uncertainty to large-scale systems, resulting in less conservative designs and models suitable for optimization under uncertainty.

Outside the aforementioned contributions, UQ of reactive systems predominantly focuses on parameter uncertainty. Notably, Chaffart et al. [56] propagated parameter uncertainty in a catalytic reactor model to the concentration of the reactor products using power series expansion [56]. Similarly, Kiamev et al. [57] motivated the investigation of parameter uncertainty by studying chemical vapor deposition and catalytic flow reactors using polynomial chaos expansion and multilevel Monte Carlo. We advocate that consideration of parameter and epistemic uncertainty for calibration and prediction is critical in the data-limited regime of reactive systems. Moreover, decision-making frameworks for optimization under uncertainty that leverage the demonstrated benefits of KOH models with GP discrepancy functions are mainly absent.

2.3 Decision-making under uncertainty

Optimization under uncertainty is a cornerstone of process systems engineering, often utilizing stochastic programming, chance-constrained optimization, robust optimization, and constraint back-off approaches. In stochastic programming, uncertain parameters are modeled using probability distributions and the objective optimizes an expected value across all the realizations of this uncertainty [58]. The parameter uncertainty in a stochastic program is approximated as scenarios, which are the discrete realizations of a probability distribution. Popular chemical engineering applications of stochastic programming include flowsheet optimization [59], supply chain management [60], energy systems [61], control [62], and beyond [63]. Chance-constrained optimization seeks to optimize an objective ensuring that the constraints are satisfied within a specified probability range, relying on probability distributions to capture the uncertainty in parameters [64]. This framework enables custom definition of risk levels, but can suffer in computational tractability [65, 66]. Chance constrained optimization may be regarded as a generalization of robust optimization [67], which avoids the need for probability distributions by defining parameter uncertainty

using sets, thereby safeguarding against a worst case scenario in this set.

Recent data-driven extensions of these techniques (see review by Ning and You [15]) bridge the gap between uncertainty modeling and decision making. For example, the emerging paradigm of distributionally robust optimization (DRO) safeguards against the worst-case in an ambiguous set of probability distributions inferred using statistics and big data analytics. DRO was used by Shang and You [68] in a process network planning and scheduling problem demonstrating less conservative solutions, potentially contributing to higher profits. Another extension is data-driven adaptive robust optimization, which integrates Bayesian ML in the form of a nonparametric Dirichlet process mixture model in an optimization problem to account for uncertainty by exploiting big data from process industries [69]. The final approach uses constraint back-offs in optimization problems to prevent constraint violation under uncertainty [70]. At large, this method has the lowest computational burden for optimization under uncertainty [71] despite the use of iterative techniques for the calculation of back-off terms, defined using the second statistical moment of the active constraints [72, 73]. Optimization with constraint back-offs have been demonstrated to robustify designs for fixed bed reactors [74], polymerization processes [75], enzyme catalyzed reactions [76], nonlinear model predictive control [77], and model based experimental campaigns [78]. Despite the unprecedented capabilities offered by the reviewed techniques, we identify two limitations, specifically optimization under epistemic (model-form) uncertainty and leveraging small or sparse data sets are not considered.

Optimization with hybrid models is closely related to Bayesian optimization (BO) [79], a class of adaptive sampling algorithms often deployed for the sequential design of experiments and derivative-free optimization. Importantly, BO can model prior information about the uncertainty of the process, making it a natural optimization algorithm for hybrid models. Recently, González and Zavala [80] developed a level-set partitioning algorithm for parallel sequential design of ex-

periments, in which the algorithm proposes multiple experiments in a single iteration. Using a reactor case study, the authors empirically demonstrated that their approach reduced search time and increased the probability of identifying a globally optimal solution while overcoming a known challenge of repeated experiments. Similarly, Cosenza et al. [81] experimentally validated a multi-source BO algorithm for cell culture media optimization. Their algorithm reduced experimental effort by 38% compared to a traditional DOE. Finally, Folch et al. [82] developed and empirically validated a novel BO algorithm for multi-source and asynchronous experiments, in which the algorithm selects new experiments before revealing prior results.

Regarding theoretical contributions to BO algorithms of hybrid models, Paulson and Lu [83] have made several to their novel COnstrained Bayesian optimizAtion of computationaLly expensive grey-box models exploiting derivaTive information (COBALT) algorithm. Moreover, COBALT is a one-step Bayes optimal algorithm that aims to tackle efficient constrained global optimization of multivariate composite functions (hybrid models). Using a bioreactor calibration case study, the authors initially demonstrated the promising performance of the algorithm. In the absence of constraints and the limit of infinite samples, COBALTs convergence is guaranteed. At the time, however, theoretical performance was not established on the bounds of the convergence rate for finite cases. Toward these ends, the authors recently proposed a novel Constrained Upper Quartile Bound (CUQB) algorithm [84], which provides improved theoretical guarantees on convergence rate bounds to the optimal global solution under mild regularity assumptions. Closely related, model-based design of experiments (MBDoE) provides a framework to sequentially optimize data collection to minimize uncertainty or discern between candidate models or both [85]. However, numerous research opportunities exist to integrate statistical theory supporting MBDoE with information theory and algorithms for BO. These advances would provide a theoretical foundation for optimization under uncertainty with hybrid models. In this context, this paper explores the

impact of hybrid model architectures on decision-making.

3 Methods: Bayesian hybrid modeling framework

Regarding notation, we use lowercase bold font to indicate vectors and uppercase bold font to indicate matrices. We use the notation $f(\cdot)$ to denote a function f with one input variable. We use hats to denote point estimates, e.g., $\hat{y}$ is a point estimate of the random variable y.

3.1 Kennedy-O'Hagan models for uncertainty quantification

In their seminal work, statisticians Kennedy and O'Hagan proposed a Bayesian framework to calibrate computationally expensive computer models from measured outputs of a physical system. The first statistical model in this framework is a true process model:

$$\mathbf{y}_i = \boldsymbol{\zeta}(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon). \tag{3.1}$$

Here, $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,d}]^\intercal \in \mathbb{R}^d$ is a vector of observations from an experiment $i \in \mathbb{N} : i \leq n$. The observations are assumed to be generated by an unknown process $\boldsymbol{\zeta}(\cdot)$ which is a function of the experiment's controlled (i.e., independent) variables $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,m}]^\intercal \in \mathbb{R}^m$, which we call controls. For all experiments, the outputs of the unknown process are corrupted by independent and identically distributed (i.i.d.) Gaussian measurement error $\boldsymbol{\epsilon}$ with an unknown variance-covariance matrix $\boldsymbol{\Sigma}_\varepsilon$.

Though we can formulate high-fidelity representations of the true process $\boldsymbol{\zeta}(\cdot)$ from expert scientific knowledge, building a model that captures every detail of a real-world system is impossible. Moreover, there will always be uncertainty between a mathematical model and the true

process. Toward these ends, KOH propose an additive statistical model to explain the true process:

$$\boldsymbol{\zeta}(\mathbf{x}_i) = \boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\theta}) + \mathbf{a}\, \delta(\mathbf{x}_i; \boldsymbol{\phi}). \tag{3.2}$$

The first component in Eq. (3.2) is a mechanistic model $\boldsymbol{\eta} : \mathbb{R}^m \times \Theta \to \mathbb{R}^d$ that represents the modeler's working knowledge of the underlying physics of the true process. The mechanistic model has two inputs: known control variables $\mathbf{x}_i$ and unknown physically meaningful parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. The second component of the HM is a data-driven discrepancy term $\mathbf{a}\, \delta(\cdot)$ that accounts for bias between the true process and the mechanistic model as a function of the control variables and unknown hyperparameters $\boldsymbol{\phi}$. Hence, Eq. (3.2) can be thought of as a "hybrid model" because it combines physical intuition with hidden trends from data through $\boldsymbol{\eta}(\cdot, \cdot)$ and $\mathbf{a}\, \delta(\cdot)$, respectively.

Though vector-valued representations of the discrepancy term exist for KOH models in literature [86], we limit this introduction to scalar discrepancies that propagate in a known way to a multivariate system. Moreover, we use a known projection coefficient $\mathbf{a} \in \mathbb{R}^d$ to project the discrepancy of the true process to the system outputs. The discrepancy is modeled using a Gaussian Process (GP) in the original KOH framework. KOH employed a GP for the same reason normal distributions are often used in statistics. That is, GP's are convenient, flexible, and often realistic. That being said, joint normality should be a feature of prior beliefs about $\delta(\cdot)$ for this to be a reasonable modeling choice. If this condition is not met, other nonparametric methods can be used.

Combining Eq. (3.1) with Eq. (3.2) yields what we shall refer to as the (Bayesian) hybrid model ((B)HM) from here on out:

$$\mathbf{y}_i = \boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\theta}) + \mathbf{a}\, \delta(\mathbf{x}_i; \boldsymbol{\phi}) + \boldsymbol{\varepsilon}_i. \tag{3.3}$$

In this work, we deploy the KOH HM for cases where the mechanistic model is computationally inexpensive and misspecified in functional form. Furthermore, we use the KOH framework to quantify bias between an inadequate first-principles model and experimental data.

3.2 Gaussian process regression

A Gaussian process (GP) is a time-continuous stochastic process (i.e., a set of random variables indexed by a continuous variable) for which every finite subset of random variables follows a multivariate normal distribution. The distribution of a GP is the joint distribution of all the its random variables, and as such, a GP can be thought of as an infinite-dimensional generalization of a multivariate normal distribution. Furthermore, a GP is a distribution over functions with a continuous domain [87, 88, 89, 90, 91]. We use the notation $f(\cdot) \sim \mathcal{GP}(m(\cdot), c(\cdot,\cdot))$ to denote that $f(\cdot)$ follows a GP distribution with mean and covariance functions

$$m(\mathbf{x}) := \mathbb{E}[f(\mathbf{x})] \tag{3.4}$$

and

$$c(\mathbf{x}, \mathbf{x}') := \text{Cov}[f(\mathbf{x}), f(\mathbf{x}')], \tag{3.5}$$

respectively. To model $m(\cdot)$ and $c(\cdot,\cdot)$, we assume that the process is stationary. That is, we assume that the process does not change when shifted in time. We impose this belief in the mean function by setting $m(\cdot) = 0$. We choose a zero mean function for ease of notation, though any function that satisfies the property $m(\mathbf{x}) = m(\mathbf{x} + d\mathbf{x})$ may be used. Furthermore, the covariance $c(\cdot,\cdot)$ is modeled as

$$c(\mathbf{x}, \mathbf{x}') = \sigma_f^2 k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 k(\mathbf{x} - \mathbf{x}') \tag{3.6}$$

where σ_f^2 is an unknown common variance of the process and $k(\cdot,\cdot)$ is a chosen correlation function that satisfies $k(\mathbf{0}) = 1$.

For n evaluations of $f(\cdot)$, this GP is equivalent to the multivariate normal distribution:

$$[f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)] \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{3.7}$$

$$\mathbf{K} = (K_{ij})_{i,j=1}^{n}, \quad K_{ij} = \sigma_f^2 k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\phi}), \quad \forall (i,j) \in \{1, \ldots, n\}^2. \tag{3.8}$$

When the process is corrupted by Gaussian noise, i.e.,:

$$g(\mathbf{x}_i) = f(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_1, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \tag{3.9}$$

the additive property of Gaussian distributions allows us to write:

$$[g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n)] \sim \mathcal{N}(\mathbf{0}, \mathbf{K} + \boldsymbol{\Sigma}_\varepsilon), \tag{3.10}$$

$$\boldsymbol{\Sigma}_\varepsilon = (\Sigma_{\varepsilon_{ij}})_{i,j=1}^{n}, \quad \Sigma_{\varepsilon_{ij}} = \sigma_\varepsilon^2 \Delta_{ij}, \quad \forall (i,j) \in \{1, \ldots, n\}^2, \tag{3.11}$$

where Δ_{ij} is the Kronecker delta function.

3.2.1 Kernel functions

We consider two correlation functions $k(\cdot,\cdot)$ to define the GP. The first is the **Radial basis function** (RBF) also known as the Exponentiated Quadratic or Squared Exponential kernel [92]:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{1}{2}\left(\frac{||\mathbf{x} - \mathbf{x}'||}{\ell}\right)^2\right]. \tag{3.12}$$

Here, ℓ is a length scale parameter that determines how far apart $\mathbf{x}$ and $\mathbf{x}'$ need to be before $f(\mathbf{x})$ can be very different from $f(\mathbf{x}')$. Moreover, ℓ controls the smoothness of the process. A generalization of Eq. (3.12) would be to write:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left[-(\mathbf{x} - \mathbf{x}')^\intercal \mathbf{\Lambda} (\mathbf{x} - \mathbf{x}')\right], \tag{3.13}$$

where $\mathbf{\Lambda}$ is an unknown symmetric positive definite matrix with the form $\mathbf{\Lambda} = \text{diag}(\ell_1, \ldots, \ell_m)$.

The second kernel we consider is the **Matern 3/2 kernel**:

$$k(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{\sqrt{3||\mathbf{x} - \mathbf{x}||^2}}{\ell}\right) \exp\left[-\frac{\sqrt{3||\mathbf{x} - \mathbf{x}||^2}}{\ell}\right]. \tag{3.14}$$

This correlation function is defined using a positive parameter $\nu = 3/2$ [93]. As $\nu \to \infty$, the Matern kernel is equivalent to the RBF kernel.

3.2.2 Likelihood model

For the training inputs $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_n^\intercal]$, we write the corresponding observed values as $\mathbf{g} = [g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n)]$. For predictions outside the training data $\mathbf{X}^* = [\mathbf{x}_1^{*\intercal}, \ldots, \mathbf{x}_q^{*\intercal}]$, let $\mathbf{f}^* = [f(\mathbf{x}_1^*), \ldots, f(\mathbf{x}_q^*)]$ denote the corresponding function evaluations. The joint distribution of the training data and the predictions is:

$$\begin{pmatrix} \mathbf{g} \\ \mathbf{f}^* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K} + \mathbf{\Sigma}_\varepsilon & \mathbf{K}|_{\mathbf{X},\mathbf{X}^*} \\ \mathbf{K}|_{\mathbf{X}^*,\mathbf{X}} & \mathbf{K}|_{\mathbf{X}^*,\mathbf{X}^*} \end{pmatrix}\right). \tag{3.15}$$

Conditioning on the observed values $\mathbf{g}$, the predictive distribution is:

$$\mathbf{f}^*|\mathbf{X}, \mathbf{g}, \mathbf{X}^* \sim \mathcal{N}(\boldsymbol{\mu}_f^*, \mathbf{K}_f^*). \tag{3.16}$$

where $\boldsymbol{\mu}_f^*$ and $\mathbf{K}_f^*$ are the predictive mean and variance, given by

$$\boldsymbol{\mu}_f^* = \mathbf{K}|_{\mathbf{X}^*,\mathbf{X}}(\mathbf{K} + \boldsymbol{\Sigma}_\varepsilon)^{-1}\mathbf{g} \tag{3.17}$$

and

$$\mathbf{K}_f^* = \mathbf{K}|_{\mathbf{X}^*,\mathbf{X}^*} - \mathbf{K}|_{\mathbf{X}^*,\mathbf{X}}(\mathbf{K} + \boldsymbol{\Sigma}_\varepsilon)^{-1}\mathbf{K}|_{\mathbf{X},\mathbf{X}^*}, \tag{3.18}$$

respectively.

3.3 Model calibration

For the HM (Eq. (3.3)) to be fully defined, we need to estimate all unknown model parameters $\boldsymbol{\omega} = [\boldsymbol{\theta}^\intercal, \boldsymbol{\phi}^\intercal, \sigma_\varepsilon]$, also referred to as model calibration. For ease of notation, we use a scalar case of the hybrid model.

3.3.1 Bayesian approach

We use Bayes' rule to perform Bayesian calibration of the HM:

$$\underbrace{p(\boldsymbol{\omega}|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{D}|\boldsymbol{\omega})}^{\text{likelihood}}\ \overbrace{p(\boldsymbol{\omega})}^{\text{prior}}}{\underbrace{\int p(\mathcal{D}|\boldsymbol{\omega})p(\boldsymbol{\omega})d\boldsymbol{\omega}}_{\text{evidence}}}. \tag{3.19}$$

In Eq. (3.19), we use $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ to denote the training data. In the Bayesian approach, the unknown HM parameters $\boldsymbol{\omega}$ are treated as random variables. Furthermore, prior information about $\boldsymbol{\omega}$ is incorporated through a probability distribution $p(\boldsymbol{\omega})$ called the prior. The training data $\mathcal{D}$ are used to construct likelihood model $p(\boldsymbol{\omega}|\mathcal{D})$, which is always independent of the prior. The evidence integral (i.e., marginal likelihood) quantifies the agreement between the data and the

prior. Use of these terms as governed by Eq. (3.19) yields the posterior distribution $p(\boldsymbol{\omega}|\mathcal{D})$ or the updated probability of observing the parameters after considering the data.

3.3.2 Simultaneously calibrated Bayesian hybrid model

The simultaneously calibrated BHM (hereafter referred to as the *simultaneous model*) is trained by jointly inferring the unknown physically-meaningful parameters $\boldsymbol{\theta}$ and the model discrepancy hyperparameters $\boldsymbol{\phi}$. The Bayesian hierarchical model is:

$$y|\delta(\mathbf{x}), \varepsilon \sim \mathcal{GP}(\eta(\mathbf{x}, \boldsymbol{\theta}), c(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}) + \sigma_\varepsilon^2 \Delta_{\mathbf{x},\mathbf{x}'}), \tag{3.20a}$$

$$\delta(\mathbf{x})|\boldsymbol{\phi} \sim \mathcal{GP}(0, c(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi})), \quad \varepsilon|\sigma_\varepsilon^2 \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{3.20b}$$

$$\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}(\cdot), \quad \boldsymbol{\phi} \sim p_{\boldsymbol{\phi}}(\cdot), \quad \sigma_\varepsilon^2 \sim p_{\sigma_\varepsilon^2}(\cdot), \tag{3.20c}$$

where $p_z(\cdot)$ denotes the probability distribution for some random variable z.

3.3.3 Incrementally calibrated Bayesian hybrid model

Kahrs and Marquardt [94] proposed the incremental identification of hybrid models by decomposing a penalized least squares parameter estimation approach into a series of more straightforward subproblems. Similarly, Wong, Storlie, and Lee [95] developed a frequentist approach to computer model calibration that theoretically justifies dividing the model calibration problem into two steps.

Building upon these methods, we define the incrementally calibrated Bayesian hybrid model (hereafter referred to as the *incremental model*). First, the glass-box model is fit to the experimental observations ignoring the discrepancy. That is, we (incorrectly) assume the model explains the observations:

$$y = \eta(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon_\eta \tag{3.21}$$

where ε_η is a measurement error term. The hierarchical model for Eq. (3.21) is:

$$y|\boldsymbol{\theta}, \varepsilon_\eta \sim \mathcal{N}(\eta(\mathbf{x}, \boldsymbol{\theta}), \sigma^2_{\varepsilon_\eta}), \tag{3.22a}$$

$$\varepsilon_\eta|\sigma^2_{\varepsilon_\eta} \sim \mathcal{N}(0, \sigma^2_{\varepsilon_\eta}), \tag{3.22b}$$

$$\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}(\cdot), \quad \sigma^2_{\varepsilon_\eta} \sim p_{\sigma^2_{\varepsilon_\eta}}(\cdot). \tag{3.22c}$$

In Eq. (3.22b), we impose a mean-zero Gaussian structure for the error, which we discuss the implications of below.

In step two, the GP is fit to the residuals between experimental observations and the predictions of the glass-box model $r(\mathbf{x}) = y - \eta(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ where $\boldsymbol{\theta}|\mathcal{D}$ is a draw from the posterior $p(\boldsymbol{\theta}|\mathcal{D})$. The model for the residuals is:

$$r(\mathbf{x}) = \delta(\mathbf{x}) + \varepsilon_\delta \tag{3.23}$$

where ε_δ is another measurement error term. The hierarchical model for step two is:

$$r(\mathbf{x})|\delta(\mathbf{x}), \varepsilon_\delta \sim \mathcal{GP}(0, c(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi}) + \sigma^2_{\varepsilon_\delta}\Delta_{\mathbf{x},\mathbf{x}'}), \tag{3.24a}$$

$$\delta(\mathbf{x})|\boldsymbol{\phi} \sim \mathcal{GP}(0, c(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi})), \quad \varepsilon_\delta|\sigma^2_{\varepsilon_\delta} \sim \mathcal{N}(0, \sigma^2_{\varepsilon_\delta}), \tag{3.24b}$$

$$\boldsymbol{\phi} \sim p_{\boldsymbol{\phi}}(\cdot), \quad \sigma^2_{\varepsilon_\delta} \sim p_{\sigma^2_{\varepsilon_\delta}}(\cdot). \tag{3.24c}$$

Again, we impose a mean-zero Gaussian error structure for ε_δ which is technically misspecified. This is the critical difference between the incremental model proposed in Eq. (3.21)-(3.24) and the work of Wong, Storlie, and Lee. Furthermore, Wong, Storlie, and Lee proposed a more theoretically sound frequentist approach in which nonparametric bootstrap is used to build confidence intervals for $\hat{\boldsymbol{\theta}}$ and $\hat{\delta}(\cdot)$ thereby circumventing the need to assume an error structure. The framework proposed here does not use this method, as bootrapping is computationally unrealistic for

iterative model development and online optimization. These more rigorous approaches should be considered for applications with significant model discrepancies.

3.3.4 Frequentist approach to incremental calibration

Without prior information, frequentist inference is a computationally less demanding alternative to MCMC. In step one, Bayesian calibration can be replaced with (non)linear least squares regression to obtain estimates of the physically meaningful parameters $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n} M(\boldsymbol{\theta}), \quad M(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i, \boldsymbol{\theta}))^2, \tag{3.25}$$

$$\hat{\sigma}^2_{\varepsilon_\eta} = \left(\frac{1}{n-p}\right) M(\hat{\boldsymbol{\theta}}). \tag{3.26}$$

In step two, the remaining (hyper)parameters $\boldsymbol{\psi} = [\sigma^2_{\varepsilon_\delta}, \boldsymbol{\phi}^\intercal]$ can be estimated with maximum likelihood estimation (MLE):

$$\hat{\boldsymbol{\psi}}_{\text{MLE}} = \arg\max_{\boldsymbol{\psi}\in\Psi} \mathcal{L}(\boldsymbol{\psi}), \quad \mathcal{L}(\boldsymbol{\psi}) = \log p(\mathbf{r}|\mathbf{X}) \tag{3.27}$$

where $\log p(\mathbf{r}|\mathbf{X})$ is the log marginal likelihood:

$$\log p(\mathbf{r}|\mathbf{X}) = -\frac{1}{2}\mathbf{r}^\intercal(\mathbf{K} + \sigma^2_{\varepsilon_\delta}\mathbf{I})^{-1}\mathbf{r} - \frac{1}{2}\log|\mathbf{K} + \sigma^2_{\varepsilon_\delta}\mathbf{I}| - \frac{n}{2}\log 2\pi. \tag{3.28}$$

Here, $\mathbf{r}$ is an n-dimensional vector of the residuals $[y_i - \eta(\mathbf{x}_i, \hat{\boldsymbol{\theta}})]_{i=1}^n$, $\mathbf{I}$ is the identity matrix, and $|\cdot|$ is the determinant.

3.3.5 Posterior approximation

Because the evidence integral is high-dimensional, Eq. (3.19) is most often analytically intractable, and the target posterior must be sampled from an empirical posterior constructed with MCMC.

We compare three approximations for approximating the HM posterior below. All Bayesian model calibration was performed using the open source `Python` package `PyMC3` [96]. Specifically, Hamiltonian Monte Carlo implemented in the No-U-Turn Sampler NUTS) [97] was used via the `PyMC3.sample()` method to estimate posteriors using 4 chains of 1000 samples each.

M0: Full estimation uses the traceplots of the model parameters $\boldsymbol{\omega}$ obtained from MCMC to generate a sample set S of the empirical distribution. For the GP, each sample s is propagated through the conditional mean and variance formulas, which necessitates recalculating the prediction mean and variance across $|S| = \mathcal{O}(10^3)$ samples of parameters. This computation in `PyMC3` [96] can take about 10 s per sample or about 3 hours to fully evaluate the model prediction and uncertainty with 1000 samples. This method is included for completeness but is not discussed further due to its computationally intensive nature. Moreover, three hours is too long for many online applications and extremely inconvenient for iterative model development.

M1: Composite estimation uses the maximum-a-posteriori (MAP) (i.e., posterior mode) estimate of the GP hyperparameters:

$$\hat{\boldsymbol{\phi}}_{\text{MAP}} \propto \underset{\boldsymbol{\phi}\in\Phi}{\arg\max}\, p(\delta|\boldsymbol{\phi})\; p(\boldsymbol{\phi}) \tag{3.29}$$

and the samples from the sample set S of the remaining parameters.

M2: Composite estimation neglecting GP uncertainty only uses samples of the standard deviation of the random noise model to calculate the hybrid model uncertainty. This method has computational complexity similar to that of M1 and helps study the performance of a hybrid model in which the discrepancy function is a conventional machine-learning model that does not provide uncertainty estimates. This is sometimes done under the name of Kriging interpolation where the statistical interpretation of the GP kernel matrix is ignored [98].

3.4 Benchmark models

We consider two references models to benchmark the HM.

The *simple physics model* represents a set of equations that do not completely capture all the phenomena governing a process due to lack of knowledge, coarse-graining for tractability, etc. These simplifications result in a systematic bias between experimental observations and model predictions, leading to epistemic uncertainty in the system. We define the simple physics model as:

$$y = \eta(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{3.30}$$

The *GP-only* model is a purely data-driven surrogate model; it is calibrated using experimental observations contained in a training set and does not incorporate scientific knowledge or physical intuition (unless a specialized kernel is employed). In our framework, the GP-only model is defined as:

$$y = \delta(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{3.31}$$

3.5 Optimization under uncertainty with Bayesian hybrid models

Ultimately, we wish to use the Bayesian hybrid models for decision-making. Let a general utility function $u(y)$ encode the consequences of a decision $\mathbf{x}$. For example, in the first case study, we consider a ballistic trajectory, and $u(\cdot)$ is the proximity to the target. In the second case study, $u(\cdot)$ is the value of a reactor effluent. In the context of Bayesian optimization, $u(\cdot)$ is an acquisition function such as expected improvement, probability of improvement, or the lower confidence bound.

A decision-maker seeks to maximize the expected value of the utility function over the random variables $\boldsymbol{\omega}$:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \quad \mathbb{E}_{\boldsymbol{\omega}}[u(y(\mathbf{x}))], \tag{3.32a}$$

$$\text{s.t.} \qquad y(\mathbf{x}) = \eta(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) + \delta(\mathbf{x}; \boldsymbol{\phi}|\mathcal{D}) + \varepsilon|\sigma_\varepsilon^2. \tag{3.32b}$$

Eq. (3.32) is a single-stage stochastic program if there is no recourse.

