Investigating Model Form Error Estimation for Sparse Data

Kyle D. Neal[1], Mohammad Khalil[2], and Teresa Portone[3]

[1,3] *Sandia National Laboratories, Albuquerque, NM, 87123, USA*
kneal@sandia.gov
tporton@academic.edu

[2] *Sandia National Laboratories, Livermore, CA, 94550, USA*
mkhalil@sandia.gov

Abstract

Computational simulations of dynamical systems often involve the use of mathematical models and algorithms to mimic and analyze complex real-world phenomena. By leveraging computational power, simulations enable researchers to explore and understand systems that are otherwise challenging to study experimentally. They offer a cost-effective and efficient means to predict and analyze the behavior of engineering, biological, and social systems. However, model form error arises in computational simulations from simplifications, assumptions, and limitations inherent in the mathematical model formulation. Several methods for addressing model form error have been proposed in the literature, but their robustness in the face of challenges inherent to real-world systems has not been thoroughly investigated. In this work, a data assimilation-based approach for model form error estimation is investigated in the presence of sparse observation data. An extension for including physics-based domain knowledge to improve estimation performance is proposed. A computational simulation based on the Lotka-Volterra equations is used for demonstration.

1. Introduction

Model form error (MFE) is a significant challenge in computational simulations, where mathematical models are used to represent complex physical phenomena. It refers to the difference between the mathematical model and the true behavior of the system being simulated. This error can arise from various sources, such as neglecting certain physical phenomena, using simplified mathematical equations, or making assumptions about parameter values. MFE is a fundamental aspect of the model itself and is independent of any specific data or observations. Addressing MFE involves refining the mathematical representation of the system (Oberkampf, DeLand, Rutherford, Diegert, & Alvin, 2002).

A related but distinct concept from MFE is model discrepancy. Model discrepancy refers to the difference between the simulation results obtained from a particular model and the observed or experimental data. It represents the difference between the model predictions and the actual behavior of the system. Model discrepancy can arise due to various factors, including measurement errors, uncertainties in input data, or limitations in the experimental setup. Model discrepancy is typically quantified by comparing the simulation results with experimental data and can be influenced by both MFE and other sources of uncertainty (Kennedy & O'Hagan, 2001).

To further illustrate the distinction between MFE and model discrepancy, let's consider an example in the context of fluid dynamics simulations. In fluid dynamics, MFE would refer to the simplifications and assumptions made in the mathematical equations used to represent fluid flow. For instance, the Navier-Stokes equations, which govern fluid flow, often require assumptions such as incompressibility, isotropy, and neglecting certain small-scale turbulent effects (Reynolds, 1976). These simplifications may introduce differences between the model and the true behavior of fluid flow in specific scenarios.

Model discrepancy, on the other hand, would refer to the differences between the predictions of the fluid dynamics model and the observed flow behavior in a specific system. This discrepancy can arise due to various factors, including measurement errors in data collection, uncertainties in estimating model parameters, or unaccounted-for physical phenomena that influence fluid flow. Model discrepancy captures the overall difference between the model predictions and the actual behavior of the fluid flow, taking into account both MFE and other sources of uncertainty.

An ongoing research challenge is that MFE cannot be directly estimated since the true equations governing a real-world system are unknown. At best, all that is available from the true system is observations of system response quantities. Hence historically, most research has focused on addressing model discrepancy. One of the most popular approaches was proposed by Kennedy and O'Hagan where they represent model discrepancy using a Gaussian process model (GPM) that incorporates test settings and estimates the hyper parameters of the GPM simultaneously with uncertain model parameters from measured system responses (Kennedy & O'Hagan, 2001). In their study, Neal et al. (Neal, Hu, Mahadevan, & Zumberge, 2019) addressed model discrepancy in a time-dependent simulation of an air cycle machine by utilizing state estimation.

Recently there has been a push to correct models at the source, meaning to address MFE in the mathematical model underpinning the computational simulation. Oliver et al. (Oliver, Terejanu, Simmons, & Moser, 2015) suggest that correcting simulations at the source of the error improves prediction accuracy beyond the observed data. Sargsyan et al. (Sargsyan, Najm, & Ghanem, 2015) embed a correction within the model by augmenting certain model parameters with probabilistic correction terms. Morrison et al. (Morrison, Oliver, & Moser, 2018) represented MFE by a finite-dimensional operator acting on a vector of state variables, which was further investigated by Portone and Moser (Portone & Moser, 2022) for a contaminant transport problem through heterogeneous media. Subramanian and Mahadevan (Subramanian & Mahadevan, 2019) used Bayesian state estimation to estimate MFE as an additive forcing term from available experimental data of system responses.