We now discuss approximations for the expected value. We start by drawing samples $s \in S$ from trace of the posterior.

$$\mathbb{E}_{\boldsymbol{\omega}}[u(\boldsymbol{\theta}, \delta, \varepsilon)] = \int_\Theta \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(\boldsymbol{\theta}, \delta, \varepsilon)\ p(\boldsymbol{\theta}, \delta, \varepsilon)\ d\delta\ d\varepsilon\ d\boldsymbol{\theta} \tag{3.33a}$$

$$= \int_\Theta \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(\boldsymbol{\theta}, \delta, \varepsilon)\ p(\delta, \varepsilon|\boldsymbol{\theta})\ p(\boldsymbol{\theta})\ d\delta\ d\varepsilon\ d\boldsymbol{\theta} \tag{3.33b}$$

$$\approx \frac{1}{|S|} \sum_{s \in S} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(\boldsymbol{\theta}_s, \delta, \varepsilon)\ p(\delta, \varepsilon|\boldsymbol{\theta}_s)\ d\delta\ d\varepsilon \tag{3.33c}$$

where $p(\delta, \varepsilon|\boldsymbol{\theta}_s)$ is the joint conditional distribution for δ and ϵ given a sample s of the marginal posterior of $\boldsymbol{\theta}$.

Next, we exploit the fact that the kernels of the GP model $\delta(\cdot)$ and observation error ε are additive to define the random variable $f := \delta + \varepsilon$ which follows a Gaussian distribution with first and second moments μ_f and σ_f^2, respectively. This allows us to approximate the remaining integral with a 7-degree Gauss-Hermite quadrature [99, 100], implemented via NumPy `polynomial.hermite.hermgauss()`:

$$\mathbb{E}_{\boldsymbol{\omega}}[u(\boldsymbol{\theta},\delta,\varepsilon)] \approx \frac{1}{|S|}\sum_{s\in S}\int_{-\infty}^{\infty} u(\boldsymbol{\theta}_s, f)\; p(f|\boldsymbol{\theta}_s)\; df \tag{3.34a}$$

$$\approx \frac{1}{\sigma_f\sqrt{2\pi}|S|}\sum_{s\in S}\int_{-\infty}^{\infty} u(\boldsymbol{\theta}_s, f)\exp\left[-\left(\frac{f-\mu_f}{2\sigma_f^2}\right)^2\right]\; df. \tag{3.34b}$$

404 As Eq. (3.34b) does not exactly correspond to a Hermite polynomial, we use a change of variables:

$$\xi := \frac{f-\mu_f}{\sqrt{2}\sigma_f^2} \iff f = \mu_f + \sqrt{2}\sigma_f\xi \tag{3.35}$$

405 and integration by substitution to obtain:

$$\mathbb{E}_{\boldsymbol{\omega}}[u(\boldsymbol{\theta},f)] \approx \frac{1}{\sqrt{\pi}|S|}\sum_{s\in S}\int_{-\infty}^{\infty} u(\boldsymbol{\theta}_s, \mu_f+\sqrt{2}\sigma_f\xi)\exp(-\xi^2)\; d\xi \tag{3.36a}$$

$$\approx \frac{1}{\sqrt{\pi}|S|}\sum_{s\in S}\sum_{j\in J} w_j u(\boldsymbol{\theta}_s, \mu_f+\sqrt{2}\sigma_f\xi_j), \tag{3.36b}$$

406 where w_j and ξ_j are the weights and nodes contained in the set of quadrature points J and $1/\sqrt{\pi}$

407 normalizes the Gauss-Hermite quadrature rule.

408 Using these approximations, we assemble the following optimization problem:

$$\arg\max_{\mathbf{x}} \quad \frac{1}{|S|\sqrt{\pi}}\sum_{s\in S}\sum_{j\in J} w_j\, u_{s,j}, \tag{3.37a}$$

$$\text{s.t.}\;\; u_{s,j} = u(y_{s,j}), \quad \forall s\in S, j\in J \tag{3.37b}$$

$$y_{s,j} = \eta(\mathbf{x},\boldsymbol{\theta}_s) + m(\mathbf{x}|\boldsymbol{\phi}) + \sqrt{2(\sigma_\delta^2|\boldsymbol{\phi}_s + \sigma_\varepsilon^2|\boldsymbol{\phi}_s)}\;\xi_j, \quad \forall s\in S, j\in J. \tag{3.37c}$$

409 One disadvantage of optimization problem (Eq. (3.37)) is that the conditional GP mean and vari-

ance must be evaluated for all samples s from the posterior trace. This is computationally expensive and can be avoided by using a point estimate for ϕ such as the MAP.

Fig. 1 summarizes the alternative modeling and computational strategies in the proposed modeling framework.

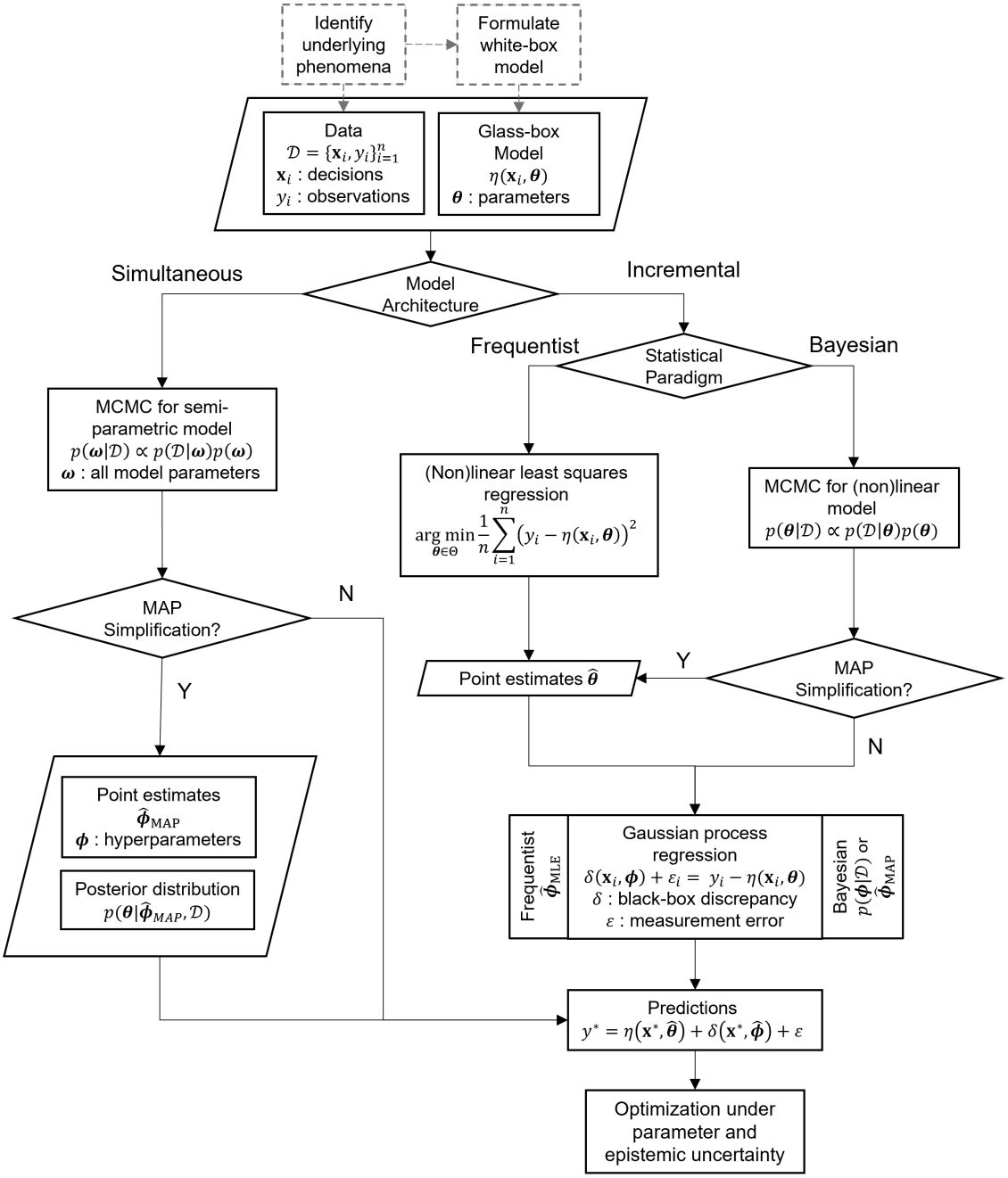


Figure 1: Proposed framework for hybrid model calibration and optimization. This framework supports both simultaneous and incremental hybrid model architectures as well as frequentist or Bayesian statistical paradigms.

4 Case study: Ballistics with Bayesian hybrid models

We compare several Bayesian hybrid model architectures and calibration methods in this first case study using a ballistics decision-making problem.

4.1 Problem statement, models, and data

Consider a projectile with constant and known mass m (kg) launched from the origin ($y = 0$, $z = 0$) using a cannon. The gunner manipulates two control variables: the firing angle ψ (°) and the initial velocity v_0 (m s^{-1}), to hit a target at $y^\dagger = 100$ m. However, the value of acceleration due to gravity g (m s^{-2}) and the full physics of projectile motion are unknown. Using information observed in prior experiments, the gunner must recommend a firing angle ψ^* and velocity v_0^* to hit the target despite parameter and epistemic uncertainty.

4.1.1 Ground truth model and training data

A complete physical model was used to simulate ground truth experiments. Assuming the projectile experiences drag due to air resistance, its flight may be described using four coupled differential equations along with initial conditions:

$$m\frac{dv_y}{dt} = -C_D v_y^2, \quad \frac{dy}{dt} = v_y, \quad v_y(0) = v_0\cos(\psi), \quad y(0) = 0 \tag{4.38}$$

$$m\frac{dv_z}{dt} = -m\,g - C_D v_z|v_z|, \quad \frac{dz}{dt} = v_z, \quad v_z(0) = v_0, \quad z(0) = 0 \tag{4.39}$$

where t (s) is time, C_D (kg m^{-1}) is the coefficient of drag and v_y and v_z (m s^{-1}) are projectile velocity's horizontal and vertical components, respectively. We solve this initial value problem by splitting the flight into two time domains: upward motion ($t \in [0, t_p]$) and downward motion ($t \in (t_p, t_f]$). We refer the reader to our prior work for a full description of the final solution for the

horizontal displacement of the projectile [101]:

$$\zeta(\mathbf{x}_i) = \frac{m}{C_D} \ln\left(C_D\, v_0\, \cos(\psi)\, t_f + m\right) \tag{4.40}$$

where $\mathbf{x}_i = [v_0, \psi]^\intercal$. The training data shown in Table 1 and Fig. 2 were generated via true process $\zeta(\cdot)$ using m = 1 kg, C_D = 0.01 kg m^{-1}, and g = 9.8 m s^{-2} and then corrupted with i.i.d. mean zero Gaussian measurement error with standard deviation $\sigma_\varepsilon = 5$.

Table 1: Training data for the projectile motion experiments.

Legend	v_0 (m s^{-1})	ψ (°)	y_i (m)
a	71	85	43.24
b	60	25.7	118.18
c	75	60	143.21
d	70	30	159.79
e	80	36	174.14
f	90	45	181.67

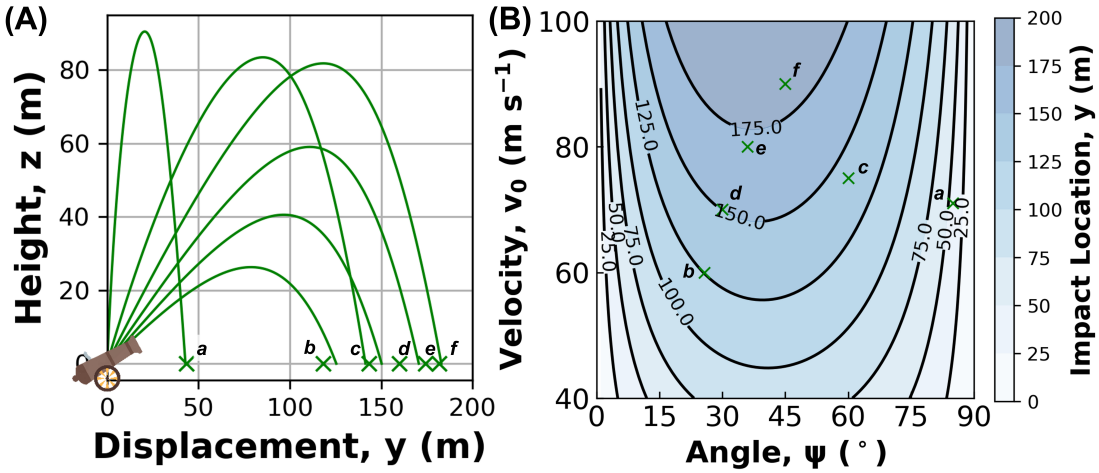


Figure 2: True trajectories simulated using the ground truth model. **(A)** Trajectories and impact locations calculated using ground truth simulations (Eq. (4.40)). Green ×'s are the six training data labeled a-f. **(B)** The firing velocity and angle are plotted on the vertical and horizontal axes, respectively. The contours of the heat map indicate the horizontal distance traveled by the projectile fired from location y = 0.

Fig. 2 shows the importance of air resistance when simulating the ballistic trajectory. Specifi-

cally, Fig. 2 (B) shows that at a constant velocity, firing the cannon using a shallow angle ($\psi < 45°$) instead of its steeper complement will result in farther horizontal displacement of the projectile. This case study aims to demonstrate how Bayesian hybrid models correct the bias of neglecting air resistance in the glass-box model.

4.1.2 Bayesian projectile motion models

We consider four model architectures for decision-making.

(1) The **simple physics model** predicts an ideal parabolic trajectory by neglecting air-resistance:

$$\eta(\mathbf{x}, \theta) = \frac{2v_0^2}{g} \sin(\psi)\cos(\psi). \tag{4.41}$$

Thus, the simple physics model is linear in parameters when $\theta = g^{-1}$. The hierarchical representation of the simple physics model is:

$$y|\theta, \varepsilon \sim \mathcal{N}(\eta(\mathbf{x}, \theta), \sigma_\epsilon^2), \tag{4.42a}$$

$$\varepsilon|\sigma_\varepsilon^2 \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{4.42b}$$

$$\theta \sim \mathcal{U}(10^{-3}, 1), \quad \sigma_\varepsilon \sim \mathcal{U}(1, 9). \tag{4.42c}$$

For all models, we use uniform priors to represent weak prior information about the behavior of the parameters in some known range of the values.

(2) The data-driven **GP-only model** does not explicitly use physical information to predict the distance traveled. The hierarhical model is:

$$y|\delta(\mathbf{x}), \varepsilon \sim \mathcal{GP}(0, \sigma_\delta^2 k(\mathbf{x}, \mathbf{x}'; \ell) + \sigma_\varepsilon^2 \Delta_{\mathbf{x},\mathbf{x}'}), \tag{4.43a}$$

$$\delta(\mathbf{x})|\sigma_\delta^2, \ell \sim \mathcal{GP}(0, \sigma_\delta^2 k(\mathbf{x}, \mathbf{x}'; \ell)), \quad \varepsilon|\sigma_\varepsilon^2 \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{4.43b}$$

$$\sigma_\delta \sim \mathcal{U}(1, 15), \quad \sigma_\varepsilon \sim \text{Gamma}(25, 5), \tag{4.43c}$$

$$\ell \sim \mathcal{U}(1, 15). \tag{4.43d}$$

Eq. (4.43a) says that at any point $\mathbf{x}$, $y(\mathbf{x})$ is Gaussian with a mean of 0 and variance $\sigma_\delta^2 k(\mathbf{x}, \mathbf{x}'; \ell) + \sigma_\varepsilon^2 \Delta_{\mathbf{x},\mathbf{x}'}$. In Eq. (4.43b) we set the mean of $\delta(\cdot)$ to 0 because we do not have a prior expectation that the observations are more likely to be positive than they are to be negative. The variance σ_δ^2 expresses a prior belief that $\delta(\cdot)$ is not likely to be outside the range $\pm 2\sigma_\delta$. The correlation $k(\cdot, \cdot)$ is governed by the Matern 3/2 kernel (Eq. (3.14)) with a single unknown length scale parameter ℓ. The wide range of the length scale ℓ reflects weak prior information about the smoothness of the function. The Gamma prior enforces that the standard deviation of the noise is strictly positive. In `PyMC`, the default Gamma distribution uses the rate-shape parameterization (i.e., $\mathbb{E}[\sigma_\varepsilon] = 5$, $\mathbb{V}[\sigma_\varepsilon] = 1$). Empirical evaluations of predictive performance demonstrated that the Matern 3/2 kernel was superior to the RBF kernel. Thus, the GP-only model uses the Matern 3/2 kernel.

(3) The **simultaneous hybrid model** augments the simple physics model $\eta(\cdot, \cdot)$ with a GP. The hierarchical model is:

$$y|\theta, \varepsilon \sim \mathcal{GP}(\eta(\mathbf{x}, \theta), \sigma_\delta^2 k(\mathbf{x}, \mathbf{x}'; \ell) + \sigma_\varepsilon^2 \Delta_{\mathbf{x},\mathbf{x}'}), \tag{4.44a}$$

$$\varepsilon|\sigma_\varepsilon^2 \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{4.44b}$$

$$\theta \sim \text{Gamma}(0.56, 5.6), \quad \sigma_\delta \sim \mathcal{U}(1, 15), \quad \sigma_\varepsilon \sim \text{Gamma}(25, 5), \tag{4.44c}$$

$$\ell \sim \mathcal{U}(1, 15). \tag{4.44d}$$

Here, $k(\cdot,\cdot)$ is the RBF kernel (Eq. (3.12)) with a single unknown length scale parameter ℓ. We used the RBF kernel because the simple physics model is smooth.

(4) The **incremental hybrid model** borrows the simple physics components of the simultaneous hybrid model in step one:

$$y|\theta,\varepsilon_\eta \sim \mathcal{N}(\eta(\mathbf{x},\theta),\tau_{\varepsilon_\eta}^{-1}), \tag{4.45a}$$

$$\varepsilon_\eta|\tau_{\varepsilon_\eta}^{-1} \sim \mathcal{N}(0,\tau_{\varepsilon_\eta}^{-1}), \tag{4.45b}$$

$$\theta \sim \text{Gamma}(0.56,5.6), \quad \tau_{\varepsilon_\eta}^{-1} \sim \text{Gamma}(5,5), \tag{4.45c}$$

where $\tau_{\varepsilon_\eta} = 1/\sigma_{\varepsilon_\eta}^2$ is the precision. In step two, the discrepancy function is fit to the residuals:

$$y-\eta(\mathbf{x},\hat{\theta})|\delta(\mathbf{x}),\varepsilon_\delta \sim \mathcal{GP}(0,\sigma_\delta^2 k(\mathbf{x},\mathbf{x}';\ell)+\sigma_{\varepsilon_\delta}^2\Delta_{\mathbf{x},\mathbf{x}'}), \tag{4.46a}$$

$$\delta(\mathbf{x})|\ell \sim \mathcal{GP}(0,\sigma_\delta^2 k(\mathbf{x},\mathbf{x}';\ell)), \quad \varepsilon_\delta|\sigma_{\varepsilon_\delta}^2 \sim \mathcal{N}(0,\sigma_{\varepsilon_\delta}^2), \tag{4.46b}$$

$$\sigma_\delta \sim \mathcal{U}(1,15), \quad \sigma_{\varepsilon_\delta} \sim \text{Gamma}(25,5), \tag{4.46c}$$

$$l \sim \mathcal{U}(1,15). \tag{4.46d}$$

To compare the incremental model with the simultaneous performance, the incremental model also used an RBF kernel.

4.1.3 Optimization under uncertainty for decision-making

We consider optimization problem (Eq. (3.37)) with the goal of hitting a target located at $y^\dagger$ = 100 m:

$$\eta_s = \frac{2\, v_0^2}{g_s}\, \sin(\psi)\, \cos(\psi), \quad \forall s \in S \tag{4.47a}$$

$$\delta_j = m(\mathbf{x}|\hat{\boldsymbol{\phi}}_{\text{MAP}}) + \xi_j \sqrt{2\sigma_\delta^2|\hat{\boldsymbol{\phi}}_{\text{MAP}}}, \quad \forall j \in J \tag{4.47b}$$

$$y_{s,j} = \eta_s + \delta_j + \varepsilon_j, \quad \Delta y_{s,j} = y_{s,j} - y^\dagger, \quad u_{s,j} = 1 - \frac{1}{y^\dagger}\min(|\Delta y_{s,j}|, y^\dagger), \quad \forall s \in S, j \in J. \tag{4.47c}$$

The utility function $u(\cdot)$ scales linearly from unity for a direct hit to zero for a miss of $y^\dagger$ m or more. In this work, we solved Eqs. (3.37) and (4.47) by enumerating solutions $v_0 \in [40, 100]$ (m s^{-1}) and $\psi \in [1, 90]$ (°) on a mesh grid 5490 points with a step size of 1 in each dimension to facilitate visualization. For larger problems, we recommend gradient-based computation optimization [102]. We highlight that the mechanistic model η_s is only indexed over posterior trace samples $s \in S$, and the GP plus observation error is indexed over quadrature nodes $j \in J$. Thus, the MAP approximation and Eq. (3.37) avoid the computational bottleneck of evaluating the GP prediction mean and variance for all trace samples.

4.1.4 True model error

To evaluate the proposed models, we define the true model error (true error) as:

$$y_{\text{err}}(\mathbf{x}) = y - y^*(\mathbf{x}) \tag{4.48}$$

where y^* is the hybrid model prediction and y is the data.

4.2 Results and discussion

Using the ballistic case study, we now demonstrate the advantages of the hybrid models and discuss best practices.

4.2.1 Hybrid models outperform the glass-box and black-box models by overcoming systematic bias utilizing limited data

Fig. 3 compares the performance of the Bayesian hybrid, simple physics, and GP-only models for decision-making. The left column, Fig. 3 (A-E), are the results from the simple physics model; the middle column, Fig. 3 (F-J), are the results using the GP-only model; and the right column, Fig. 3 (K-O), are results of the simultaneous hybrid model. Row 1, Fig. 3 (A, F, K), plots the distance traveled as predicted by the model. Row 2, Fig. 3 (B, G, L), shows the uncertainty in predictions. Row 3, Fig. 3 (C, H, M), is the objective value of the single-stage stochastic program. The optimum decisions are marked with purple dots. Row 4, Fig. 3 (D, I, N), shows y_{err} and the model used in the framework Fig. 3 (A, F, K). Row 5, Fig. 3 (E, J, O), is the absolute value of the true error plotted in row 4, which is representative of the utility function for the ballistic experiments.

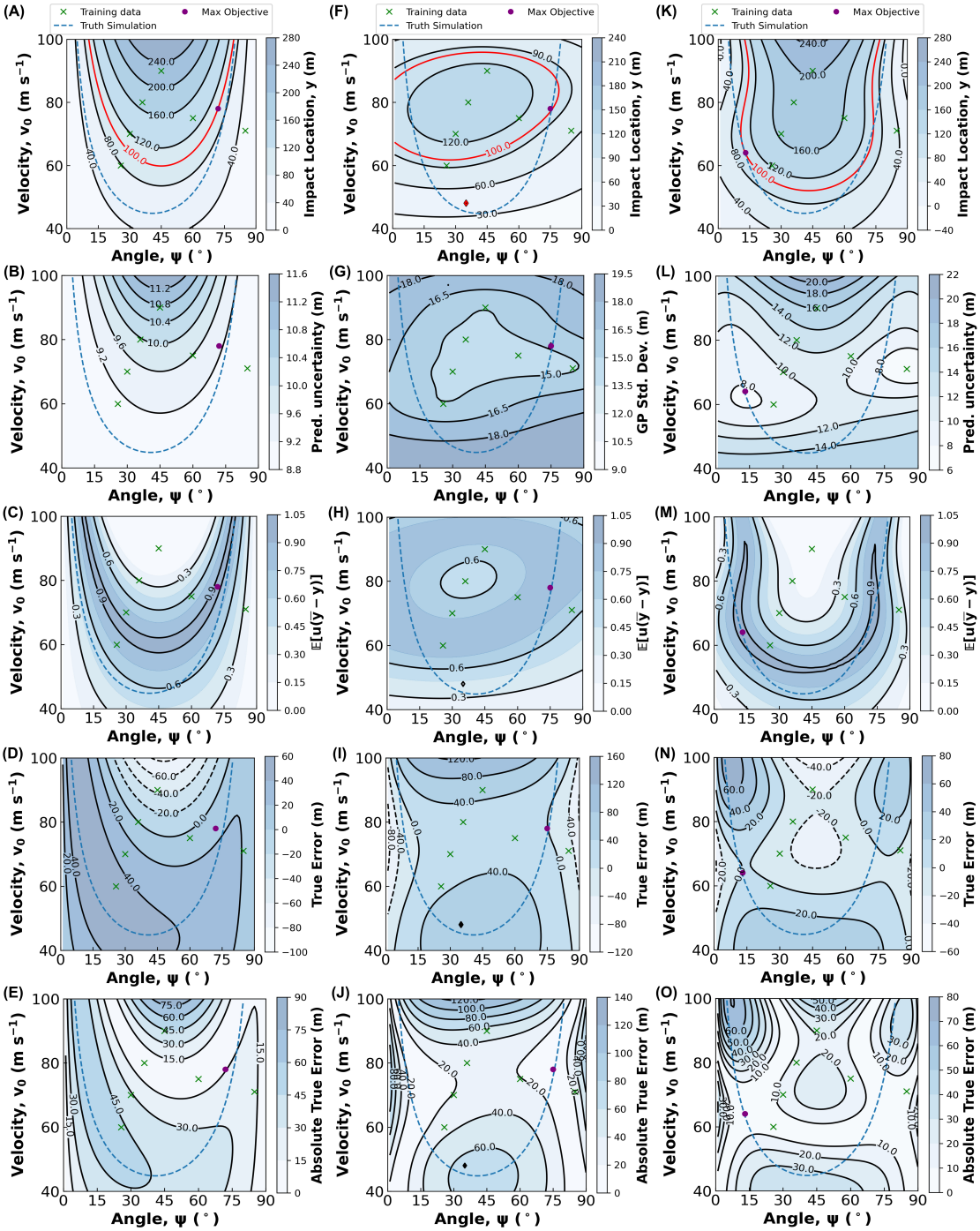


Figure 3: Results from the Bayesian decision-making framework. In each plot, the two control variables, firing velocity v_0 (m s^{-1}) and angle ψ (°), are plotted on the vertical and horizontal axes, respectively. The dashed blue contour shows true experimental conditions needed to satisfy the objective of hitting the target and corresponds to the $y^\dagger = 100$ m contour in Fig. 2. Green $\times$'s are the training data, and the purple dot is the optimum decision. The **left column (A-E)** shows the simple physics model, the **middle column (F-J)** shows the GP-only model and the **right column (K-O)** shows the hybrid model. **Row 1 (A, F, K)** shows the distance traveled predictions of the model. The red contour highlights decisions corresponding to the prediction $y = 100$ m. **Row 2 (B, G, L)** shows the uncertainties in the model prediction. **Row 3 (C, H, M)** displays the values of the optimization objective. **Row 4 (D, I, N)** shows the error between the ground truth and the model prediction. **Row 5 (E, J, O)** plots the absolute errors between the truth and model predictions.

The simple physics model misses the target with the highest absolute error of 14.2±9.3 m by recommending an experiment at ψ^*=72° and v_0^*=78 m s^{-1} as shown by the purple dot in Fig. 3 (A-E). As highlighted in Fig. 3 (A), the simple model predictions are symmetric around ψ= 45°, whereas the ground truth model (dashed blue line) which includes air resistance is not. Fig. 3 (B) shows the prediction uncertainty for the simple physics model increases with v_0. However, these uncertainties are much smaller than the absolute error shown in Fig. 3 (E) because the simple physics model neglects model-form (epistemic) uncertainty.

Decision-making using the data-driven GP-only model misses the target by 0.9±10 m with an experiment proposed at ψ^*=75° and v_0^*=78 m s^{-1}, shown by the purple dot in Fig. 3 (F-J). The order of magnitude improvement in absolute error over the simple physics model is because this surrogate model interpolates the optimum experiment between observations in the training set at [60°, 75 m s^{-1}] and [85°, 71 m s^{-1}] (Fig. 3 (J)). However, the exclusion of physical information in the GP-only framework leads to the possibility of several physically impractical predictions. For example, at ψ = 45°, increasing the velocity above 90 m s^{-1} (topmost green $\times$ in Fig. 3 (F)) leads to a decrease in the horizontal displacement of the projectile which is contrary to physical intuition irrespective of the consideration of air-resistance effects. Additionally, the GP cannot learn the sensitivity of impact location to firing angle ψ as noted by the mismatch between the elliptical and U-shaped contours of the GP-only model predictions and the truth, respectively. In general, due to their reliance on interpolation, pure ML models often perform poorly for prediction and decision-making using small datasets or weakly informative priors [103, 104, 105, 106, 107]. Experimental observations are often sparse or expensive for many (chemical) engineering applications, and informative priors are challenging to specify. These limitations motivate hybrid models.

The simultaneous hybrid model is superior with the least absolute prediction error of 0.5±7.9 m for an experiment recommended at ψ^*=13° and v_0^*=64 m s^{-1} shown by the purple dot in Fig. 3

(K-O). The hybrid model's glass-box (simple physics) component encodes physics-based information, e.g., the form of the ideal trajectory of the projectile. The black-box (GP) component leverages observations to learn the discrepancy between the simple glass-box model and the real process. The horseshoe-shaped contours in Fig. 3 (K) highlight the hybrid model predictions are a linear combination of the U-shaped simple physics model prediction and elliptical GP-only model prediction. Moreover, the simultaneous hybrid model's predictions, shown by the red contour in Fig. 3 (K), overlap the dashed blue truth contour at several points, in contrast to Fig. 3 (A, F). Fig. 3 (L) shows the hybrid model's prediction uncertainty also combines the shapes of the glass-box and black-box models, Fig. 3 (B, G). In Fig. 3 (L), there are two regions with prediction uncertainties less than 8 m, which are near training data. As a consequence of these improved predictions, Fig. 3 (M) shows that two segments of the truth contour (dashed blue) pass through regions with objective values above 0.9 or higher (recall that 1 is a direct hit, i.e., $y_{\text{err}} = 0$). Fig. 3 (N, O) shows the true error and absolute true error contours both incorporate visual features from similar contours for the glass-box Fig. 3 (D, E) and black-box Fig. 3 (I, J) models. Thus, in summary, Fig. 3 illustrates how hybrid models combine the advantages of both glass-box and black-box models to enable more accurate predictions with limited training data.

4.2.2 Explicitly accounting for epistemic uncertainty leads to better decision-making.

Next, we compare two instances of the hybrid model: M1, which is the standard model, and M2, ignores the GP uncertainty (i.e., treats it as a Kriging model). Fig. 4 compares the top 1% decisions calculated with M1 and M2. Fig. 4 (A, B) are parity plots that compare the absolute values of the true error $|y_{\text{err}}|$ and the model prediction error $|y^{\dagger} - y^{*}|$ on the horizontal and vertical axes respectively. As expected, the means and standard deviations of absolute prediction errors are larger for M1. Moreover, Fig. 4 (B) shows a handful of optimal decisions for M2 have absolute true errors

above 15 m, which is not the case for M1 in Fig. 4 (A). Fig. 4 (C) sorts these top 1% decisions for M1 and M2 from lowest to higher absolute error to facilitate comparisons. For example, approximately 78% of the top decisions for M1 have an absolute error of less than 6 m compared to only 61% for M2. Ultimately, Fig. 4 shows that including the GP prediction uncertainty in M1 helps safeguard against poor decisions.

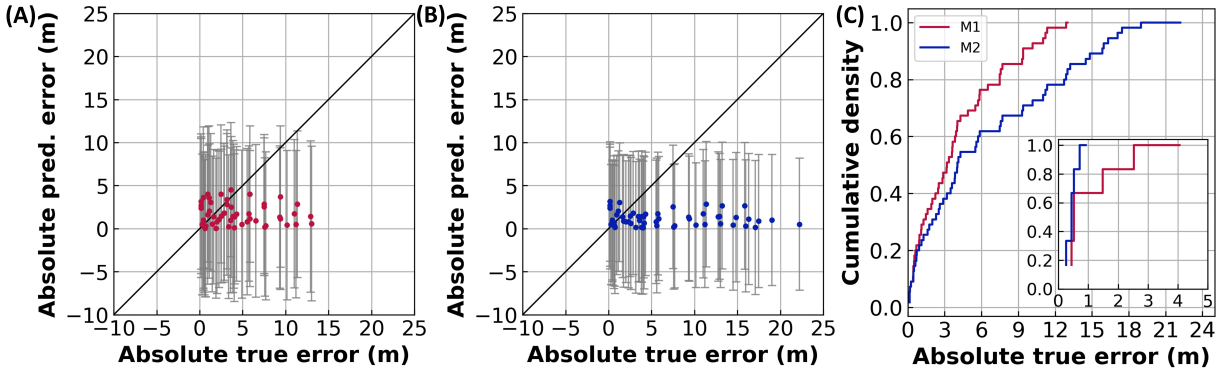


Figure 4: Considering epistemic uncertainty improves decision-making with hybrid models. The best 55 decisions (top 1% of objective values) are shown in this figure. Parity plots for the simultaneous hybrid model trained via **(A)** M1 (composite estimation) and **(B)** M2 (composite estimation neglecting GP uncertainty). The horizontal and vertical axes represent the absolute values of the true and model prediction errors, respectively. Each dot corresponds to a decision from the Bayesian framework, and the error bars indicate the uncertainty corresponding to the decision. **(C)** Compares M1 and M2 utilizing the cumulative density of the absolute true errors from the parity plots (A, B) and its inset repeats the same comparison using a smaller batch of top 0.1% of decisions.

To better understand the importance of epistemic uncertainty, Fig. 5 shows the objective and uncertainty contours for M2 and is analogous to Fig. 3 (K, L) for M1. Neglecting GP prediction uncertainty in M2 leads to large deviations between the best objective function value (darkest solid blue contours) and ground truth predictions (dashed blue contour) in regions (I), (II), and (III) of Fig. 5 (A). The prediction uncertainties of M2, shown in Fig. 5 (B), are now a function only of the parameter variability, which leads to a maximum of 12 m uncertainty in regions (I), (II), and (III). In the same regions of Fig. 3 (L), the inclusion of the GP uncertainty causes the total prediction uncertainty to be as high as 16 m. The low prediction uncertainty values from M2 are

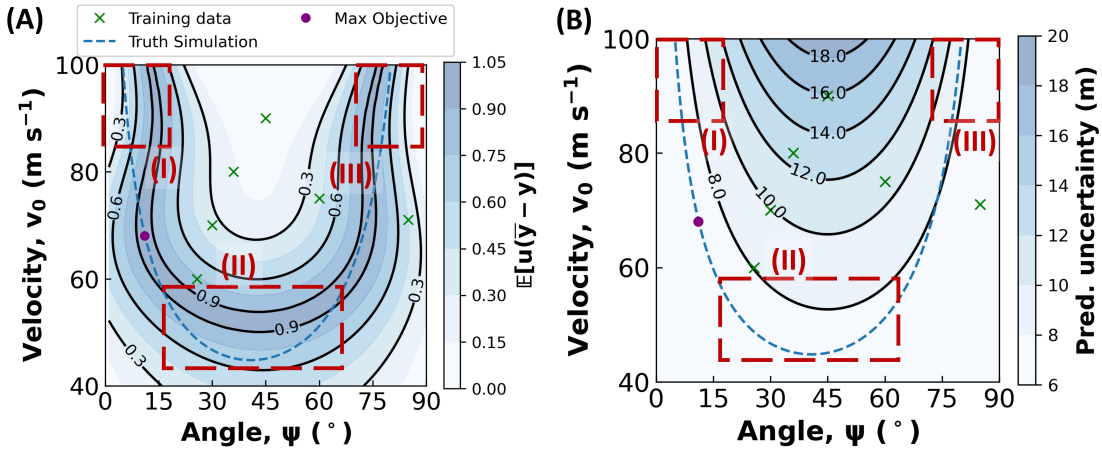


Figure 5: Neglecting epistemic uncertainty leads to large deviations from the truth due to low uncertainty predictions. In both figures, the two control variables of the ballistic experiment, firing velocity v_0 and angle ψ are plotted on the vertical and horizontal axes, respectively. The dashed blue contour shows true experimental conditions needed to satisfy the objective of hitting the target and corresponds to the 100 m contour in Fig. 2. Green $\times$'s are the training data used to calibrate the model and the purple dot is the optimum decision predicted by the framework. Simultaneous hybrid model prediction means **(A)** and uncertainties **(B)** using M2 (composite estimation neglecting GP uncertainty). Regions of largest deviation between the best model predictions (darkest solid blue contours corresponding to objective values ≥ 0.9) and the truth (dashed blue contour) are highlighted using red boxes and labeled (I), (II), and (III) and may be compared with analogous coordinates in Fig 3 (K, L, O). Note that Fig. 3 (O), the plot for absolute true errors is common for both M1 and M2.

not significant enough to impact decisions in these regions, which also correspond to true errors as high as 80 m as seen in Fig. 3 (O) thus leading to the higher maximum absolute true error for M2.

4.2.3 Decision-making with incremental hybrid models

A modeler faces several nuanced choices while building a Bayesian hybrid model, including the model structure. We now compare the performance of incremental hybrid models to the simultaneous hybrid models (analyzed previously). Fig. 6 explores an incremental M1 hybrid model and is analogous to the columns in Fig. 3. For completeness, Fig. S1 in the SM explores an incremental M2 hybrid model.

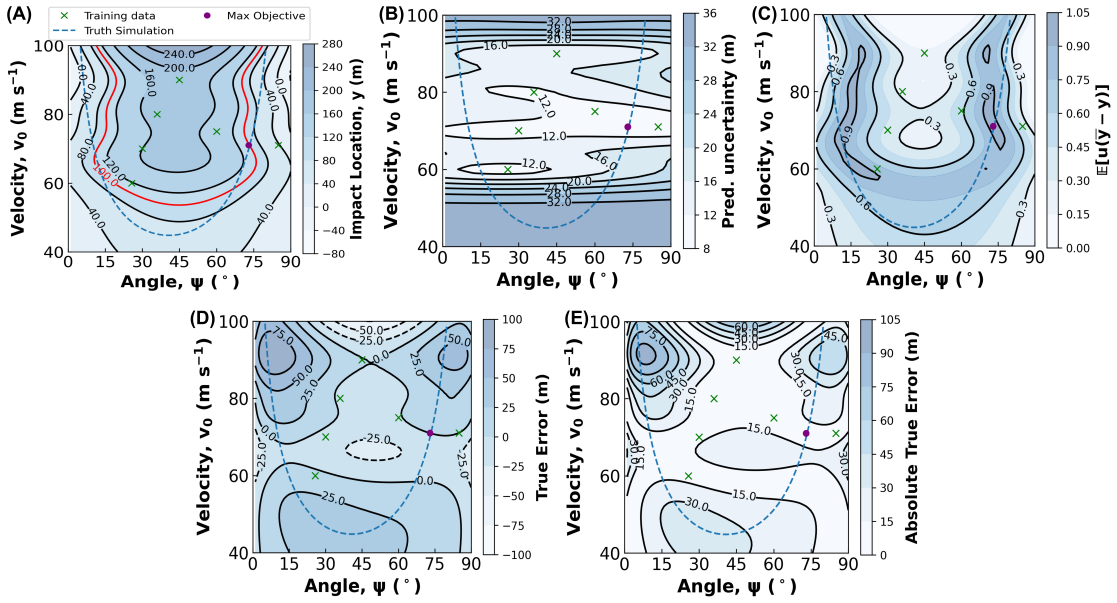


Figure 6: Results from the Bayesian decision-making framework using the incremental hybrid model and M1 (composite estimation). The control variables of the ballistics experiment, firing velocity v_0 and angle ψ, are plotted on the vertical and horizontal axes, respectively. The dashed blue contour shows true experimental conditions needed to satisfy the goal of the experiment, corresponding to the 100 m contour in Fig. 2. Green $\times$'s are the training data and the purple dot is the optimum recommendation of the framework. **(A)** incremental model predictions. **(B)** Uncertainty in incremental model predictions. **(C)** Objective values calculated via M1 (composite estimation). **(D)** True error. **(E)** Absolute true error.

The incremental hybrid model surpasses the simultaneous model with an absolute error of 0.1 ± 10.7 m for an experiment predicted at ψ^*=73° and v_0^*=71 m s^{-1} as shown by the purple dot in Fig. 6. As shown in Fig. 3 (K), the simultaneous model recommends a decision far away from the training data. In contrast, Fig. 6 (A) shows the incremental model recommends a decision between two observations (i.e., interpolates). Fig. 6 (B) shows the incremental model uncertainties and has a maximum value of 35.1 m at [45°, 41 m s^{-1}]; this is 1.6 times higher than the maximum simultaneous model uncertainty, which is the next highest and is shown in Fig. 3 (L). The regions of lowest uncertainty surround all but one of the data points, the exception lies in the second lowest region due to the influence of glass-box uncertainty. Outside of the range of the training set, near the top and bottom of Fig. 6 (B), the contour values rapidly increase and adopt a horizontal

shape further indicating the incremental model's inability to extrapolate the effect of firing angle ψ on the systematic bias. In general, the contours in Fig. 6 (A-E) (incremental model) have similar shapes to Fig. 3 (K-O). Recall, that the hybrid model shapes are a combination of the glass-box and black-box models (see SM Fig. S2), and for the incremental hybrid model, the influence of the black-box model is more pronounced (see SM Fig. S3). We hypothesize this is because the incremental hybrid model trains the GP after calibrating (and fixing) the glass-box model. As such, the black-box model needs to correct for larger residuals.

4.2.4 Summary

Finally, Fig. 7 compares the top 1% of decisions generated from all studied model architectures. The horizontal axis shows the absolute true error $|y_{\text{err}}|$. The vertical axis is the cumulative density, i.e., the fraction of top decisions with an absolute true error less than the value on the horizontal axis. Fig. 7, shows that the inadequate glass-box (simple physics) model makes the worst decisions, followed by the black-box (GP-only) model for the reasons discussed above. The incremental hybrid models perform the next best, and the simultaneous hybrid models perform the best. Thus, in summary, this case study shows the benefits of hybrid models for more accurate predictions and decision-making under uncertainty. The difference between including (M1) and neglecting (M2) GP uncertainty is less influential than using an incremental versus simultaneous hybrid model architecture. The incremental models sacrifice modest performance for computationally more straightforward inference calculations by decomposing the training into two steps. In the following case study, we explore nonlinear regression to infer parameters in the glass-box model. This further simplifies the computational workflow by eliminating the need for Bayesian model calibration for highly nonlinear models.

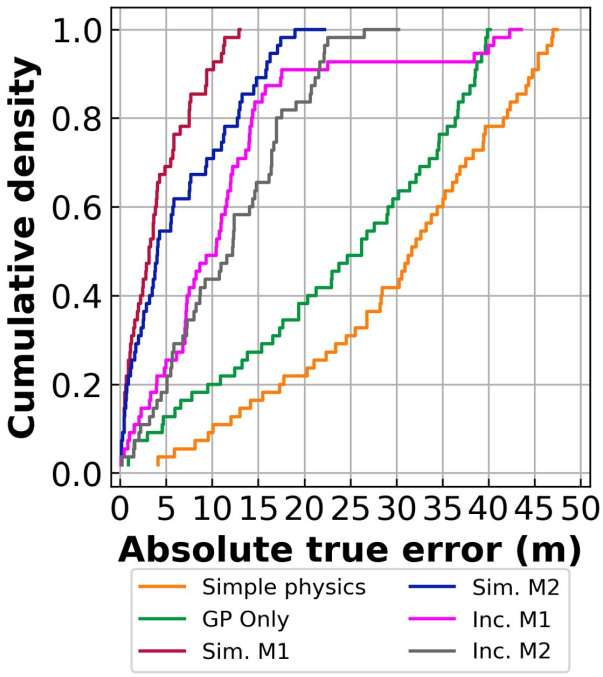


Figure 7: Hybrid models outperform the glass-box and black-box models in decision-making. The top 1% of decisions (N=55) recommended using alternate model architectures are compared in this figure utilizing the cumulative density of the absolute true errors. The red and blue are for the hybrid simultaneous models (M1 and M2), the pink and grey curves correspond to incremental hybrid models (M1 and M2), the green curve is for the GP-only model, and the black curve is for the simple physics model.

5 Case study: Reactor optimization

5.1 Problem statement, models, and data

As a second illustrative case study, we consider parameter and epistemic uncertainty in a series-like kinetic model used to optimize the controls of an isothermal batch reactor. In the kinetic model, species A reacts to form the desired product B via a non-elementary reversible reaction, which subsequently decomposes to undesired product C. The goal is to maximize the final concentration of B and minimize the concentrations of A and C without knowing the true kinetic mechanism of the productive reaction. This is done by manipulating three decision variables: the

starting concentration of species A, c_{A0} (M), the temperature of the reactor, T (K), and the duration of the reaction, t (h).

5.1.1 True kinetic model and training data

The true kinetic model (TKM) takes the form:

$$\mathrm{A} \underset{\kappa_3}{\overset{\kappa_1}{\rightleftharpoons}} \mathrm{B} \xrightarrow{\kappa_2} \mathrm{C} \tag{5.49}$$

where $\boldsymbol{\kappa} = [\kappa_1, \kappa_2, \kappa_3]$ are the rate constants for each reaction computed with the Arrhenius equation:

$$\kappa_h = \alpha_h \, \exp\left(-\frac{\nu_h}{RT}\right), \quad h \in \{1, 2, 3\}. \tag{5.50}$$

Here, $\boldsymbol{\alpha} = [200\ \mathrm{h}^{-1}, 100\ \mathrm{M}^{-2}\mathrm{h}^{-1}, 50\ \mathrm{h}^{-1}]$ are the pre-exponential factors, $\boldsymbol{\nu} = [10, 20, 15]$ (J mol^{-1}) are the activation energies, and R (J mol^{-1} K^{-1}) is the universal gas constant.

The rates of reaction are described by a system of ordinary differential equations:

$$\frac{dc_A}{dt} = -\kappa_1 c_A + \kappa_3 c_B, \tag{5.51}$$

$$\frac{dc_B}{dt} = \kappa_1 c_A - \kappa_2 c_B^3 - \kappa_3 c_B, \tag{5.52}$$

$$\frac{dc_C}{dt} = \kappa_2 c_B^3, \tag{5.53}$$

which are solved numerically with initial conditions

$$c_A(t=0) = c_{A0}, \quad c_B(t=0) = 0, \quad c_C(t=0) = 0. \tag{5.54}$$

In this formulation, we have assumed that the kinetics are non-elementary. Following notation introduced in Eq. (3.3), the vector-valued TKM $\boldsymbol{\zeta}(\cdot)$ is written as the numerical solution to the

ODE system:

$$\boldsymbol{\zeta}(\mathbf{x}_i) = [c_A(\mathbf{x}_i), c_B(\mathbf{x}_i), c_C(\mathbf{x}_i)], \tag{5.55}$$

$$\mathbf{x}_i = [t, T, c_{A0}], \tag{5.56}$$

$$t \in [0, 1], \quad T \in [293, 493], \quad c_{A0} \in [0, 5.0]. \tag{5.57}$$

The observation noise is assumed to be independent and identically distributed for all species and time, i.e.,:

$$\boldsymbol{\varepsilon}_i = [\varepsilon_{A,i}, \varepsilon_{B,i}, \varepsilon_{C,i}], \quad i \in \{1, \ldots, n\}, \tag{5.58}$$

$$\boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}, \tag{5.59}$$

where $\sigma_{\varepsilon} = 0.03$ M.

Following Eq. (3.3), the observations $\mathbf{y}_i$ are:

$$\mathbf{y}_i = \boldsymbol{\zeta}(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i. \tag{5.60}$$

Training data $\boldsymbol{\mathcal{D}}$ is generated by sampling the design space $T \times c_{A0}$ with a Latin hypercube design of twenty experiments. Experiments are randomly selected in sets of four to serve as the training data for ten different cases, as described in Table S1 of the SM.

5.1.2 Hybrid kinetic model

A simple kinetic model (SKM) has series kinetics,

$$\mathrm{A} \xrightarrow{\mathrm{k_1}} \mathrm{B} \xrightarrow{\mathrm{k_2}} \mathrm{C}\,, \tag{5.61}$$

where $\mathbf{k} = [k_1, k_2]$ (h^{-1}) is the rate constant of reaction and is computed with pre-exponential factor $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]$ (h^{-1}) and activation energy $\boldsymbol{\rho} = [\rho_1, \rho_2]$ (J mol^{-1}) following Eq. (5.50):

$$k_b = \lambda_b \exp\left(\frac{\rho_b}{RT}\right), \quad b \in \{1, 2\}. \tag{5.62}$$

The rates of reaction are derived assuming the proposed kinetics are elementary,

$$\frac{dc_A}{dt} = -k_1 c_A, \tag{5.63}$$

$$\frac{dc_B}{dt} = k_1 c_A - k_2 c_B, \tag{5.64}$$

$$\frac{dc_C}{dt} = k_2 c_B, \tag{5.65}$$

with initial conditions in Eq. (5.54). This system of equations has a known analytical solution:

$$c_A(t) = c_{A0} \exp\left(-k_1 t\right), \tag{5.66}$$

$$c_B(t) = \frac{k_1}{k_2 - k_1} c_{A0} \left(\exp(-k_1 t) - \exp(-k_2 t)\right), \tag{5.67}$$

$$c_C(t) = c_{A0} - c_B(t) - c_A(t). \tag{5.68}$$

Thus, the SKM $\boldsymbol{\eta}(\cdot, \cdot)$ is

$$\boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\theta}) = [c_A(\mathbf{x}_i, \boldsymbol{\theta}), c_B(\mathbf{x}_i, \boldsymbol{\theta}), c_C(\mathbf{x}_i, \boldsymbol{\theta})], \tag{5.69}$$

$$\boldsymbol{\theta} = [\boldsymbol{\lambda}^\intercal, \boldsymbol{\rho}^\intercal]. \tag{5.70}$$

In this workflow, nonlinear regression (Eqs. (3.25) and (3.26)) is used to fit observations generated from the decisions to the SKM in Eqs. (5.56) and (5.69), respectively, yielding estimates of the parameters $\hat{\boldsymbol{\theta}}$. Following frequentist inference of the SKM, the hybrid kinetic model (HKM)

discrepancy $\mathbf{a}\ \delta(\cdot)$ is trained on species B only, i.e., $\mathbf{a} = [0, 1, 0]$ in Eq. (3.3), using Bayesian GP regression. The GP was constructed with a zero mean function $m(\mathbf{x}) = 0$ and the Matern 3/2 kernel (Eq. (3.14)) for the covariance term $c(\mathbf{x}, \mathbf{x}'; \boldsymbol{\phi})$. The hierarchical model for Bayesian inference is thus:

$$y|\delta(\mathbf{x}), \varepsilon \sim \mathcal{GP}(0, \sigma_\delta^2 k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\ell}) + \sigma_\varepsilon^2 \Delta_{\mathbf{x},\mathbf{x}'}), \tag{5.71a}$$

$$\delta(\mathbf{x})|\boldsymbol{\phi} \sim \mathcal{GP}(0, \sigma_\delta^2 k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\ell})), \quad \varepsilon|\sigma_\varepsilon^2 \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{5.71b}$$

$$\sigma_\delta \sim \text{Gamma}(2.5, 1), \quad \sigma_\varepsilon \sim \text{Gamma}(9, 300), \tag{5.71c}$$

$$\ell_1, \ell_2, \ell_3 \overset{\text{ind.}}{\sim} \mathcal{U}(0.1, 1). \tag{5.71d}$$

In Eq. (5.71), we used the generalized RBF kernel (Eq. (3.13)) to encode a prior belief in the smoothness along each input variable. Moreover, a uniform prior over the length scale hyperparameters was chosen to reflect weak prior knowledge in the smoothness of the process. Gamma priors were chosen for the standard deviations because the Gamma distribution has a positive support. The `PyMC3` marginal likelihood implementation was used for GP regression. `PyMC3.sample()` provides the MAP estimate of the hyperparameters $\hat{\boldsymbol{\phi}}_{\text{MAP}}$ where $\boldsymbol{\phi} = [\sigma_\delta, \boldsymbol{\ell}^\intercal]$.

5.1.3 Optimization with the Hybrid Model

We now optimize the batch reactor to promote the formation and proliferation of the desired product B by manipulating reactor temperature and batch time:

$$\arg\max_{\mathbf{x}} \quad \mathbf{w}^\intercal \mathbf{c}(\mathbf{x}_i), \tag{5.72a}$$

$$\text{s.t.} \quad t \in [0, 1], T \in [293, 493], \tag{5.72b}$$

where $\mathbf{w} = [-1/2, 1, -1/2]^\intercal$ and the optimized concentration profiles $\mathbf{c}(\mathbf{x}_i)$ are generated from the model choice:

$$\mathbf{c}(\mathbf{x}_i) = \begin{cases} \boldsymbol{\eta}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) & \text{simple kinetics model,} \quad (5.73a) \\ \boldsymbol{\eta}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) + \mathbf{a}\, \delta(\mathbf{x}_i; \hat{\boldsymbol{\phi}}) & \text{hybrid kinetics model.} \quad (5.73b) \end{cases}$$

5.2 Results and discussion

5.2.1 Benchmarking hybrid kinetic model predictive performance

Fig. 8 compares the SKM (8a, solid lines) and HKM (8b, dashed lines) concentration predictions for case one at (i) 300, (ii) 350, (iii) 400, and (iv) 450 K holding the initial concentration at 2.0 M. Shaded regions represent point-wise prediction intervals constructed using a nominal coverage probability (NCP) of 68% [108]. The predicted concentration profiles of the remaining case studies can be found in SM Figs. S4-S12.