In real-world applications, there is often limited observation data of system states. Observations require instrumentation and monitoring of systems, which can be expensive or infeasible given physical limitations in the environments of interest. Sparse data creates a challenge for data assimilation methodologies like the one proposed in (Subramanian & Mahadevan, 2019). However, there may be domain knowledge available for real-world applications. An expert in a particular field may have an idea of the missing physics even if the exact equations are unknown. Inclusion of domain knowledge offers the potential to improve MFE estimation.

In this work, the MFE estimation approach developed by (Subramanian & Mahadevan, 2019) is investigated in the presence of sparse observation data and is extended to incorporate subject matter expert (SME) knowledge about the MFE.

2. DEMONSTRATION PROBLEM

The Lotka-Volterra equation, also known as the predator-prey equation, is a mathematical model that describes the interaction between two species in an ecosystem. It was developed independently by Alfred J. Lotka and Vito Volterra in the early 20th century. The equation consists of a pair of first-order nonlinear differential equations, one representing the population growth of the prey species and the other representing the population decline of the predator species. The model assumes that the prey population grows exponentially in the absence of predators, while the predator population declines proportionally to the rate at which it consumes the prey. The Lotka-Volterra equation provides valuable insights into the dynamics of predator-prey relationships and has applications in various fields, including ecology, population biology, and economics.

The two species Lotka-Volterra equation is

$$\begin{aligned} \frac{dx}{dt} &= \alpha x - \beta xy \\ \frac{dy}{dt} &= \delta xy - \gamma y\,, \end{aligned} \tag{1}$$

where x is the number of prey and y is the number of predators. For this demonstration, the model parameters are defined as

$$[\alpha, \beta, \delta, \gamma] = [1.5, 1, 3, 1]\,, \tag{2}$$

where α is the maximum prey per capita growth rate, β is the effect of the presence of predators on the prey growth rate, δ is the effect of the presence of prey on the predator's growth rate, and γ is the predator's per capita death rate. For demonstration, the initial conditions for the two states are set at

$$[x_0, y_0] = [1, 1]\,. \tag{3}$$

Solving the coupled ODEs defined in Eqs. 1-3 using a numerical integration scheme produces the time-dependent system states shown in Fig. 1.

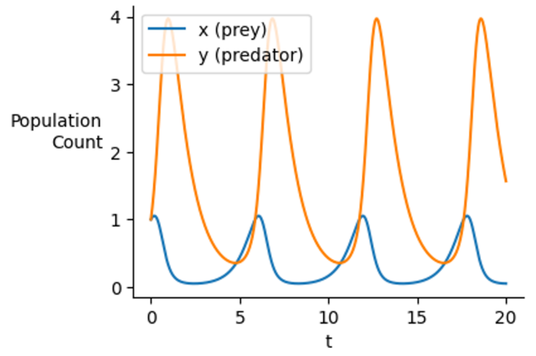

Figure 1. True system states for the Lotka-Volterra model

Synthetic experimental data is generated from the true system states in Fig. 1 by specifying an observation time step. Two data scenarios are considered: dense observation data shown in Fig. 2 and sparse observation data shown in Fig. 3. The effect of measurement noise is not considered in this example but could be studied in the future.

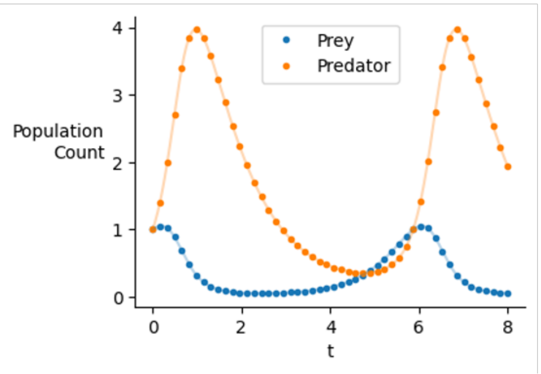

Figure 2. Dense observation data

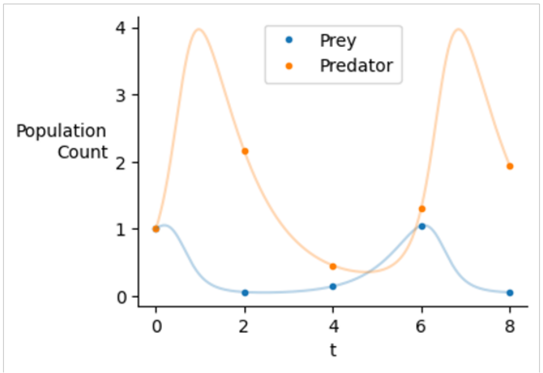

Figure 3. Sparse observation data

To emulate the scenario where the modeled mathematical equations do not capture the true behavior of the real system, the following deficient mathematical model is proposed

$$\begin{aligned}\frac{dx}{dt} &= \alpha x\\ \frac{dy}{dt} &= \delta xy - \gamma y\,,\end{aligned} \tag{4}$$

where the second term in the first equation has been omitted.