Fig. 9 reports the mean absolute error (MAE, pink), root mean squared error (RMSE, blue), and actual coverage probability (ACP) of the SKM (solid) and HKM (stripes) predictions made in Fig. 8 for case one. We define the ACP as the percent of total predictions within their respective prediction interval across all temperatures. This process was repeated for all ten case studies, whose respective bar charts are shown in SM Figs. S13-S21.

Figs. 8 and 9 show the structural inadequacy of the SKM. The model-form uncertainty violates the fundamental assumption of (non)linear least squares regression, thus motivating the HKM framework. In Fig. 8b, the prediction bands of the HKM capture all of the observations, whereas those of the SKM (Fig. 8a) do not. Furthermore, Fig. 9 reports the ACP of the SKM and HKM as 69.7% and 100%, respectively. All of the HKM and 80% of the SKM cases show conservative prediction, i.e., ACP $>$ NCP. Namely, the ACP of cases six and eight are 67.4% and 66.7% and are

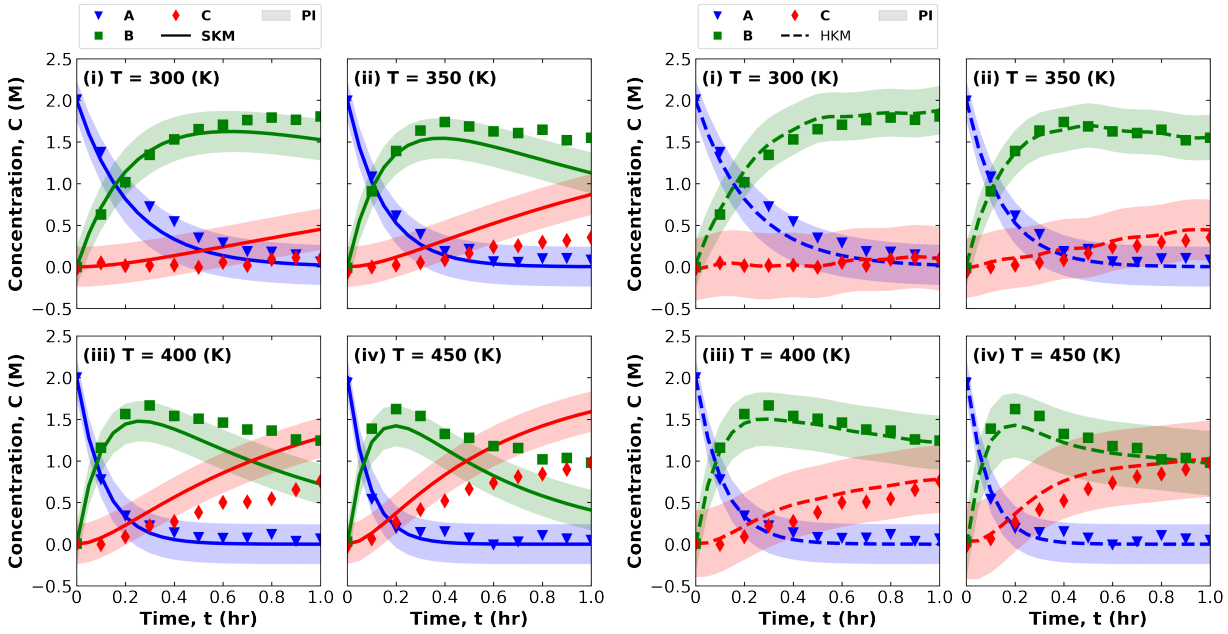


Figure 8: Case one concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C is shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).

not permissible as the maximum type I error (i.e., false positive) rate exceeds the nominal level.

Figs. 8 and 9 demonstrate that the HKM outperforms the SKM in predictive accuracy. In Fig. 8a, the SKM prediction lack of fit to the observations is evident. In contrast, the HKM predictions (Fig. 8b) better fit the observations, indicating that the GP successfully corrects for epistemic uncertainty in the SKM. This is supported by the MAE and RMSE values reported in Fig. 9. Moreover, Fig. 9 represents the observed trend in predictive error, in which the SKM error far exceeds the HKM error. Across all ten cases, we observed 97.5% of the HKM MAE values were less than those of SKM, averaging 2.6 times less error. Similarly, 95% of the HKM RMSE values were substantially less than those of the SKM, averaging 3.2 times lower error.

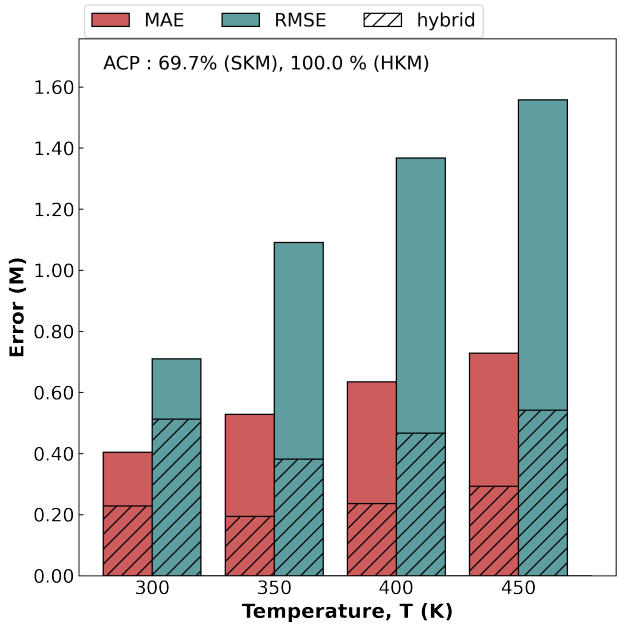


Figure 9: Case one mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

5.2.2 Optimization with the hybrid model

Next, we compared the HKM and the SKM for optimization under uncertainty. Table 2 shows the error, defined as positive (negative) for underprediction (overprediction), of each model in optimizing the objective function (Eq. (5.72a)) and the corresponding time and temperature decisions for all ten cases. An error value of zero occurs when the optimal value is at a global minimum (*) or maximum (**) in both the ground truth and model predictions. Experiments represent the collection of training data generated from the Latin hypercube design decisions (Table S1) for each case study.

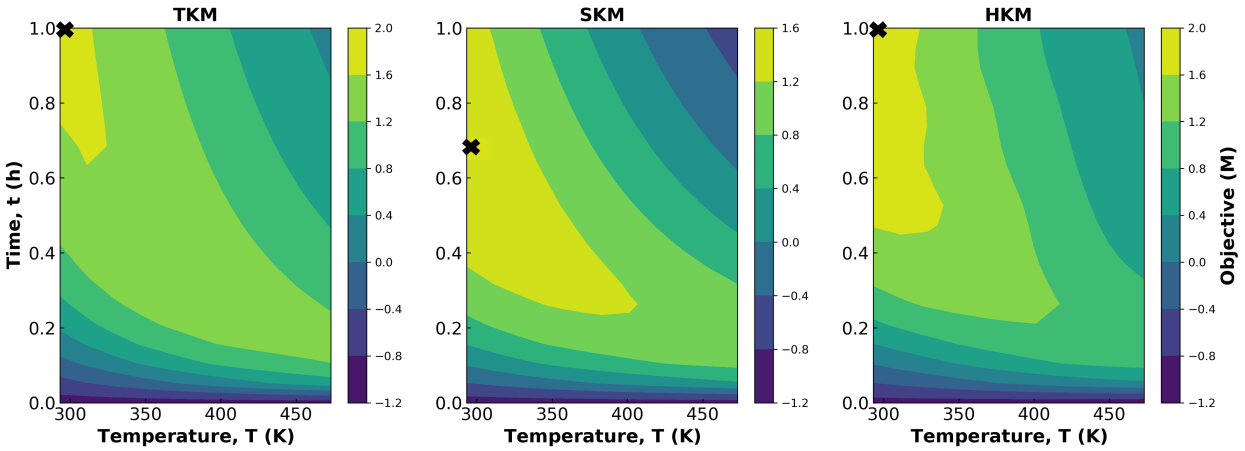
Table 2 confirms that superior predictive performance of the HKM translates to optimization under uncertainty. Here, the SKM underpredicts the TKM optimal time for all cases, whereas the HKM misses 70% with 4.5 times smaller MAE. Moreover, the HKM MAE for computing the maximum objective is 1.6 times smaller than that of the SKM. Cases five and nine are unique

	Temperature, T (K)		Time, t (h)		Objective (M)		
Case	SKM	HKM	SKM	HKM	SKM	HKM	Experiments
1	0.00*	0.00*	0.32	0.00**	0.22	-0.19	1, 10, 14, 19
2	0.00*	0.00*	0.37	0.11	0.44	-0.38	3, 4, 12, 18
3	0.00*	0.00*	0.16	0.11	0.27	-0.02	6, 7, 11, 13
4	0.00*	0.00*	0.21	0.00**	0.31	-0.23	1, 5, 15, 16
5	-180.00**	0.00*	0.79	0.16	0.31	-0.03	11, 12, 16, 17
6	0.00*	0.00*	0.16	0.11	0.25	-0.11	5, 13, 15, 20
7	0.00*	0.00*	0.26	0.16	0.31	-0.15	2, 6, 17, 20
8	0.00*	0.00*	0.21	0.11	0.24	-0.16	8, 17, 18, 20
9	-180.00**	0.00*	0.79	0.00**	0.16	-0.27	1, 5, 9, 16
10	0.00*	0.00*	0.26	0.11	0.10	0.09	9, 11, 13, 14

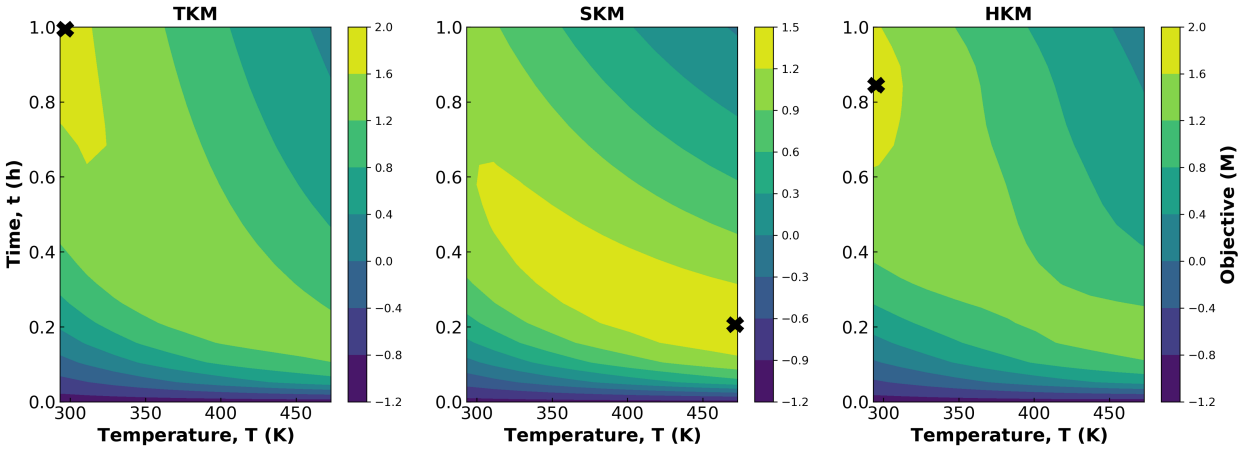
Table 2: SKM and HKM deviation from ground truth optimal time, temperature, and objective value for cases one through ten. Predictions made at global minima or maxima are indicated with * and **, respectively. A deviation value of zero occurs when an optimal value is at a global minimum or maximum for both the ground truth value and model prediction. Experiments represent the collection of training data from the Latin hypercube design.

in that the SKM predicts the optimal temperature at the global maximum instead of the TKM global minimum. We explain this behavior in further analysis and emphasize here that the HKM effectively corrects for this extreme model-form uncertainty.

To further illustrate observations made in the previous discussion, Fig. 10 compares the objective function landscape for the TKM, SKM, and HKM for cases one (Fig. 10a) and five (Fig. 10b). Case one (Fig. 10a) represents most of the SKM objective function landscapes in which the maximum objective is found at longer batch times and lower operating temperatures. Case five (Fig. 10b) was chosen arbitrarily over case nine and shows that the SKM optimal objective is achieved at shorter batch times and high-temperature operating conditions. The objective function landscapes of the remaining case studies are shown in SM Figs. S22-S29.



(a) Case One



(b) Case Five

Figure 10: Objective of the TKM, SKM, and HKM as a function of time and temperature for cases (a) one and (b) five. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

5.2.3 Why the simple kinetic model fails

To understand why the SKM predicts the optimal concentration at high temperatures and short times in cases five and nine, we turned to the parameter estimates and training data of the SKM. Table 3 reports the regressed parameters ρ_1, ρ_2, λ_1, and λ_2 of the SKM for all case studies. Fig. 11 plots the absolute true error of the objective as a function of initial concentration and temperature along with the training data (black X's) for cases one (Fig. 11a) and nine (Fig. 11b) at different points in time (top to bottom). For demonstration purposes, case nine was chosen over case five because it exhibited greater error. For consistency with prior analysis, the error of the maximum objective at 2.0 M of the TKM and SKM are indicated with a circle and triangle, respectively. In the SM, additional case studies are shown in Figs. S30-S37.

Together, Table 3 and Fig. 11 demonstrate that the SKM is not robust to random perturbations in the training data. Table 3 shows that the activation energy of the first reaction exceeds that of the second ($\lambda_1 > \lambda_2$) in cases five and nine only, which explains why the maximum objective was observed at high temperatures and long times in Fig. 10b. This is consistent with Fig. 11, in which the maximum absolute true error of case nine (Fig. 11b) exceeds that of case one (Fig. 11a). In addition, Fig. 11 demonstrates that the absolute true error increases with distance from the

Case	ρ_1 (s^{-1})	ρ_2 (s^{-1})	λ_1 (kJ mol^{-1})	λ_2 (kJ mol^{-1})
1	72.45	49.51	6.94	12.49
2	76.60	10.40	7.23	7.51
3	139.55	88.74	9.17	14.24
4	128.45	31.22	8.83	11.26
5	123.82	5.54	8.63	5.94
6	124.86	107.16	8.89	14.78
7	89.29	27.95	7.88	10.95
8	107.24	44.97	8.22	12.48
9	197.87	4.67	9.90	6.46
10	62.70	183.59	6.55	16.61

Table 3: Regressed parameters of the SKM for all ten cases.

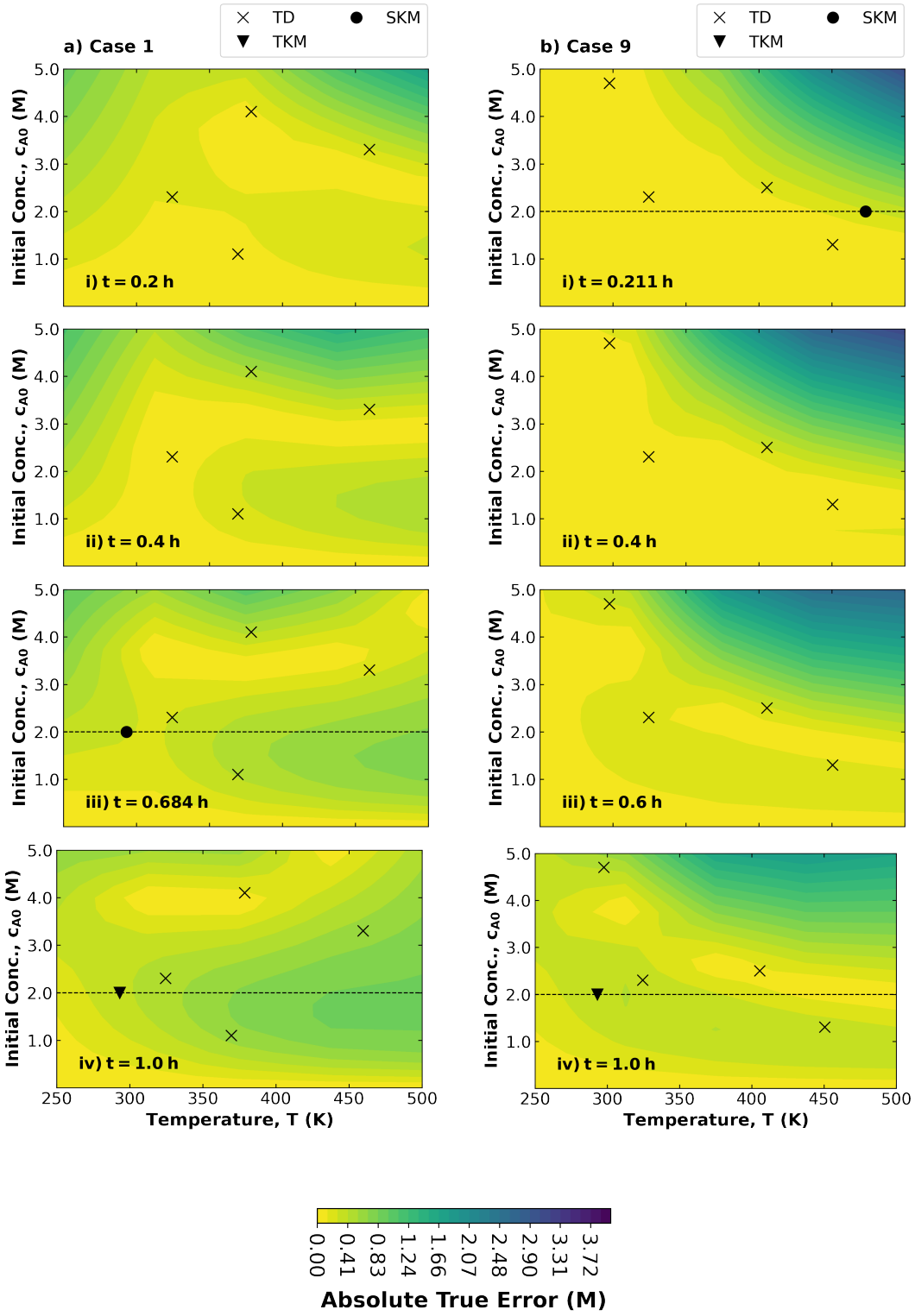


Figure 11: Absolute true error of the objective as a function of starting concentration of A, and temperature for case (a) one and (b) nine at four snapshots in time (top to bottom). Compared with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.

training data. This later observation highlights the need for more refined tools to enable the data-limited design of experiments under epistemic uncertainty.

6 Conclusions and outlook

This work proposes a Bayesian hybrid modeling computational framework for decision-making (optimization) considering aleatoric, parameter, and epistemic uncertainty for small datasets. Epistemic uncertainty is quantified using a Kennedy and O'Hagan-style Gaussian process (GP) discrepancy function, which guards against overfitting and overparameterization while providing a general hybrid model architecture that may be easily customized for many applications. The Kennedy and O'Hagan paradigm, closely related to Kriging surrogate models and grey-box hybrid models, has grown in popularity over the past decade, especially in chemical engineering. This work extends KOH hybrid models by using their posterior distribution (or an approximation) to define scenarios for stochastic programming.

This work uses two illustrative case studies — ballistics firing and reaction kinetics — to systematically explore Bayesian hybrid models for optimization under epistemic uncertainty. In both case studies, the hybrid model successfully integrates an inadequate glass-box model (e.g., neglects air resistance or assumes incorrect reaction pathways) with a GP discrepancy model to correct for the systematic bias. Both case studies show the hybrid model outperforms the alternative models with small datasets for decision-making under uncertainty. Moreover, we compare simultaneous versus incremental hybrid model architectures. We also discuss computational approximations to quickly calibrate hybrid models, which is essential for iterative model development and real-time control and optimization applications.

These two illustrative case studies highlight several future research directions beyond new applications. While both case studies consider only a single GP discrepancy function, the framework

can be extended to multiple GPs. Similarly, efficient gradient-based optimization with hybrid models is an open research area [102, 109]. Future work should explore the behavior of Bayesian hybrid models for sequences of experiments. Recently, we showed that physics-informed glass-box models for thickness control in additive manufacturing require significantly less data to train and globally gave more accurate predictions than black-box GP models [110]. Interestingly for this specific application, there was sufficient training data at thin film thickness such that the GP model was more accurate than the glass-box model. We suspect the glass-box models used in this paper and Wang et al. suffer from structural inadequacy that could not be overcome with additional data, i.e., the inadequate glass-box models will fail to converge to the true process, whereas a GP model will converge given enough data. This motivates several research questions related to the convergence of hybrid models, how to select the optimal sequence of experiments, and how to quantify the minimum dataset needed to reliably use a Bayesian hybrid model. Finally, this paper compares Bayesian hybrid models against GP models, one of the best ML architectures for small data sets. Future work should benchmark against more ML alternatives, such as deep neural networks, random forests, and ensembles, especially in higher dimensional case studies.

7 Computer code availability

Requests for copies of the computer codes for private use should be directed to the corresponding author.

8 Acknowledgements

NSF Award CBET-1941596 funded this project. E.A.E. gratefully acknowledges partial support from the Patrick and Jana Eiler's Graduate Fellowship. K.D.J. gratefully acknowledges partial

support from the Materials Innovation to Realize Innovative Technologies for Sustainability Fellowship. All authors are grateful for the support from the University of Notre Dame and the Center for Research Computing (CRC).

CRediT author statement

E.A.E. — Conceptualization (ballistics, reactions), Methodology (Bayesian studies for ballistics and reactions), Software (Bayesian studies for ballistics and reactions), Formal Analysis (ballistics), Validation (ballistics lead, reactions), Investigation (ballistics, reactions), Data-curation (ballistics, reactions), Writing – Original draft, Writing - Review & Edits, Visualization (ballistics).

K.D.J. — Conceptualization (frequentist, generalization, and integration), Methodology (frequentist, generalization), Software (frequentist, reactions), Formal Analysis (reactions), Validation (reactions lead), Investigation (reactions lead), Data-curation (reactions lead), Writing – Original draft (reactions, generalization of methods and notation), Writing - Review & Editing, Visualization (reactions).

X.G. — software (ballistics).

J.W. — software (ballistics).

A.W.D. — Conceptualization (lead), Methodology, Software, Formal analysis, Visualization, Writing – Original Draft, Writing – Review & Edits, Supervision, Project Administration, Funding acquisition

References

[1] Claire S. Adjiman et al. "Process Systems Engineering Perspective on the Design of Materials and Molecules". In: *Industrial & Engineering Chemistry Research* 60.14 (2021), pp. 5194–5206. ISSN: 0888-5885. DOI: 10.1021/acs.iecr.0c05399.

[2] Yuhe Tian et al. "An overview of process systems engineering approaches for process intensification: State of the art". In: *Chemical Engineering and Processing* 133.C (2018), pp. 160–210. DOI: 10.1016/j.cep.2018.07.014.

[3] John P. Eason and Lorenz T. Biegler. "Advanced trust region optimization strategies for glass box/black box models". In: *AIChE Journal* 64.11 (2018), pp. 3934–3943. DOI: 10.1002/aic.16364.

[4] Joel Sansana et al. "Recent trends on hybrid modeling for Industry 4.0". In: *Computers & Chemical Engineering* 151 (2021), p. 107365. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2021.107365. URL: https://www.sciencedirect.com/science/article/pii/S0098135421001435.

[5] Kevin McBride, Edgar Ivan Sanchez Medina, and Kai Sundmacher. "Hybrid Semi-parametric Modeling in Separation Processes: A Review". In: *Chemie Ingenieur Technik* 92.7 (2020), pp. 842–855. DOI: 10.1002/cite.202000025.

[6] Moritz von Stosch et al. "Hybrid semi-parametric modeling in process systems engineering: Past, present and future". In: *Computers & Chemical Engineering* 60 (2014), pp. 86–101. DOI: 10.1016/j.compchemeng.2013.08.008.

[7] George E. P Box, Alberto Luceño, and María del Carmen Paniagua-Quiñones. *Statistical Control by Monitoring and Adjustment*. 2nd ed. John Wiley & Sons, 2009.

[8] Marc C. Kennedy and Anthony O'Hagan. "Bayesian calibration of computer models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 425–464. DOI: 10.1111/1467-9868.00294. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00294.

[9] B.E. Ydstie. "Forecasting and control using adaptive connectionist networks". In: *Computers & Chemical Engineering* 14.4 (1990), pp. 583–599. DOI: 10.1016/0098-1354(90)87029-O.

[10] M.A. Kramer. "Autoassociative neural networks". In: *Computers & Chemical Engineering* 16.4 (1992), pp. 313–328. DOI: 10.1016/0098-1354(92)80051-A.

[11] Bhavik R. Bakshi and George Stephanopoulos. "Wave-net: a multiresolution, hierarchical neural network with localized learning". In: *AIChE Journal* 39.1 (1993), pp. 57–81. DOI: 10.1002/aic.690390108.

[12] Jay H. Lee, Joohyun Shin, and Matthew J. Realff. "Machine learning: Overview of the recent progress and implications for the process systems engineering field". In: *Computers & Chemical Engineering* 114 (2018), pp. 111–121. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2017.10.008.

[13] Nicholas E. Jackson, Michael A. Webb, and Juan J. de Pablo. "Recent advances in machine learning towards multiscale soft materials design". In: *Current Opinion in Chemical Engineering* 23.C (2019), pp. 106–114. DOI: 10.1016/j.coche.2019.03.005.

[14] Mojtaba Haghighatlari and Johannes Hachmann. "Advances of machine learning in molecular modeling and simulation". In: *Current Opinion in Chemical Engineering* 23 (2019), pp. 51–57. DOI: 10.1016/j.coche.2019.02.009.

[15] Chao Ning and Fengqi You. "Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming". In: *Computers & Chemical Engineering* 125 (2019), pp. 434–448. DOI: 10.1016/j.compchemeng.2019.03.034.

[16] Sohrab Zendehboudi, Nima Rezaei, and Ali Lohi. "Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review". In: *Applied Energy* 228 (2018), pp. 2539–2566. DOI: 10.1016/j.apenergy.2018.06.051.

[17] Dimitris C. Psichogios and Lyle H. Ungar. "A hybrid neural network-first principles approach to process modeling". In: *AIChE Journal* 38.10 (1992), pp. 1499–1511. DOI: doi.org/10.1002/aic.690381003.

[18] Michael L. Thompson and Mark A. Kramer. "Modeling chemical processes using prior knowledge and neural networks". In: *AIChE Journal* 40.8 (1994), pp. 1328–1340. DOI: doi.org/10.1002/aic.690400806.

[19] L.F.M. Zorzetto, R.M. Filho, and M.R. Wolf-Maciel. "Processing modelling development through artificial neural networks and hybrid models". In: *Computers & Chemical Engineering* 24.2 (2000), p. 1355. DOI: 10.1016/S0098-1354(00)00419-1.

[20] D orte Solle et al. "Between the Poles of Data-Driven and Mechanistic Modeling for Process Operation". In: *Chemie Ingenieur Technik* 89.5 (2017), pp. 542–561. DOI: 10.1002/cite.201600175.

[21] Haiyu Qi et al. "A hybrid neural network-first principles model for fixed-bed reactor". In: *Chemical Engineering Science* 54.13 (1999), pp. 2521–2526. DOI: 10.1016/S0009-2509(98)00523-5.

[22] G. Zahedi et al. "Hybrid artificial neural network-First principle model formulation for the unsteady state simulation and analysis of a packed bed reactor for CO_2 hydrogenation to methanol". In: *Chemical Engineering Journal* 115.1 (2005), pp. 113–120. DOI: 10.1016/j.cej.2005.08.018.

[23] Meisam Babanezhad et al. "Liquid-phase chemical reactors: Development of 3D hybrid model based on CFD-adaptive network-based fuzzy inference system". eng. In: *Canadian Journal of Chemical Engineering* 97.S1 (2019), pp. 1676–1684. ISSN: 0008-4034. DOI: 10.1002/cjce.23378.

[24] Amir Mosavi et al. "Prediction of multi-inputs bubble column reactor using a novel hybrid model of computational fluid dynamics and machine learning". In: *Engineering Applications of Computational Fluid Mechanics* 13.1 (2019), pp. 482–492. DOI: 10.1080/19942060.2019.1613448.

[25] A.A. Safavi, A. Nooraii, and J.A. Romagnoli. "A hybrid model formulation for a distillation column and the online optimisation study". In: *Journal of Process Control* 9.2 (1999), pp. 125–134. DOI: 10.1016/S0959-1524(98)00041-9.

[26] Francis J. Doyle III, Christopher A. Harrison, and Timothy J. Crowley. "Hybrid model-based approach to batch-to-batch control of particle size distribution in emulsion polymerization". In: *Computers & Chemical Engineering* 27.8 (2003), pp. 1153–1163. DOI: 10.1016/S0098-1354(03)00043-7.

[27] P. Georgieva, M.J. Meireles, and S. Feyo de Azevedo. "Knowledge-based hybrid modelling of a batch crystallisation when accounting for nucleation, growth and agglomeration phenomena". In: *Chemical Engineering Science* 58.16 (2003), pp. 3699–3713. DOI: 10.1016/S0009-2509(03)00260-4.

[28] Mohammed Saad Faizan Bangi and Joseph Sang-Il Kwon. "Deep hybrid modeling of chemical process: Application to hydraulic fracturing". In: *Computers & Chemical Engineering* 134.C (2020), p. 106696. DOI: 10.1016/j.compchemeng.2019.106696.

[29] Fan Yang et al. "A hybrid deep learning and mechanistic kinetics model for the prediction of fluid catalytic cracking performance". In: *Chemical Engineering Research & Design* 155 (2020), pp. 202–210. DOI: 10.1016/j.cherd.2020.01.013.

[30] Debanjan Ghosh, Jesús Moreira, and Prashant Mhaskar. "Model Predictive Control Embedding a Parallel Hybrid Modeling Strategy". In: *Industrial & Engineering Chemistry Research* 60.6 (2021), pp. 2547–2562. DOI: 10.1021/acs.iecr.0c05208.

[31] Shu Yang et al. "Hybrid Modeling in the Era of Smart Manufacturing". In: *Computers & Chemical Engineering* 140.C (2020), p. 106874. DOI: 10.1016/j.compchemeng.2020.106874.

[32] William Bradley and Fani Boukouvala. "Two-Stage Approach to Parameter Estimation of Differential Equations Using Neural ODEs". In: *Industrial & Engineering Chemistry Research* 60.45 (2021), pp. 16330–16344. DOI: 10.1021/acs.iecr.1c00552.

[33] Bram De Jaegher et al. "Enhancing mechanistic models with neural differential equations to predict electrodialysis fouling". In: *Separation and Purification Technology* 259 (2021), p. 118028. DOI: 10.1016/j.seppur.2020.118028.

[34] Benjamin Bayer, Gerald Striedner, and Mark Duerkop. "Hybrid Modeling and Intensified DoE: An Approach to Accelerate Upstream Process Characterization". In: *Biotechnology Journal* 15.9 (2020), e2000121. DOI: 10.1002/biot.202000121.

[35] Anjana Puliyanda et al. "A review of automated and data-driven approaches for pathway determination and reaction monitoring in complex chemical systems". In: *Digital Chem-*

ical Engineering 2 (2022), p. 100009. ISSN: 2772-5081. DOI: 10.1016/j.dche.2021.100009. URL: https://www.sciencedirect.com/science/article/pii/S2772508121000090.

[36] Alessandra Saraceno et al. "A hybrid neural approach to model batch fermentation of "ricotta cheese whey" to ethanol". In: *Computers & Chemical Engineering* 34.10 (2010), pp. 1590–1596. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2009.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0098135409002828.

[37] Abbas Azarpour et al. "Catalytic activity evaluation of industrial Pd/C catalyst via gray-box dynamic modeling and simulation of hydropurification reactor". In: *Applied Catalysis A: General* 489 (2015), pp. 262–271. ISSN: 0926-860X. DOI: 10.1016/j.apcata.2014.10.048. URL: https://www.sciencedirect.com/science/article/pii/S0926860X14006711.

[38] Abbas Azarpour et al. "A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors". In: *Chemical Engineering Research and Design* 117 (2017), pp. 149–167. ISSN: 0263-8762. DOI: 10.1016/j.cherd.2016.10.024. URL: https://www.sciencedirect.com/science/article/pii/S0263876216303549.

[39] Linh Bui et al. "A Hybrid Modeling Approach for Catalyst Monitoring and Lifetime Prediction". In: *ACS Engineering Au* 2.1 (2022), pp. 17–26. DOI: 10.1021/acsengineeringau.1c00015. eprint: https://doi.org/10.1021/acsengineeringau.1c00015. URL: https://doi.org/10.1021/acsengineeringau.1c00015.

[40] Y. Bard. *Nonlinear Parameter Estimation*. Academic Press, 1974. ISBN: 9780120782505. URL: https://books.google.com/books?id=fNo6MR1iS1UC.

[41] George B Arhonditsis et al. "Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management". In: *Journal of Marine Systems* 73.1-2 (2008), pp. 8–30. DOI: doi.org/10.1016/j.jmarsys.2007.07.004.

[42] A Van Griensven and T Meixner. "A global and efficient multi-objective auto-calibration and uncertainty estimation method for water quality catchment models". In: *Journal of Hydroinformatics* 9.4 (2007), pp. 277–291. DOI: doi.org/10.2166/hydro.2007.104..

[43] Ying Xiong et al. "A better understanding of model updating strategies in validating engineering models". In: *Computer methods in applied mechanics and engineering* 198.15-16 (2009), pp. 1327–1337. DOI: doi.org/10.1016/j.cma.2008.11.023.

[44] Wei Chen et al. "A design-driven validation approach using Bayesian prediction models". In: *Journal of Mechanical Design* 130.2 (2008), p. 021101. DOI: doi.org/10.1115/1.2809439.

[45] Piyush M Tagade, Byeong-Min Jeong, and Han-Lim Choi. "A Gaussian process emulator approach for rapid contaminant characterization with an integrated multizone-CFD model". In: *Building and Environment* 70 (2013), pp. 232–244. DOI: doi.org/10.1016/j.buildenv.2013.08.023.

[46] Piyush Tagade et al. "Bayesian calibration for electrochemical thermal model of lithium-ion cells". In: *Journal of Power Sources* 320 (2016), pp. 296–309. DOI: doi.org/10.1016/j.jpowsour.2016.04.106.

[47] Sayan Ghosh et al. "Advances in Bayesian Probabilistic Modeling for Industrial Applications". eng. In: *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B. Mechanical Engineering* 6.3 (2020). DOI: https://doi.org/10.1115/1.4046747.

[48] Gordon H ullen et al. "Managing uncertainty in data-driven simulation-based optimization". In: *Computers & Chemical Engineering* 136 (2020), p. 106519. DOI: 10.1016/j.compchemeng.2019.106519.

[49] David C. Miller et al. "Carbon Capture Simulation Initiative: A Case Study in Multiscale Modeling and New Challenges". In: *Annual Review of Chemical and Biomolecular Engineering* 5.1 (2014), pp. 301–323. DOI: 10.1146/annurev-chembioeng-060713-040321.

[50] Jayashree Kalyanaraman et al. "Uncertainty quantification via Bayesian inference using sequential Monte Carlo methods for CO_2 adsorption process". In: *AIChE Journal* 62.9 (2016), pp. 3352–3368. DOI: 10.1002/aic.15381.

[51] David S Mebane et al. "Bayesian calibration of thermodynamic models for the uptake of CO_2 in supported amine sorbents using ab initio priors". In: *Physical Chemistry Chemical Physics* 15.12 (2013), pp. 4355–4366. DOI: doi.org/10.1039/c3cp42963f..

[52] K. Sham Bhat et al. "Upscaling Uncertainty with Dynamic Discrepancy for a Multi-Scale Carbon Capture System". In: *Journal of the American Statistical Association* 112.520 (2017), pp. 1453–1467. URL: 10.1080/01621459.2017.1295863.

[53] Curtis B Storlie et al. "Calibration of Computational Models With Categorical Parameters and Correlated Outputs via Bayesian Smoothing Spline ANOVA". In: *Journal of the American Statistical Association* 110.509 (2015), pp. 68–82. DOI: doi.org/10.1080/01621459.2014.979993.

[54] Kuijun Li et al. "Multiscale modeling of an amine sorbent fluidized bed adsorber with dynamic discrepancy reduced modeling". In: *Reaction Chemistry & Engineering* 2.4 (2017), pp. 550–560. DOI: 10.1039/c7re00040e.

[55] Anca Ostace et al. "Probabilistic Model Building with Uncertainty Quantification and Propagation for a Dynamic Fixed Bed CO_2 Capture Process". doi.org/10.1021/acs.energyfuels.9b03250. In: *Energy & Fuels* 34.2 (2020), pp. 2516–2532.

[56] Donovan Chaffart, Shabnam Rasoulian, and Luis A. Ricardez-Sandoval. "Distributional uncertainty analysis and robust optimization in spatially heterogeneous multiscale process systems". In: *AIChE Journal* 62.7 (2016), pp. 2374–2390. DOI: https://doi.org/10.1002/aic.15215. eprint: https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.15215. URL: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.15215.

[57] Grigoriy Kimaev, Donovan Chaffart, and Luis A. Ricardez-Sandoval. "Multilevel Monte Carlo applied for uncertainty quantification in stochastic multiscale systems". In: *AIChE Journal* 66.8 (2020), e16262. DOI: https://doi.org/10.1002/aic.16262. eprint: https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.16262. URL: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.16262.

[58] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[59] J Steimel and S Engell. "Conceptual design and optimization of chemical processes under uncertainty by two-stage programming". In: *Computers & Chemical Engineering* 81 (2015), pp. 200–217. DOI: doi.org/10.1016/j.compchemeng.2015.05.016.

[60] Luis J Zeballos, Carlos A Méndez, and Ana P Barbosa-Povoa. "Design and Planning of Closed-Loop Supply Chains: A Risk-Averse Multistage Stochastic Approach". In: *Industrial*

978 *& Engineering Chemistry Research* 55.21 (2016), pp. 6236–6249. DOI: doi.org/10.1021/
979 acs.iecr.5b03647..

980 [61] Xinyue Peng, Thatcher W Root, and Christos T Maravelias. "Optimization-based process
981 synthesis under seasonal and daily variability: Application to concentrating solar power".
982 In: *AIChE Journal* 65.7 (2019), e16458.

983 [62] Jose A Renteria et al. "Optimal PID controller tuning using stochastic programming tech-
984 niques". In: *AIChE Journal* 64.8 (2018), pp. 2997–3010.

985 [63] Can Li and Ignacio E. Grossmann. "A Review of Stochastic Programming Methods for
986 Optimization of Process Systems Under Uncertainty". In: *Frontiers in Chemical Engineering*
987 2 (2021). DOI: doi.org/10.3389/fceng.2020.622241.

988 [64] Ruiwei Jiang and Yongpei Guan. "Data-driven chance constrained stochastic program". In:
989 *Mathematical Programming* 158.1-2 (2016), pp. 291–327. DOI: doi.org/10.1007/s10107-
990 015-0929-7.

991 [65] Abebe Geletu et al. "Advances and applications of chance - constrained approaches to
992 systems optimisation under uncertainty". In: *International Journal of Systems Science* 44.7
993 (2013), pp. 1209–1232. DOI: doi.org/10.1080/00207721.2012.670310.

994 [66] Abebe Geletu and Pu Li. "Recent Developments in Computational Approaches to Opti-
995 mization under Uncertainty and Application in Process Systems Engineering". In: *Chem-*
996 *BioEng Reviews* 1.4 (2014), pp. 170–190. DOI: doi.org/10.1002/cben.201400013.

997 [67] Ignacio E Grossmann et al. "Recent advances in mathematical programming techniques
998 for the optimization of process systems under uncertainty". In: *Computers & Chemical Engi-*
999 *neering* 91 (2016), pp. 3–14. DOI: oi.org/10.1016/j.compchemeng.2016.03.002.

[68] Chao Shang and Fengqi You. "Distributionally robust optimization for planning and scheduling under uncertainty". In: *Computers & Chemical Engineering* 110 (2018), pp. 53–68.

[69] Chao Ning and Fengqi You. "Data-driven adaptive nested robust optimization: General modeling framework and efficient computational algorithm for decision making under uncertainty". In: *AIChE Journal* 63.9 (2017), pp. 3790–3817. DOI: doi.org/10.1002/aic.15717.

[70] Gene A Bunin. "Constraint back-offs for safe, sufficient excitation: A general theory with application to experimental optimization". In: *Computers & Chemical Engineering* 93 (2016), pp. 353–360. DOI: doi.org/10.1016/j.compchemeng.2016.07.006.

[71] Kai Liu et al. "Dynamic optimization of natural gas pipeline networks with demand and composition uncertainty". In: *Chemical Engineering Science* 215 (2020), p. 115449. DOI: doi.org/10.1016/j.ces.2019.115449.

[72] Johannes Maußner and Hannsjörg Freund. "Optimization under uncertainty in chemical engineering: Comparative evaluation of unscented transformation methods and cubature rules". In: *Chemical Engineering Science* 183 (2018), pp. 329–345. DOI: doi.org/10.1016/j.ces.2018.02.002.

[73] B. Srinivasan et al. "Dynamic optimization of batch processes: II. Role of measurements in handling uncertainty". In: *Computers & Chemical Engineering* 27.1 (2003), pp. 27–44. DOI: 10.1016/S0098-1354(02)00117-5.

[74] Johannes Maußner and Hannsjörg Freund. "Efficient calculation of constraint back-offs for optimization under uncertainty: A case study on maleic anhydride synthesis". eng. In: *Chemical Engineering Science* 192 (2018), pp. 306–317. DOI: doi.org/10.1016/j.ces.2018.06.079.

[75] Jun Shi, Lorenz T. Biegler, and Intan Hamdan. "Optimization of grade transitions in polyethylene solution polymerization processes". In: *AIChE Journal* 62.4 (2016), pp. 1126–1142. DOI: doi.org/10.1002/aic.15113.

[76] Victor N. Emenike et al. "Robust dynamic optimization of enzyme-catalyzed carboligation: A point estimate-based back-off approach". In: *Computers & Chemical Engineering* 121 (2019), pp. 232–247. DOI: doi.org/10.1016/j.compchemeng.2018.10.006.

[77] Elton Pan et al. "Constrained model-free reinforcement learning for process optimization". In: *Computers & Chemical Engineering* 154 (2021), p. 107462. DOI: doi.org/10.1016/j.compchemeng.2021.107462.

[78] Federico Galvanin et al. "Backoff strategy for model-based experiment design under parametric uncertainty". In: *AIChE Journal* 56.8 (2010), pp. 2088–2102. DOI: doi.org/10.1002/aic.12138..

[79] Ke Wang and Alexander W Dowling. "Bayesian optimization for chemical products and functional materials". In: *Current Opinion in Chemical Engineering* 36 (2022), p. 100728.

[80] Leonardo D González and Victor M Zavala. "New paradigms for exploiting parallel experiments in Bayesian optimization". In: *Computers & Chemical Engineering* 170 (2023), p. 108110. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2022.108110. URL: https://www.sciencedirect.com/science/article/pii/S0098135422004434.

[81] Zachary Cosenza et al. "Multi-information source Bayesian optimization of culture media for cellular agriculture". In: *Biotechnology and Bioengineering* 119.9 (2022), pp. 2447–2458. DOI: 10.1002/bit.28132. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.28132. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.28132.

[82] Jose Pablo Folch et al. "Combining multi-fidelity modelling and asynchronous batch Bayesian Optimization". In: *Computers & Chemical Engineering* 172 (2023), p. 108194. ISSN: 0098-1354. DOI: https://doi.org/10.1016/j.compchemeng.2023.108194. URL: https://www.sciencedirect.com/science/article/pii/S0098135423000637.

[83] Joel A. Paulson and Congwen Lu. "COBALT: COnstrained Bayesian optimizAtion of computationaLly expensive grey-box models exploiting derivaTive information". In: *Computers & Chemical Engineering* 160 (2022), p. 107700. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2022.107700. URL: https://www.sciencedirect.com/science/article/pii/S0098135422000436.

[84] Congwen Lu and Joel A. Paulson. *No-Regret Constrained Bayesian Optimization of Noisy and Expensive Hybrid Models using Differentiable Quantile Function Approximations*. 2023. arXiv: 2305.03824 [stat.ML].

[85] Jialu Wang and Alexander W Dowling. "Pyomo.DoE: An open-source package for model-based design of experiments in Python". In: *AIChE Journal* 68.12 (2022), e17813.

[86] Dave Higdon et al. "Computer Model Calibration Using High-Dimensional Output". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 570–583. ISSN: 01621459. URL: http://www.jstor.org/stable/27640080 (visited on 09/03/2023).

[87] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[88] Eric Bradford, Artur M Schweidtmann, and Alexei Lapkin. "Efficient multiobjective optimization employing Gaussian processes, spectral sampling and a genetic algorithm". eng. In: *Journal of Global Optimization* 71.2 (2018), pp. 407–438. DOI: doi.org/10.1007/s10898-018-0609-2.

[89] Mehdi Jamei et al. "On the assessment of specific heat capacity of nanofluids for solar energy applications: Application of Gaussian process regression (GPR) approach". eng. In: *Journal of Energy Storage* 33 (2021), p. 102067. DOI: 10.1016/j.est.2020.102067.

[90] Brian Kolb et al. "Representing Global Reactive Potential Energy Surfaces Using Gaussian Processes". eng. In: *The Journal of Physical Chemistry. A, Molecules, Spectroscopy, Kinetics, Environment, & General Theory* 121.13 (2017), pp. 2552–2557. DOI: 10.1021/acs.jpca.7b01182.

[91] V.S. Mandrekar and L. Gawarecki. *Stochastic Analysis for Gaussian Random Processes and Fields: With Applications*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2015. ISBN: 9781498707824. URL: https://books.google.com/books?id=dez5CQAAQBAJ.

[92] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[93] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[94] O. Kahrs and W. Marquardt. "Incremental identification of hybrid process models". In: *Computers & Chemical Engineering* 32.4 (2008), pp. 694–705. DOI: 10.1016/j.compchemeng.2007.02.014.

[95] Raymond K. W. Wong, Curtis B. Storlie, and Thomas C. M. Lee. "A frequentist approach to computer model calibration". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 79.2 (2017), pp. 635–648. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/44682529 (visited on 07/28/2023).

[96] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (2016), e55.

[97] Matthew D. Hoffman and Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.47 (2014), pp. 1593–1623. URL: http://jmlr.org/papers/v15/hoffman14a.html.

[98] Alexander I. J Forrester, András Sóbester, and A.J Keane. *Engineering Design via Surrogate Modelling: A Practical Guide*. eng. Progress in Astronautics and Aeronautics. Chichester, West Sussex, England ; Hoboken, NJ: J. Wiley, 2008.

[99] Ryan G. McClarren. *Uncertainty Quantification and Predictive Computational Science: A Foundation for Physical Scientists and Engineers*. 1st ed. 2018. Springer International Publishing, 2018.

[100] Lorenz T Biegler. *Nonlinear Programming. Concepts, Algorithms, and Applications to Chemical Processes*. Society for Industrial and Applied Mathematics, 2010.

[101] Elvis A Eugene, Xian Gao, and Alexander W Dowling. "Learning and Optimization with Bayesian Hybrid Models". eng. In: *2020 American Control Conference (ACC)*. AACC, 2020, pp. 3997–4002. DOI: doi.org/10.23919/ACC45564.2020.9148007. URL: https://doi.org/10.23919/ACC45564.2020.9148007.

[102] Jialu Wang, Elvis A Eugene, and Alexander W Dowling. "Scalable Stochastic Programming with Bayesian Hybrid Models". In: *14th International Symposium on Process Systems Engineering*. Vol. 49. Computer Aided Chemical Engineering. 2022, pp. 1309–1314.

[103] Indranil Pan, Lachlan R. Mason, and Omar K. Matar. "Data-centric Engineering: integrating simulation, machine learning and statistics. Challenges and opportunities". In: *Chemical Engineering Science* 249 (2022), p. 117271. DOI: doi.org/10.1016/j.ces.2021.117271.

[104] Bryce Meredig. "Five High-Impact Research Areas in Machine Learning for Materials Science". In: *Chemistry of Materials* 31.23 (2019), pp. 9579–9581. DOI: doi.org/10.1021/acs.chemmater.9b04078.

[105] Kristian Kersting. "Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines". In: *Frontiers in Big Data* 1 (2018), p. 6. DOI: doi.org/10.3389/fdata.2018.00006.

[106] Miguel López-Pérez et al. "Learning from crowds in digital pathology using scalable variational Gaussian processes". In: *Scientific Reports* 11.1 (2021), p. 11612. DOI: doi.org/10.1038/s41598-021-90821-3..

[107] Jenn Brynjarsdóttir and Anthony O'Hagan. "Learning about physical parameters: The importance of model discrepancy". In: *Inverse problems* 30.11 (2014), p. 114007.

[108] Joshua Landon and Nozer D. Singpurwalla. "Choosing a Coverage Probability for Prediction Intervals". In: *The American Statistician* 62.2 (2008), pp. 120–124. ISSN: 00031305. URL: http://www.jstor.org/stable/27643988 (visited on 11/22/2022).

[109] Artur M Schweidtmann et al. "Deterministic global optimization with Gaussian processes embedded". In: *Mathematical Programming Computation* 13.3 (2021), pp. 553–581.

[110] Ke Wang et al. "When physics-informed data analytics outperforms black-box machine learning: A case study in thickness control for additive manufacturing". In: *Digital Chemical Engineering* (2022), p. 100076.

Supplementary Material for *Learning and Optimization Under Epistemic Uncertainty with Bayesian Hybrid Models*

Elvis A. Eugene, Kyla D. Jones, Xian Gao, Jialu Wang, & Alexander W. Dowling (2023), *Computers & Chemical Engineering*, https://doi.org/10.1016/j.compchemeng.2023.108430

Ballistics Case Study

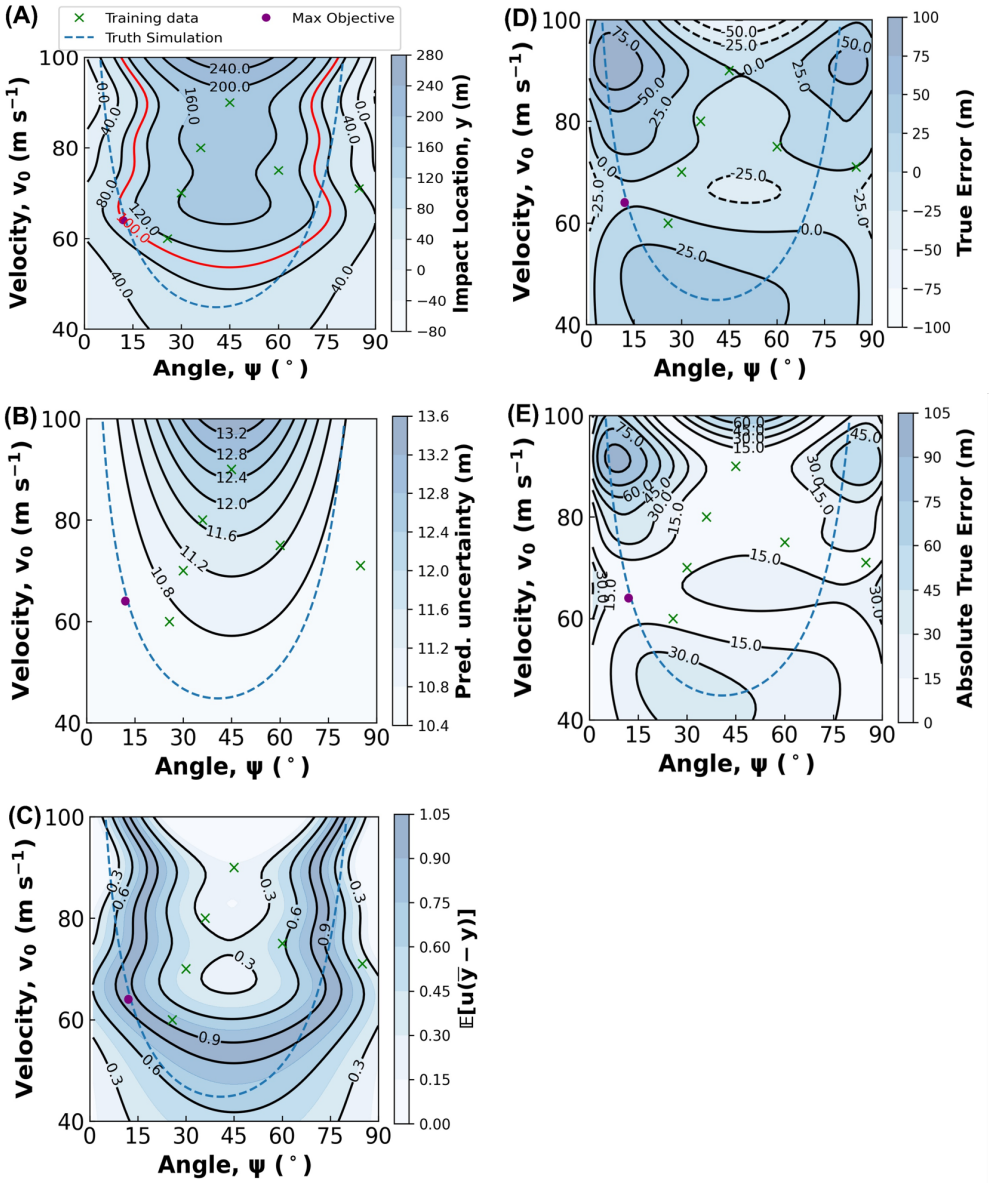


Figure S1: Results from the Bayesian decision-making framework using the incremental hybrid model and M2 (composite estimation neglecting GP uncertainty). The control variables of the ballistics experiment, firing velocity v_0 and angle ψ, are plotted on the vertical and horizontal axes, respectively. The dashed blue contour shows true experimental conditions needed to satisfy the goal of the experiment, corresponding to the 100 m contour in Fig. 2. Green $\times$'s are the training data and the purple dot is the optimum recommendation of the framework. **(A)** incremental model predictions. **(B)** Uncertainty in incremental model predictions. **(C)** Objective values calculated via M2 (composite estimation neglecting GP uncertainty). **(D)** True error. **(E)** Absolute true error.