We will assume that it is known that the MFE, Q, exists in the first equation as

$$\frac{dx}{dt} = \alpha x - Q\,, \tag{5}$$

so MFE estimation will involve estimating Q from observation data of x and y. Because this is a contrived example, the true MFE term is known and is

$$Q_{true} = \beta xy\,. \tag{6}$$

We will investigate two formulations for how Q evolves in time. In the case where there is no knowledge about how Q evolves, the hypothesized dynamics #1 is written as

$$Q_t = Q_{t-1} + N(0, 5e^{-1})\,, \tag{7}$$

where $N(0, 5e^{-1})$ indicates a normal distribution with 0 mean and a variance of 0.5. Eq. (7) means that Q at the current time step is equal to Q at the previous time step plus some Gaussian noise. The Δt is a constant $\frac{2}{150}$. Alternatively, a SME may have an inclination of how Q evolves in time but not be confident in their understanding. In this case, the hypothesized dynamics #2 is

$$Q_t = \beta xy + N(0, 1e^{-1})\,, \tag{8}$$

so the SME did indeed provide the true functional relationship to system states for Q, but their uncertainty is captured through additive 0-mean Gaussian noise with a 0.1 variance. Within data assimilation, either formulation of Q, i.e. Eq. (7) or Eq. (8), can be utilized to simulate how Q evolves over time.

3. RESULTS

Data assimilation is performed through a basic particle filter implementation (Carpenter, Clifford, & Fearnhead, 1999) to simultaneously estimate the system states, x and y, and the MFE term Q.

The resulting estimated system states are plotted next for three cases; (1) dense data and the uninformed model of Q, (2) sparse data and the uninformed model of Q, and (3) sparse data and the informed model of Q. In Fig. 4, the uncertainty grows as the model advances in time, but the uncertainty is quickly reduced at each time step where an observation is present. In Fig. 5, the sparsity of observations allows the estimated states to diverge from the true states by orders of magnitude between the five data observations. Clearly the best estimation occurs when using the SME-informed hypothesized dynamics #2 for Q, even in the presence of sparse data. In Fig. 6, the median estimates have close agreement with the known true states and the interquartile range (IRQ) shows that the variance is small.

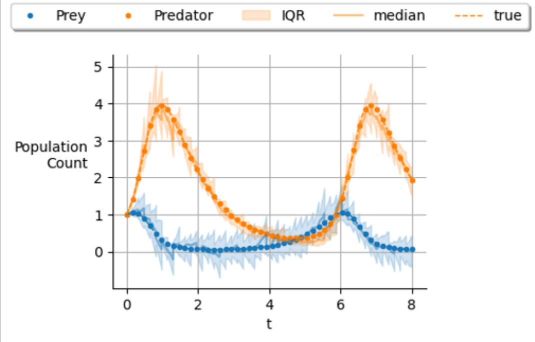

Figure 4. Estimate of system states with dense data and hypothesized dynamics #1.

Next, the estimates of Q are shown, and they tell a similar story as the state estimates. Q is different from x and y,

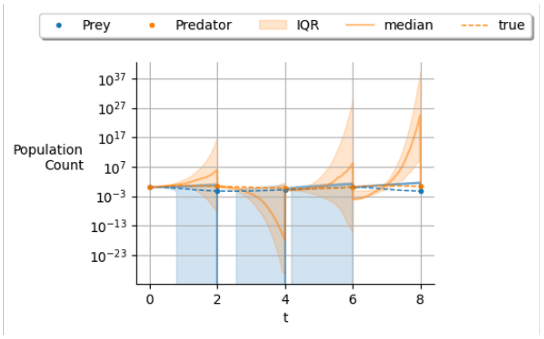

Figure 5. Estimate of system states with sparse data and hypothesized dynamics #1.

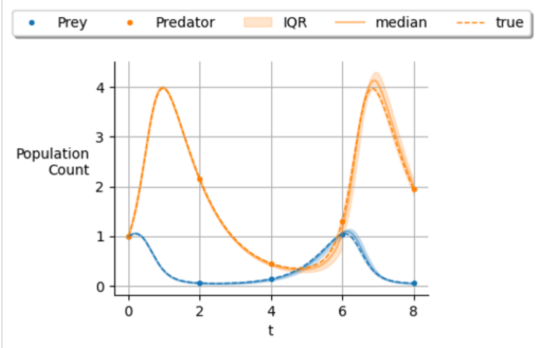

Figure 6. Estimate of system states with sparse data and hypothesized dynamics #2.