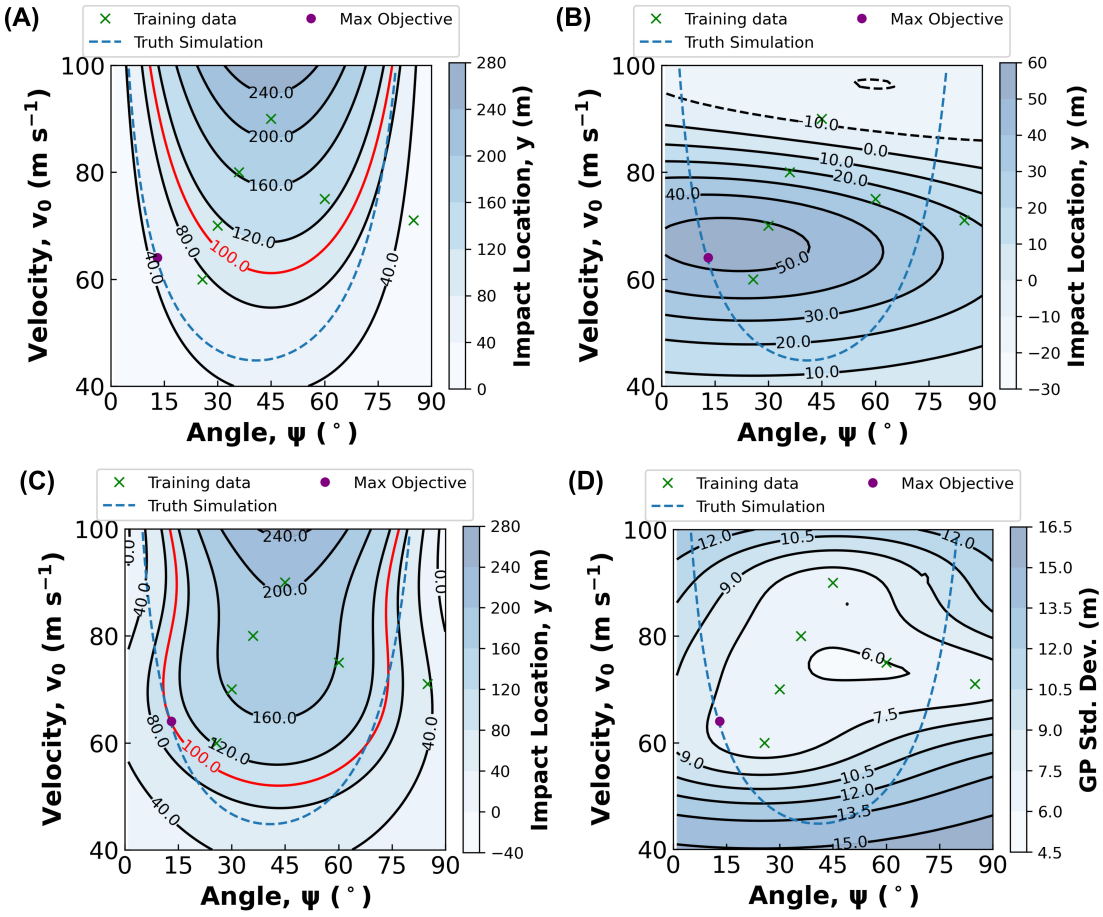


Figure S2: The additive hybrid model is a linear combination of the assumed physical glass-box model and GP discrepancy function. **(A)** Predictions of the glass-box model. Notice that the contour shapes are identical to the simple physics model predictions in Fig. 3(A). **(B)** Predictions of the GP discrepancy function. Notice the qualitative agreement with the contours for the GP-only model prediction in Fig. 3(F). **(C)** Values of the hybrid model predictions are a linear combination of the values in (A) and (B). Note that this plot is identical to Fig. 3(K).

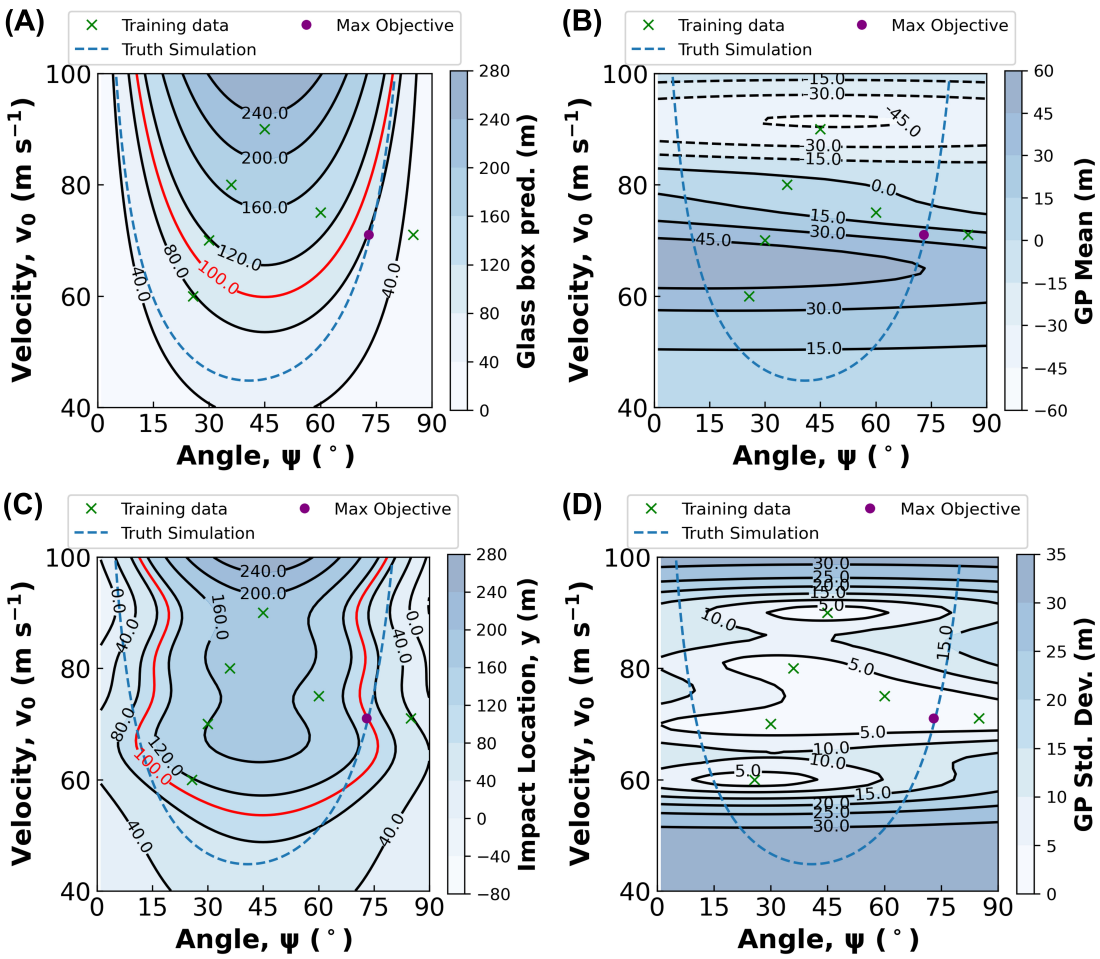


Figure S3: Component predictions of the incremental hybrid model. **(A)** Predictions of the glass-box model. Notice that the contour shapes are identical to the simple physics model predictions in Fig. 3(A). **(B)** Predictions of the GP discrepancy function. Notice the horizontal contour lines and orientation of the ellipse (unlike Fig. 3(F)) indicating high uncertainty in learning the effect of firing angle ψ on the model discrepancy. **(C)** Values of the hybrid model predictions are a linear combination of the values in (A) and (B). Note that this plot is identical to Fig. 6(C). **(D)** The GP standard deviation has irregularly shaped contours surrounding the training data but is nearly horizontal elsewhere which is another indication of the GP's inability to learn the influence of angle.

Reactions Case Study

Case	Experiment	c_{A0} (M)	T-0.65 (K)
1	10	1.1	369
	13	4.3	432
	18	3.1	306
	1	2.3	324
2	3	1.5	387
	2	3.5	468
	12	2.7	441
	18	3.1	306
3	7	1.9	396
	13	4.2	432
	6	4.5	315
	11	1.7	342
4	5	4.7	297
	15	3.7	423
	16	1.3	450
	1	2.3	324
5	17	3.9	360
	12	2.7	441
	11	1.7	342
	16	1.3	450
6	20	2.1	333
	15	3.7	423
	13	4.2	432
	5	4.7	297
7	2	3.5	468
	6	4.5	315
	20	2.1	333
	16	1.3	450
8	8	2.9	414
	17	3.9	360
	20	2.1	333
	18	3.1	306
9	9	2.5	405
	5	4.7	297
	16	1.3	450
	1	2.3	324
10	13	4.2	432
	14	4.1	378
	11	1.7	342
	9	2.5	405

Table S1: Training data for case studies 1-10 generated with a random sampling of experiments 1-20 in the Latin hypercube design.

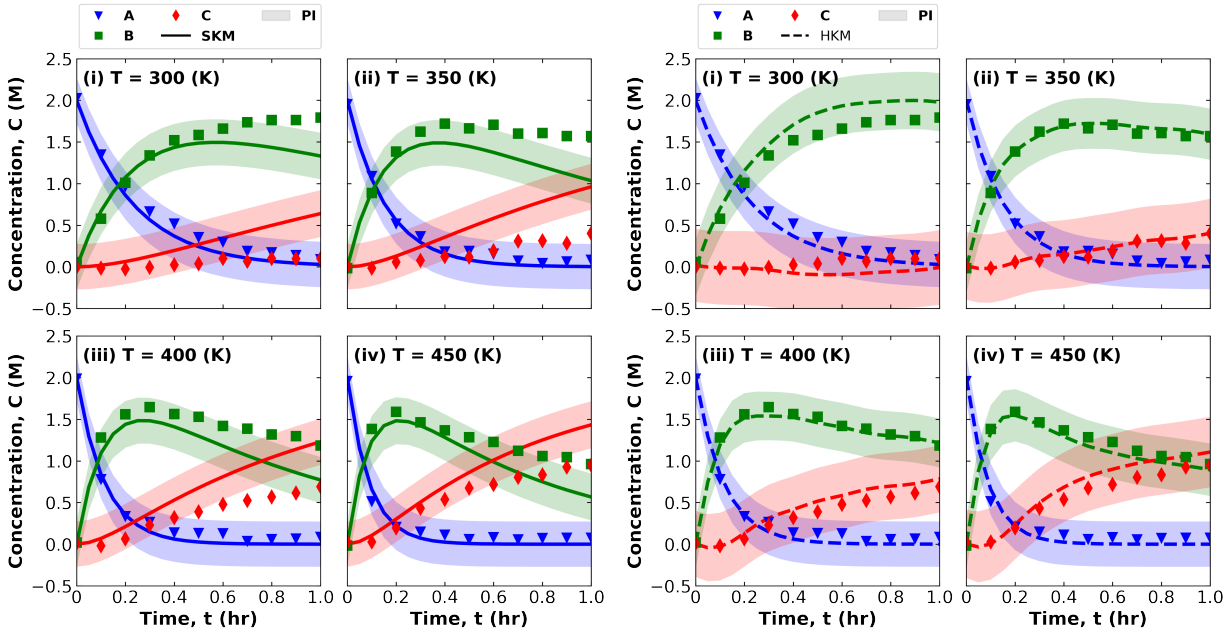
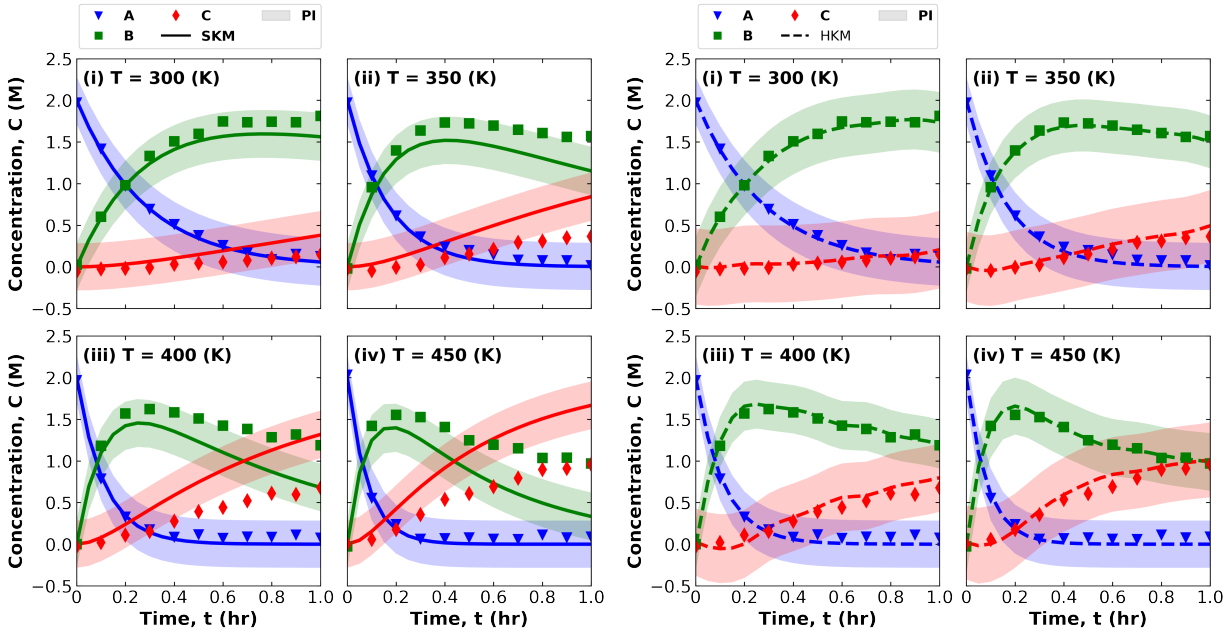
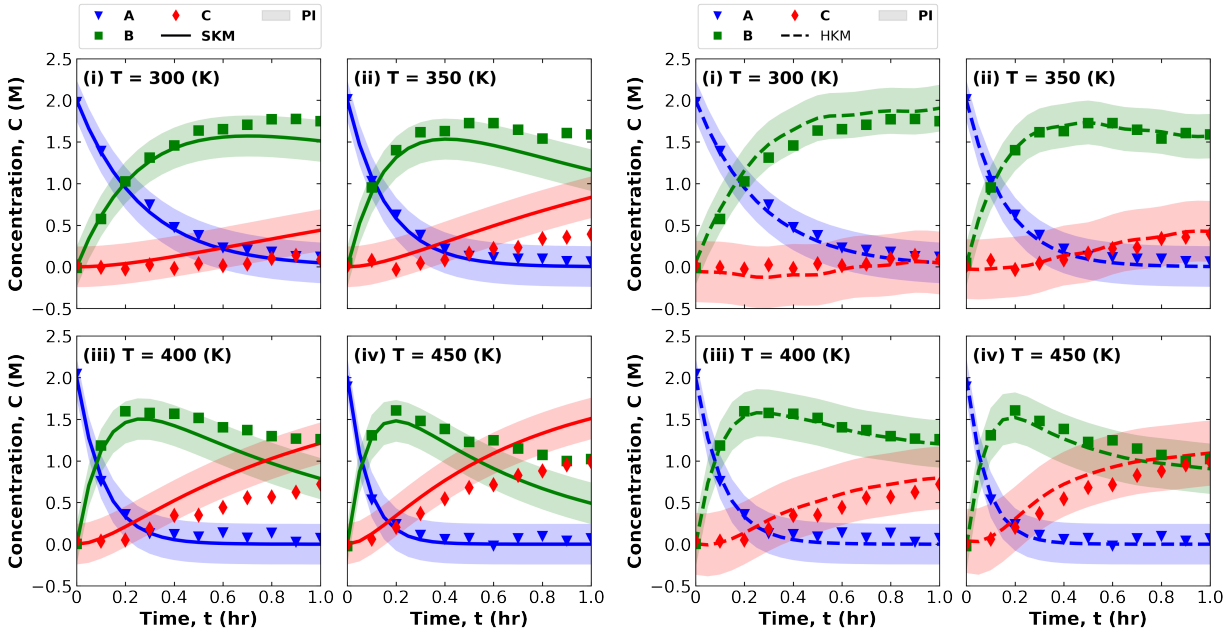


Figure S4: Case **two** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).



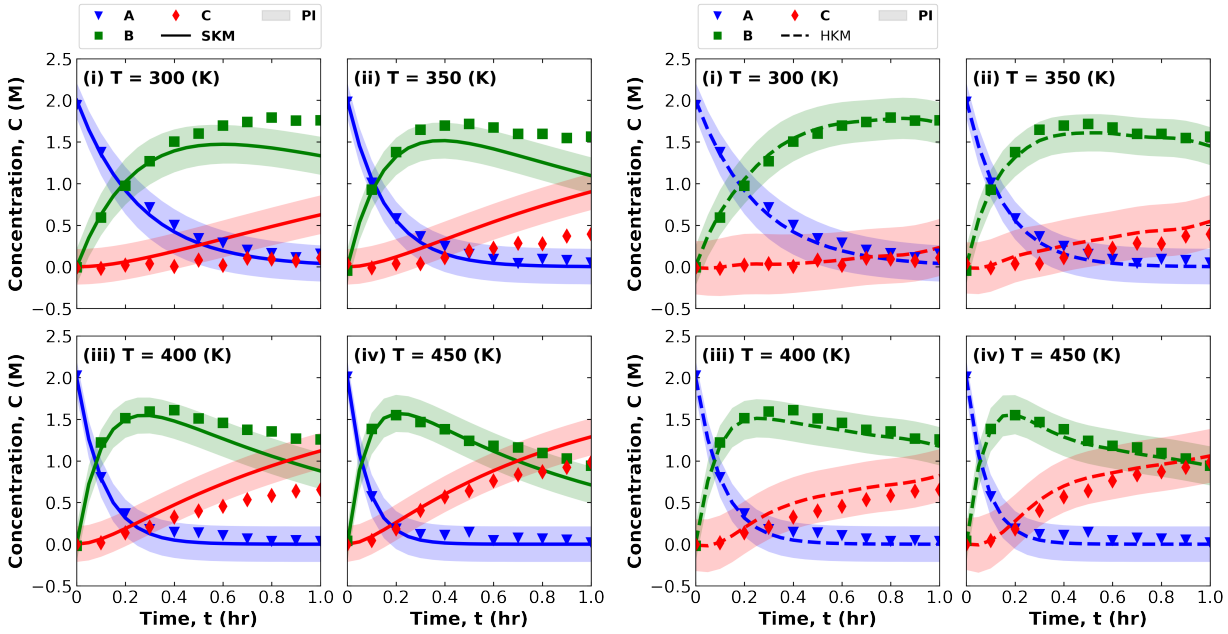
(a) (b)

Figure S5: Case **three** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).



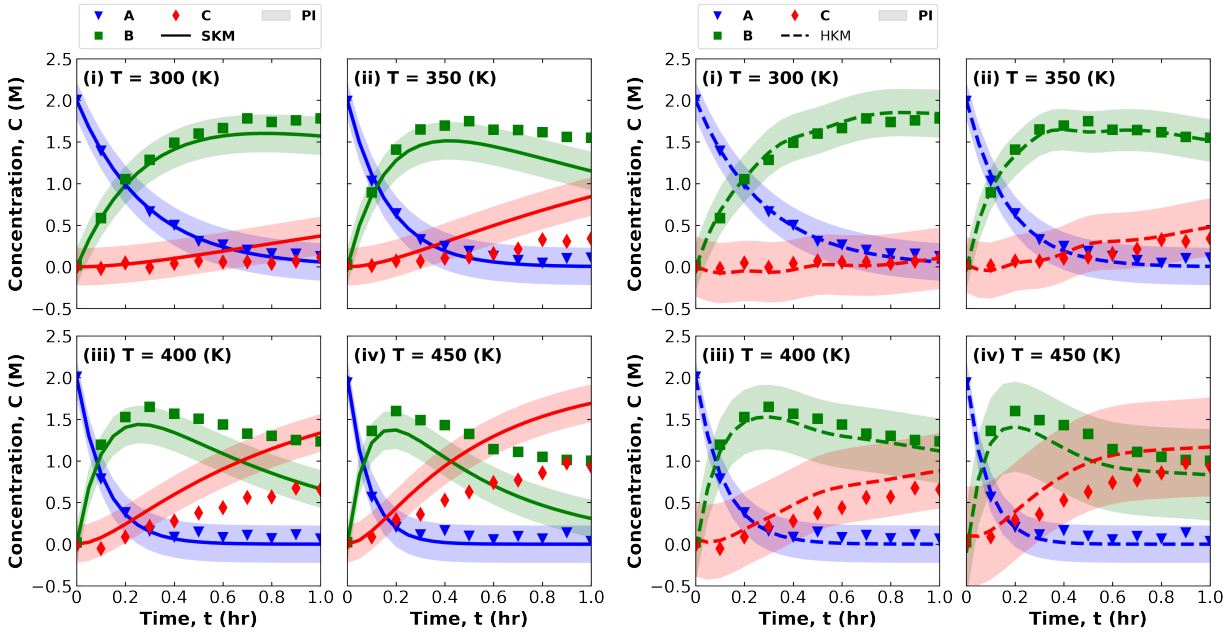
(a) (b)

Figure S6: Case **four** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).



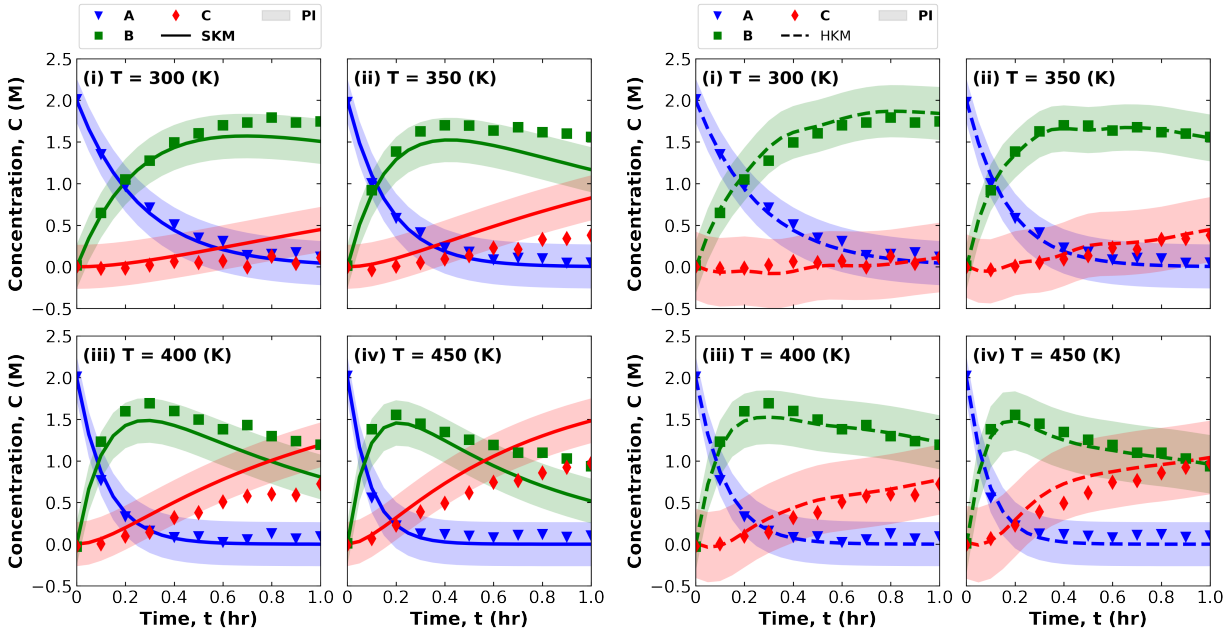
(a) (b)

Figure S7: Case **five** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).



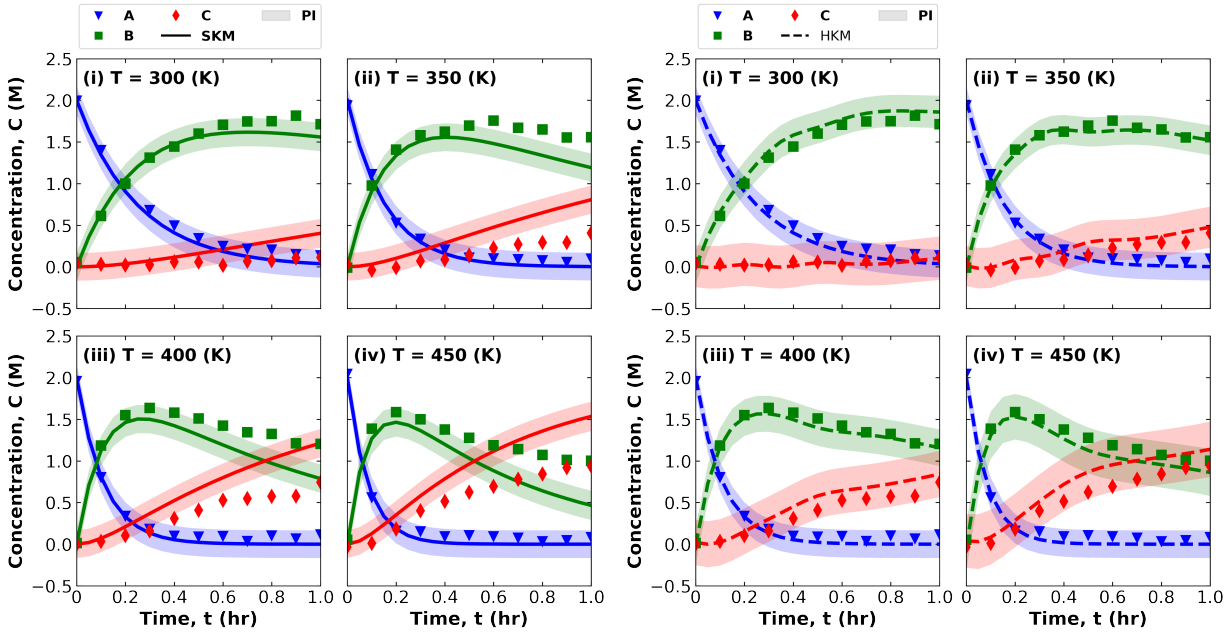
(a) (b)

Figure S8: Case **six** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).



(a) (b)

Figure S9: Case **seven** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).



(a) (b)

Figure S10: Case **eight** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).

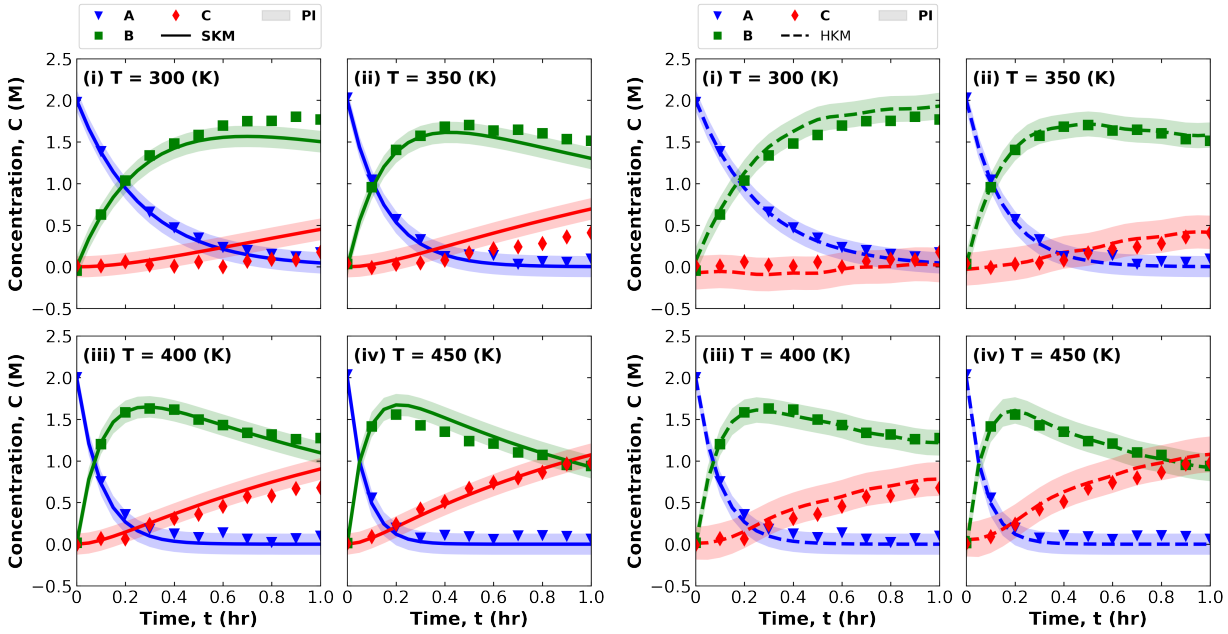
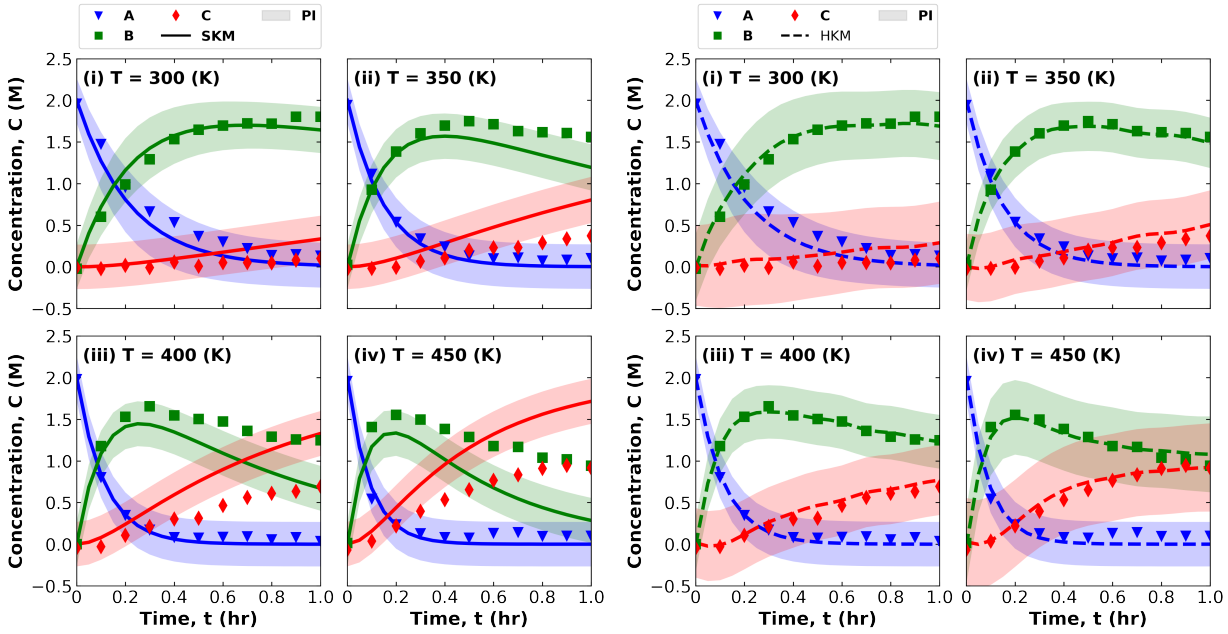


Figure S11: Case **nine** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).



(a) (b)

Figure S12: Case **ten** concentration of all chemical species vs. time for (a) SKM (solid lines) and (b) HKM (dashed lines) holding the initial concentration of A at 2.0 M for temperatures (i) 300, (ii) 350, (iii) 400, & (iv) 450 K. Observed concentration of A, B, and C are shown with red triangles, green squares, and blue diamonds, respectively. Shaded regions represent point-wise prediction intervals ($\pm\hat{\sigma}$).

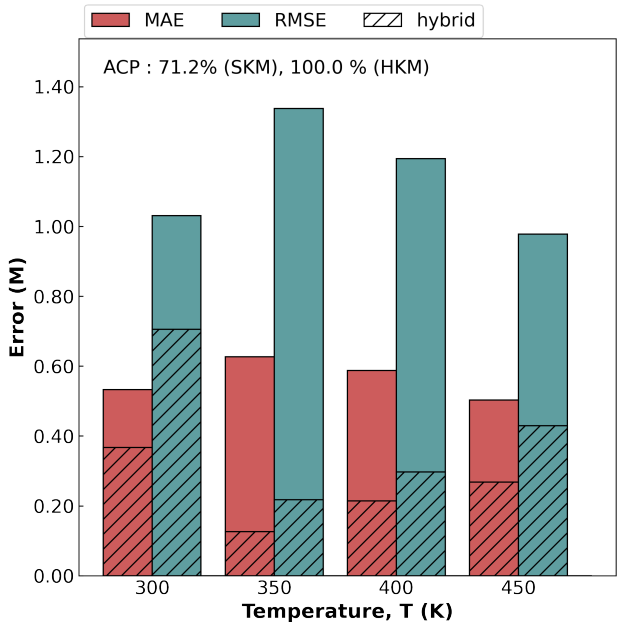


Figure S13: Case **two** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

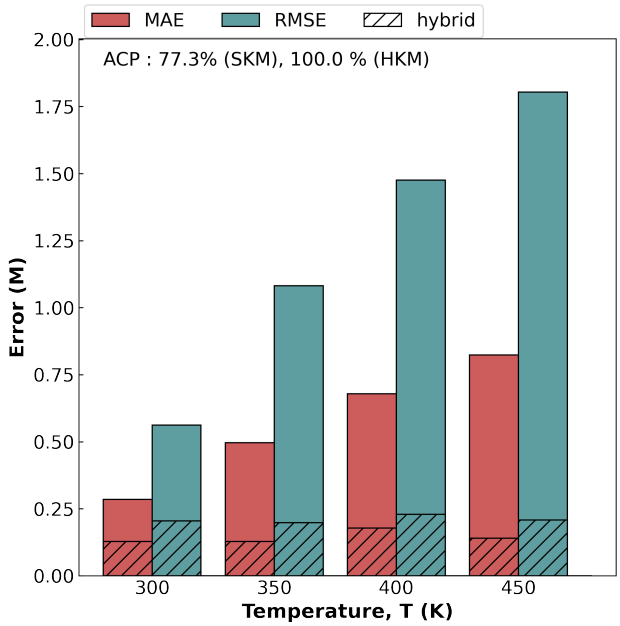


Figure S14: Case **three** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

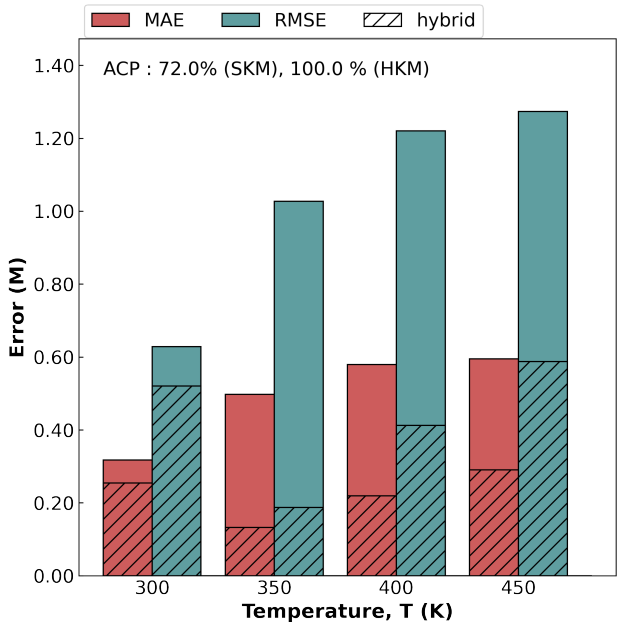


Figure S15: Case **four** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

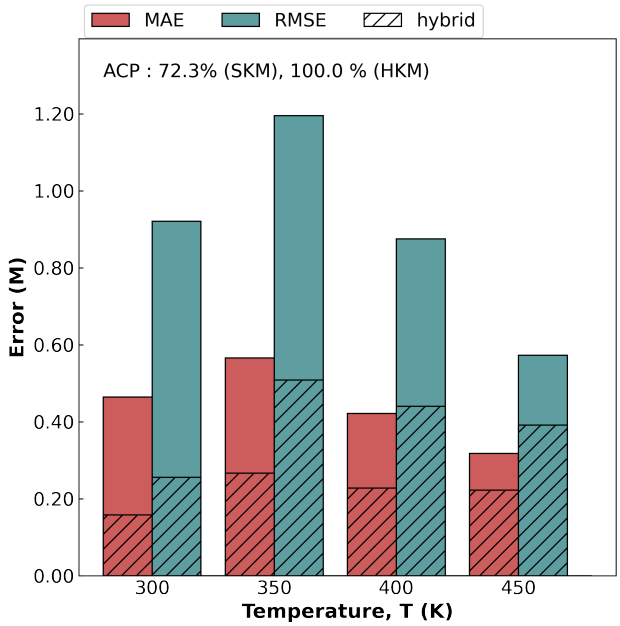


Figure S16: Case **five** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

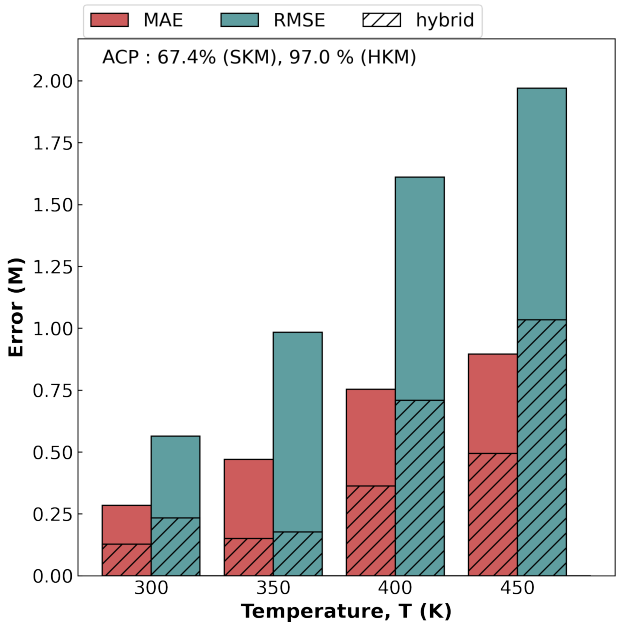


Figure S17: Case **six** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

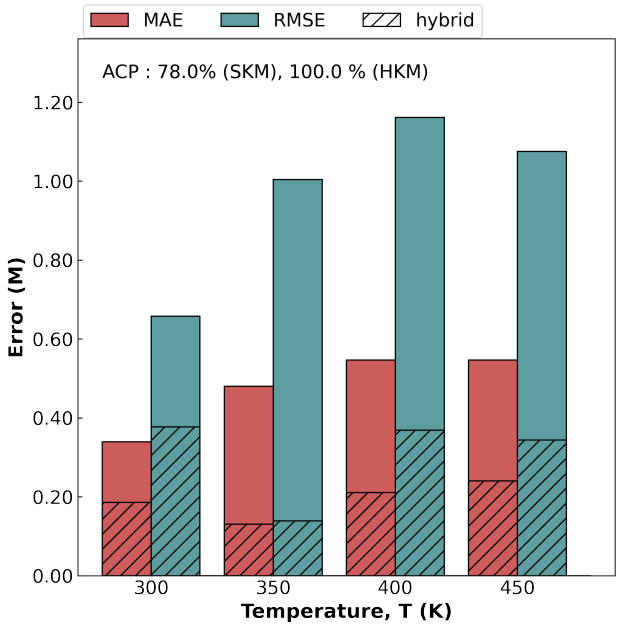


Figure S18: Case **seven** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

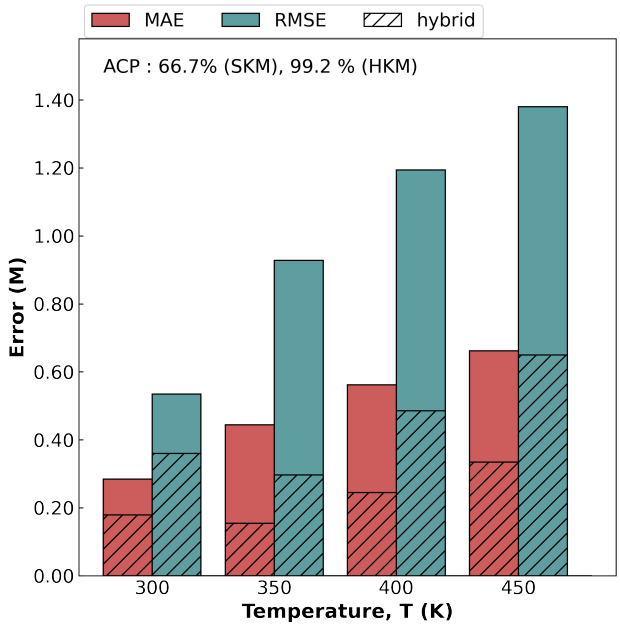


Figure S19: Case **eight** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

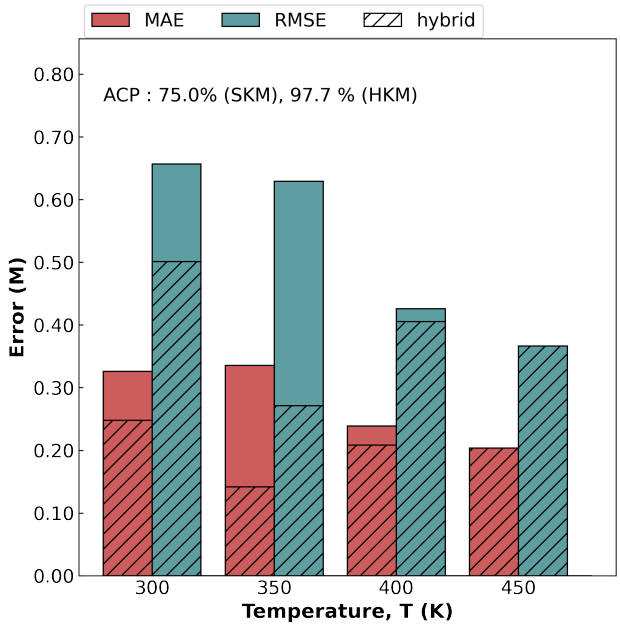


Figure S20: Case **nine** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

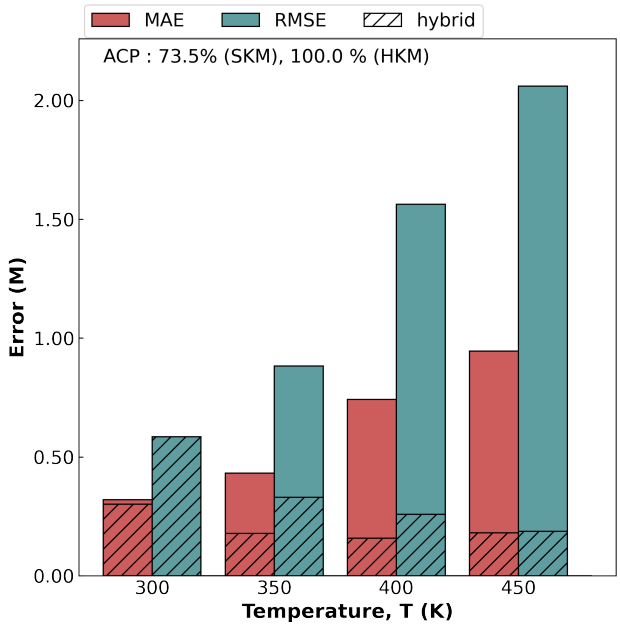


Figure S21: Case **ten** mean absolute error (MAE, pink) and root mean squared error (RMSE, blue) for the SKM (solid) and HKM (stripes) across temperatures 300, 350, 400, & 450 K. Actual coverage probability (ACP) across all temperatures is reported for both models.

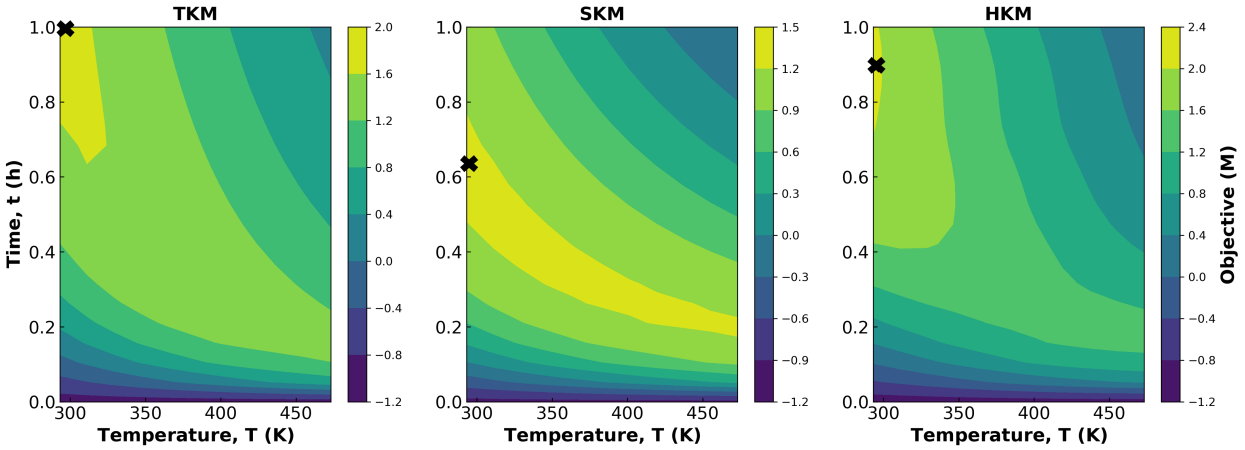


Figure S22: Objective of the TKM, SKM, and HKM as function of time and temperature for case **two**. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

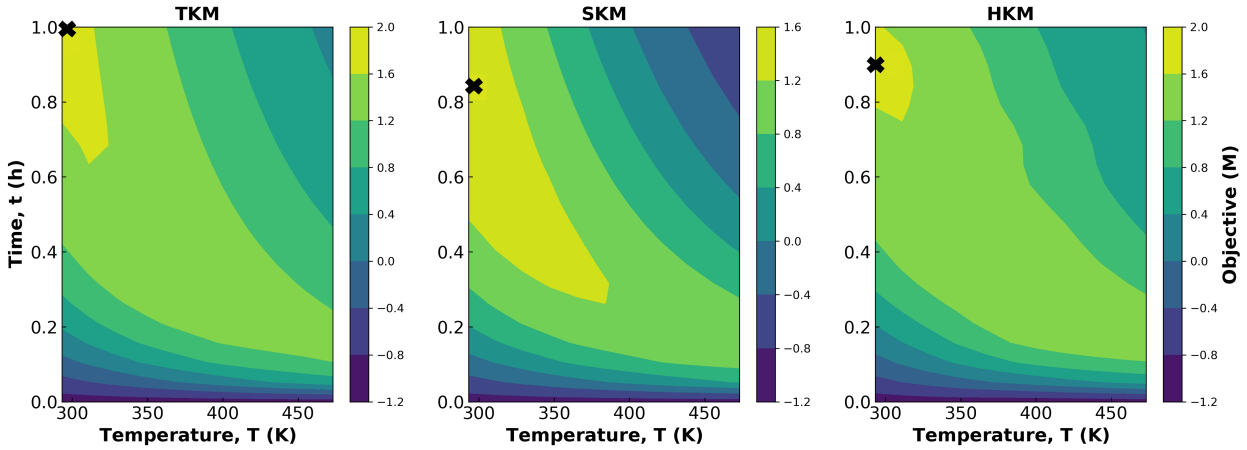


Figure S23: Objective of the TKM, SKM, and HKM as function of time and temperature for case **three**. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

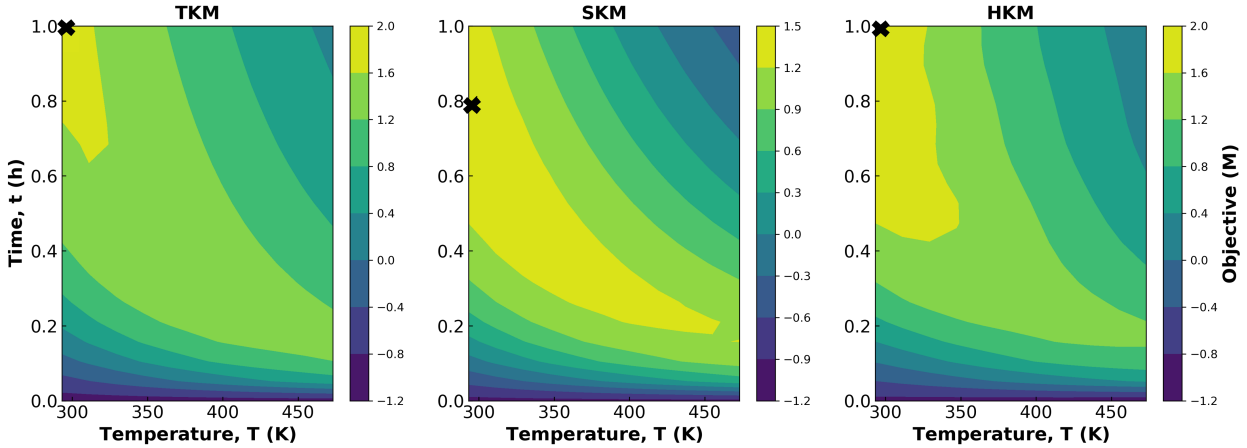


Figure S24: Objective of the TKM, SKM, and HKM as function of time and temperature for case **four**. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

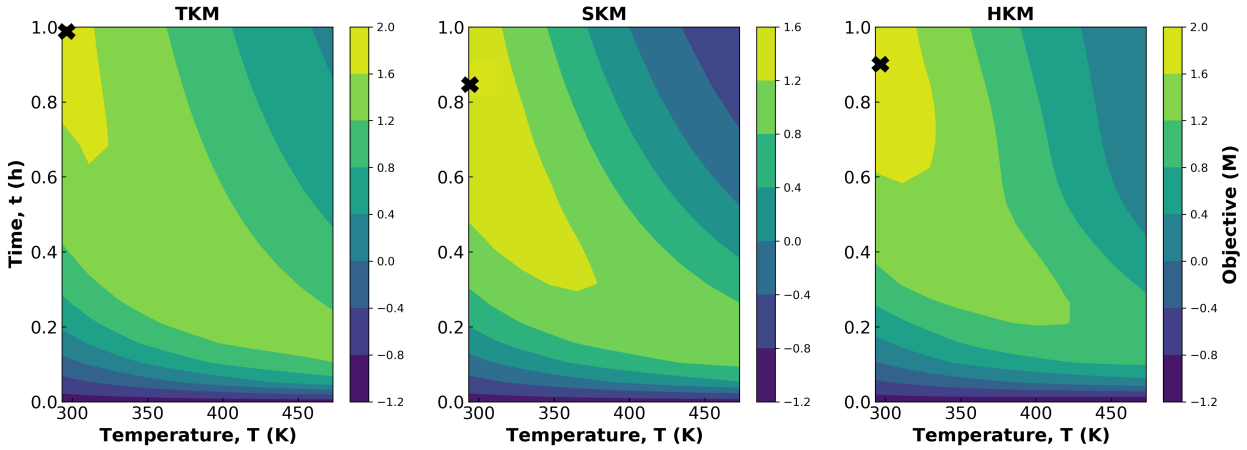


Figure S25: Objective of the TKM, SKM, and HKM as function of time and temperature for case **six**. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

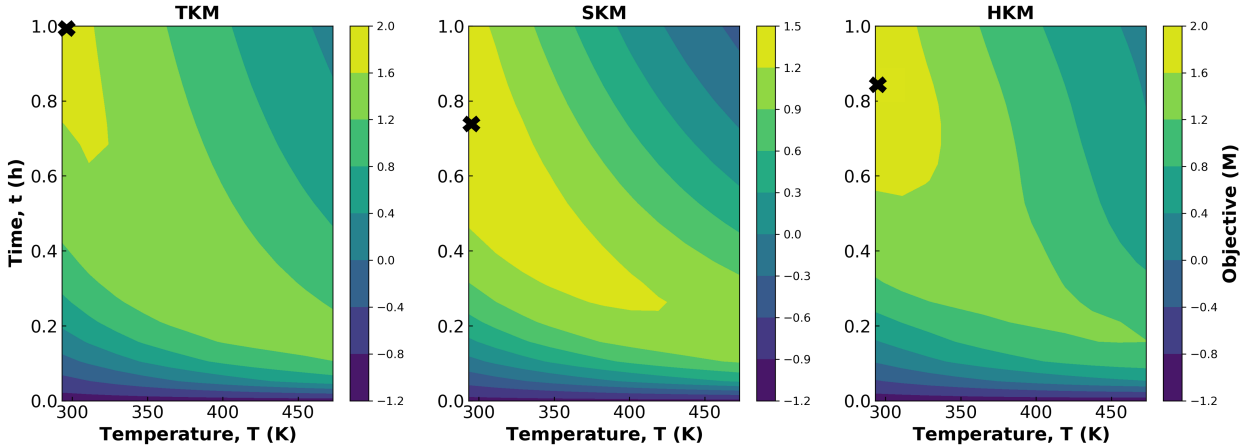


Figure S26: Objective of the TKM, SKM, and HKM as function of time and temperature for case **seven**. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

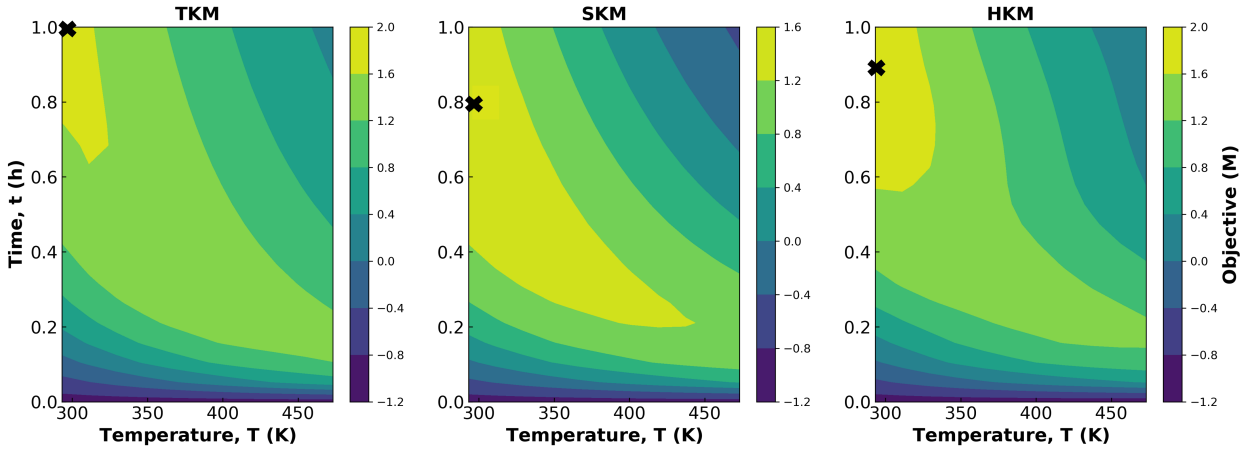


Figure S27: Objective of the TKM, SKM, and HKM as function of time and temperature for case **eight**. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

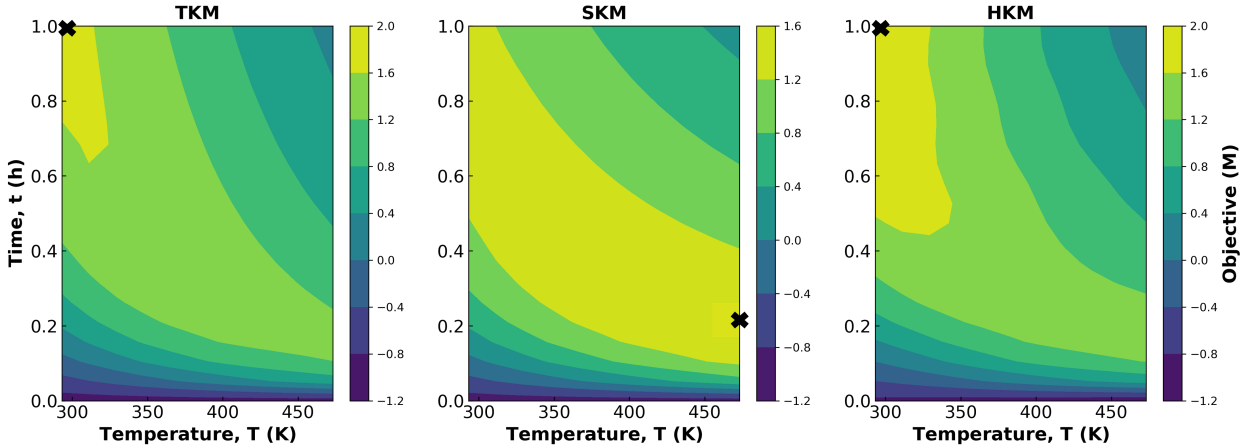


Figure S28: Objective of the TKM, SKM, and HKM as function of time and temperature for case **nine**. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

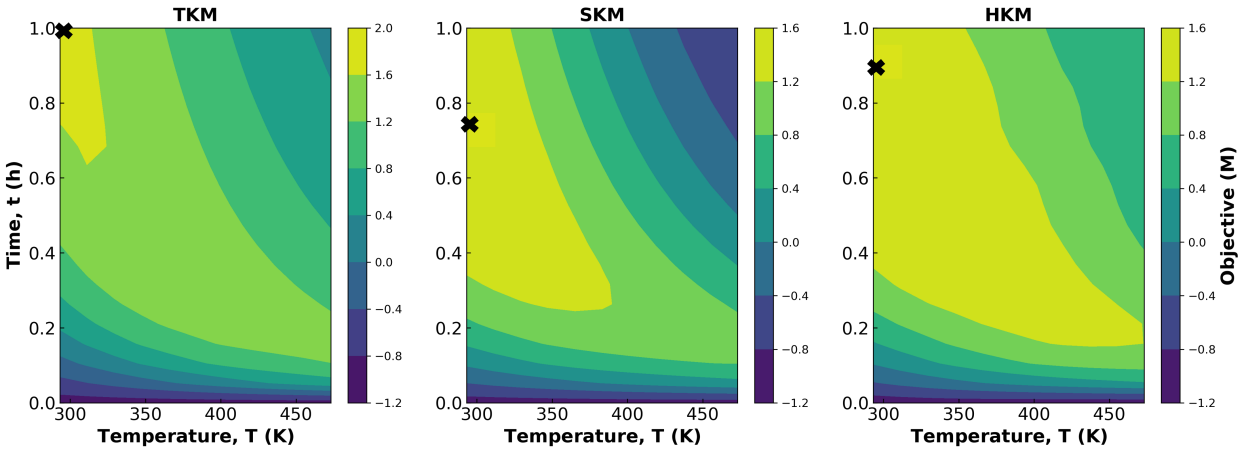


Figure S29: Objective of the TKM, SKM, and HKM as function of time and temperature for case **ten**. The objective function was computed with c_{A0} = 2.0 M. The maximum value of the objective function is marked with a black X.

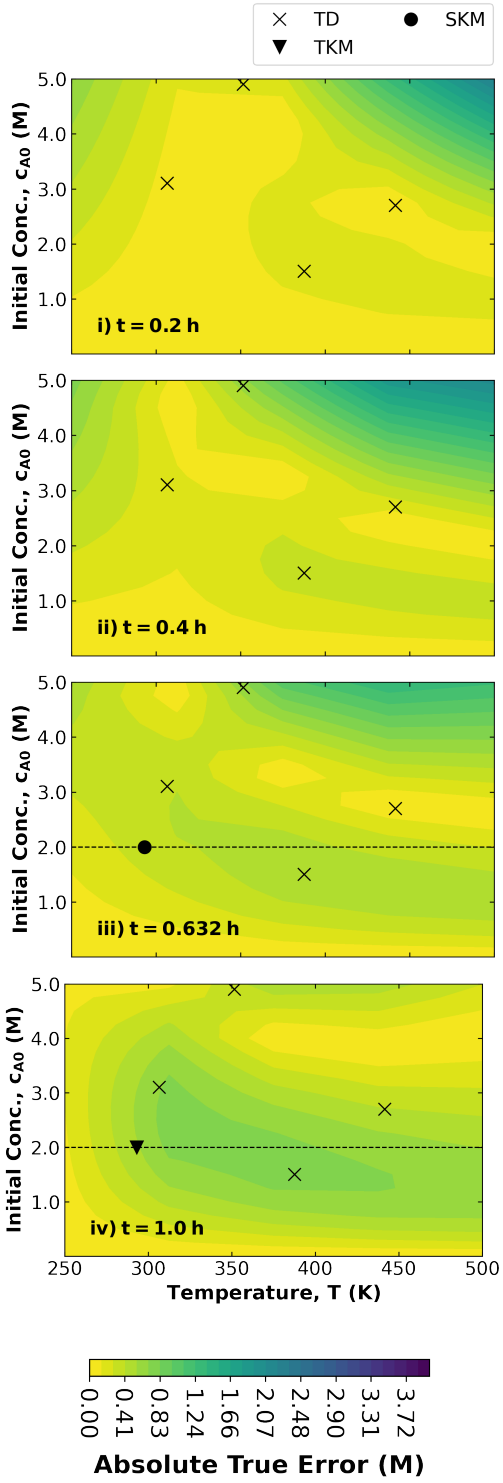


Figure S30: Absolute true error of the objective as a function of starting concentration of A, and temperature for case **two** at four snapshots in time (top to bottom). For comparison with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.

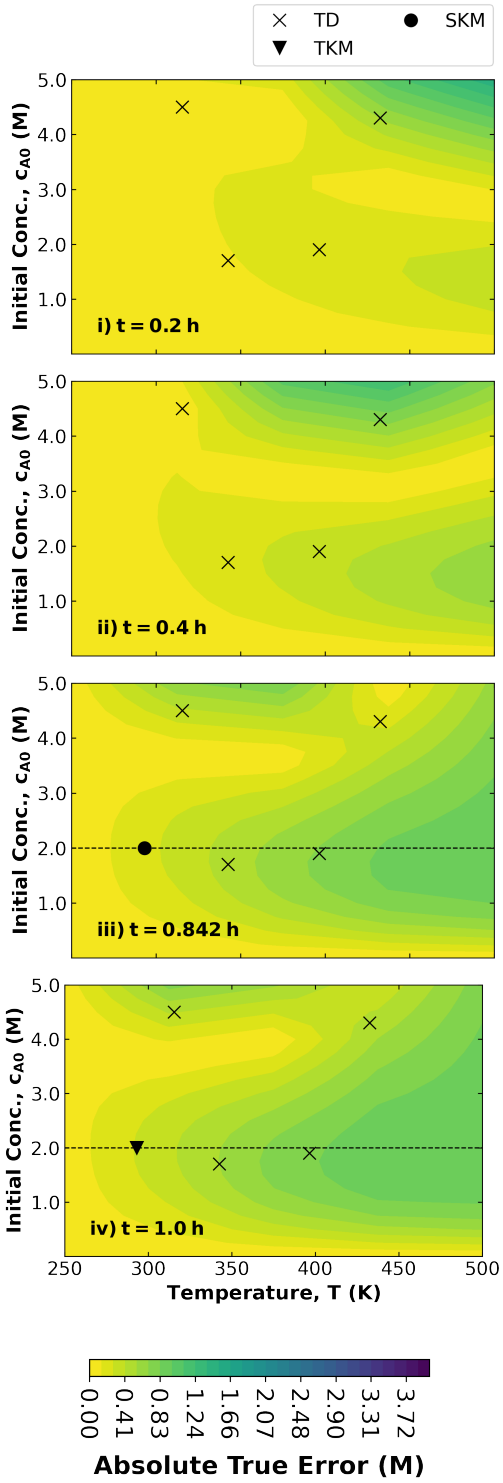


Figure S31: Absolute true error of the objective as a function of starting concentration of A, and temperature for case **three** at four snapshots in time (top to bottom). For comparison with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.

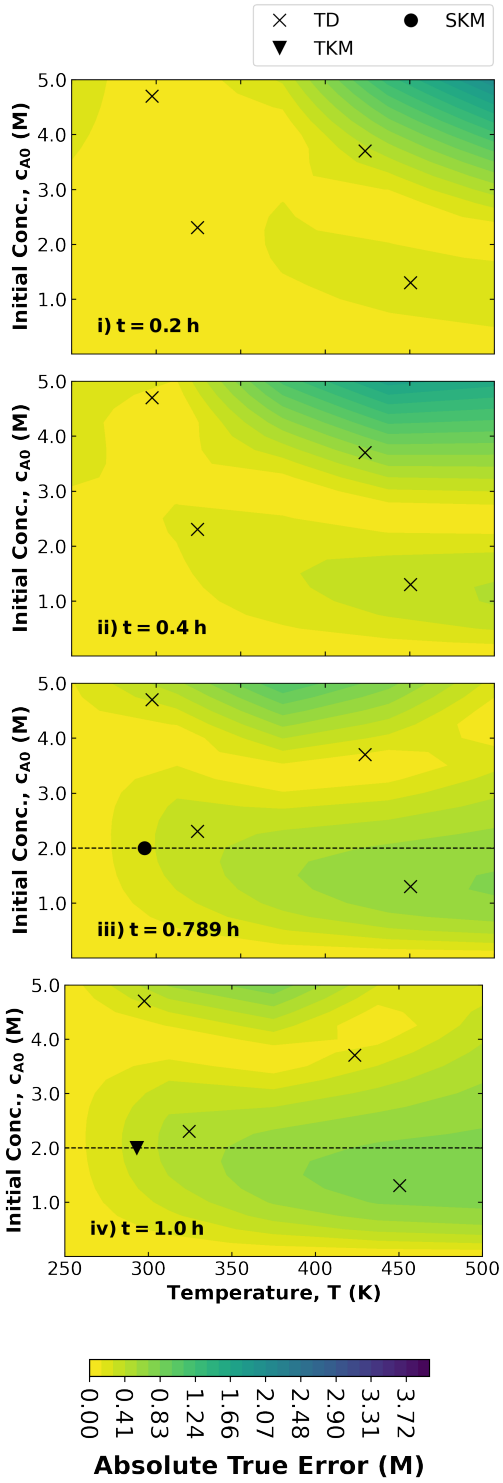


Figure S32: Absolute true error of the objective as a function of starting concentration of A, and temperature for case **four** at four snapshots in time (top to bottom). For comparison with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.

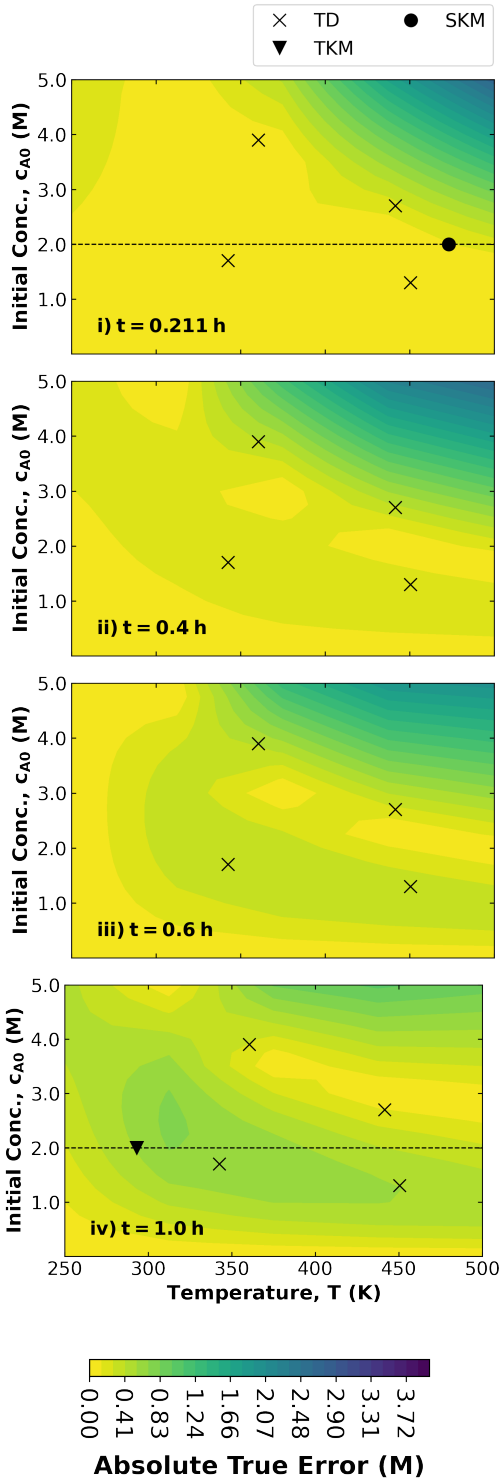


Figure S33: Absolute true error of the objective as a function of starting concentration of A, and temperature for case **five** at four snapshots in time (top to bottom). For comparison with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.

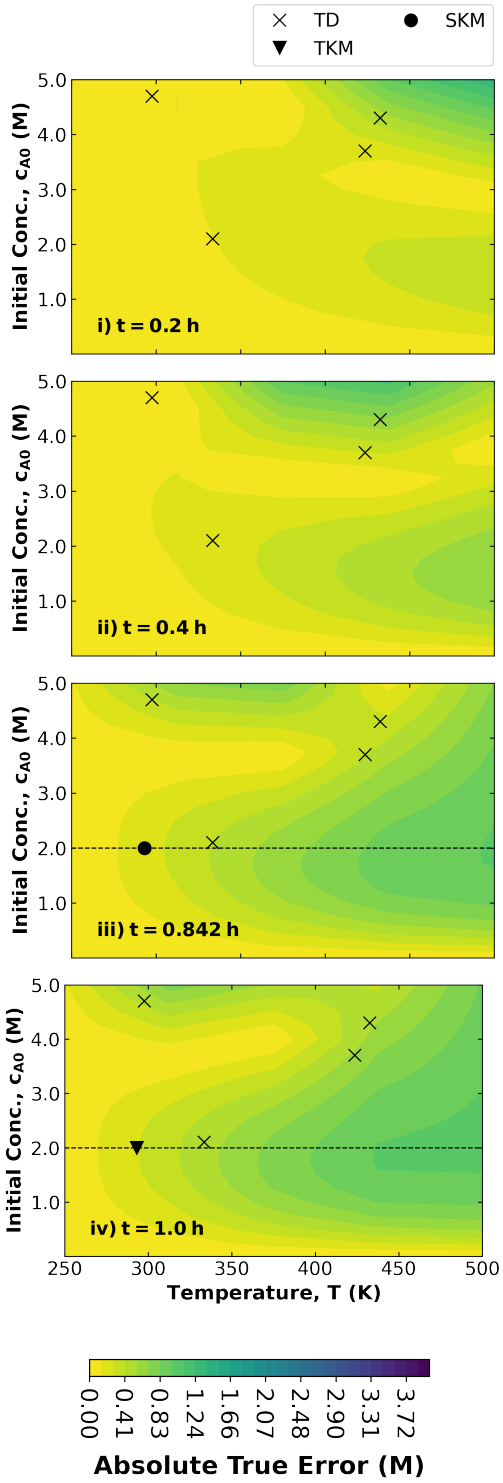


Figure S34: Absolute true error of the objective as a function of starting concentration of A, and temperature for case **six** at four snapshots in time (top to bottom). For comparison with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.

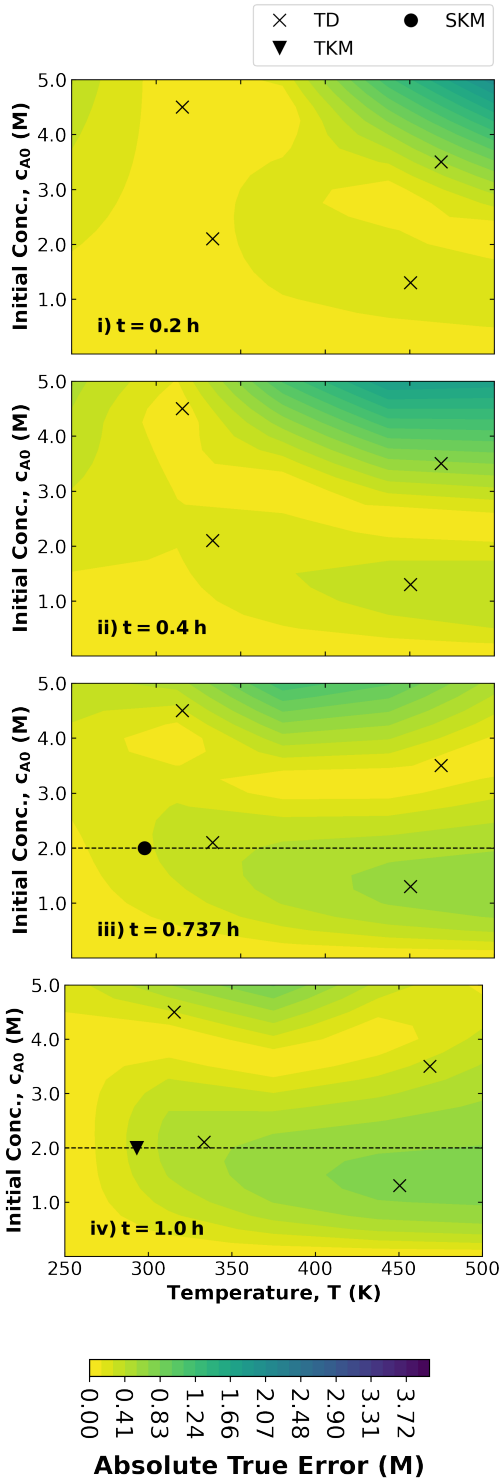


Figure S35: Absolute true error of the objective as a function of starting concentration of A, and temperature for case **seven** at four snapshots in time (top to bottom). For comparison with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.

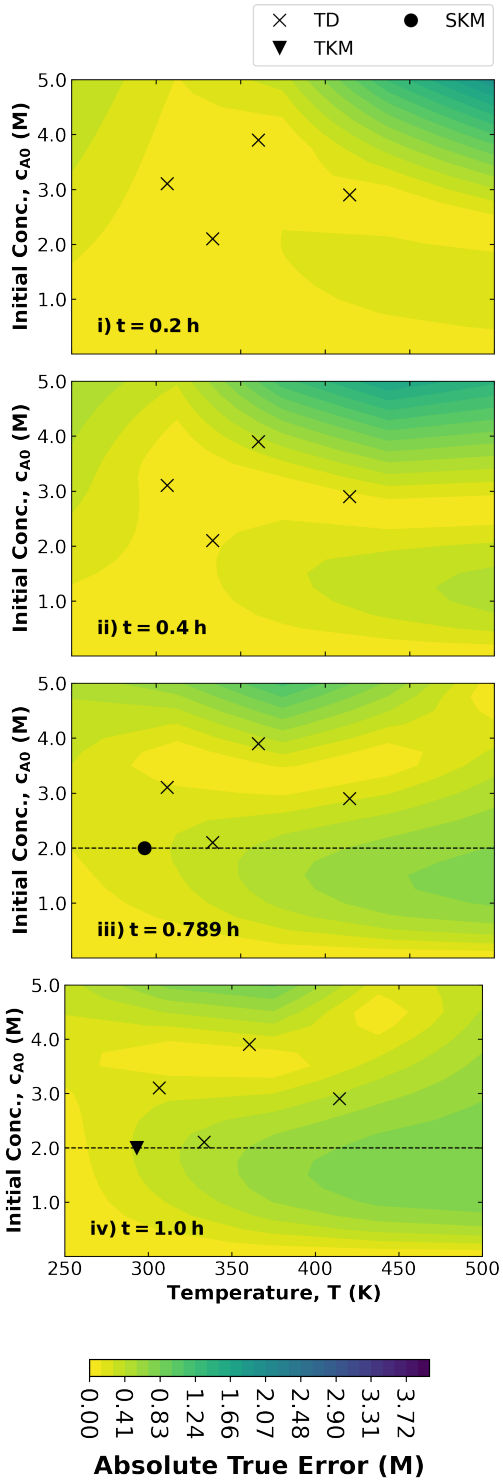


Figure S36: Absolute true error of the objective as a function of starting concentration of A, and temperature for case **eight** at four snapshots in time (top to bottom). For comparison with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.

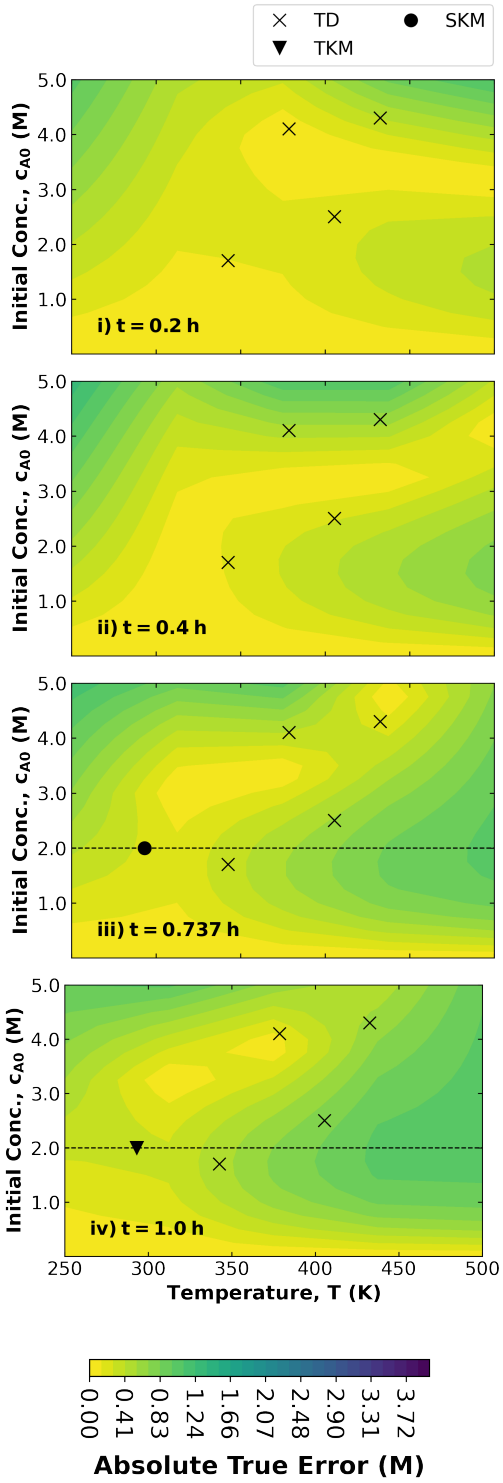


Figure S37: Absolute true error of the objective as a function of starting concentration of A, and temperature for case **ten** at four snapshots in time (top to bottom). For comparison with the previous analysis, the predicted maximum objective temperature and time holding the initial concentration at 2.0 M for the SKM and TKM are represented by a circle and triangle, respectively.