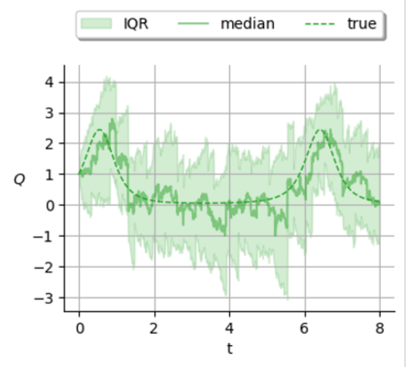

Figure 7. Estimate of the missing MFE term with dense data and hypothesized dynamics #1.

though, since it cannot be directly observed as it is a hidden state. In Fig. 8, the variance grows between observations while the median is relatively constant, which is expected given the evolution of Q defined in Eq. 7. A noticeable difference in the bias and variance of the estimates of Q exist between Fig. 7 and Fig. 9, which indicates that even dense observation data cannot compensate for the absence of SME knowledge.

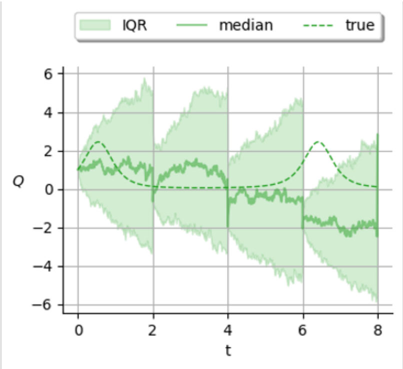

Figure 8. Estimate of the missing MFE term with sparse data and hypothesized dynamics #1.

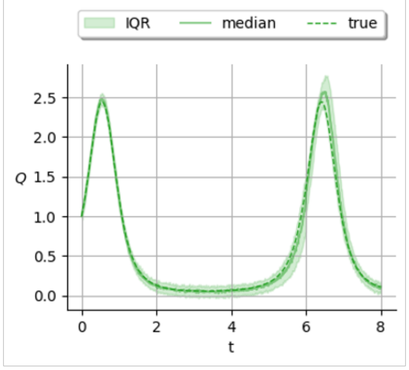

Figure 9. Estimate of the missing MFE term with sparse data and hypothesized dynamics #2.

4. CONCLUSIONS

This study implemented a recent data assimilation-based approach for MFE estimation in the context of the Lotka-Volterra (predator-prey) model. Performance of the methodology was investigated in the presence of sparse data and with the incorporation of domain knowledge. This limited study indicates that sparse data does reduce the performance of MFE estimation; however, inclusion of SME knowledge through a physics-informed process model for the evolution of MFE can more than mitigate the challenges with sparse data.

The Lotka-Volterra model examined here is a relatively simple system of two ODEs. Many real-world problems will be governed by high-dimensional PDEs, so we will pursue extending this work to PDEs in the future.

ACKNOWLEDGEMENTS

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration

(DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

The authors would also like to acknowledge the contributions of ChatGPT 3.5 Turbo (OpenAI, 2021), the AI language model developed by OpenAI, in the development of this research paper. Its assistance in generating text from user prompts has enhanced the quality and efficiency of this research endeavor. We thank the developers and researchers at OpenAI for their dedication and expertise in creating and improving ChatGPT 3.5 Turbo.

REFERENCES

Carpenter, J., Clifford, P., & Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, *146*(1), 2–7.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464.

Morrison, R. E., Oliver, T. A., & Moser, R. D. (2018). Representing model inadequacy: A stochastic operator approach. *SIAM/ASA Journal on Uncertainty Quantification*, *6*(2), 457–496.

Neal, K., Hu, Z., Mahadevan, S., & Zumberge, J. (2019). Discrepancy prediction in dynamical system models under untested input histories. *Journal of Computational and Nonlinear Dynamics*, *14*(2), 021009.

Oberkampf, W. L., DeLand, S. M., Rutherford, B. M., Diegert, K. V., & Alvin, K. F. (2002). Error and uncertainty in modeling and simulation. *Reliability Engineering & System Safety*, *75*(3), 333–357.

Oliver, T. A., Terejanu, G., Simmons, C. S., & Moser, R. D. (2015). Validating predictions of unobserved quantities. *Computer Methods in Applied Mechanics and Engineering*, *283*, 1310–1335.

OpenAI. (2021). *Sandia national laboratories chatgpt (chatgpt 3.5 turbo) [large language model].* Retrieved from `https://ai.sandia.gov/chat`

Portone, T., & Moser, R. D. (2022). Bayesian inference of an uncertain generalized diffusion operator. *SIAM/ASA Journal on Uncertainty Quantification*, *10*(1), 151–178.

Reynolds, W. C. (1976). Computation of turbulent flows. *Annual Review of Fluid Mechanics*, *8*(1), 183–208.

Sargsyan, K., Najm, H. N., & Ghanem, R. (2015). On the statistical calibration of physical models. *International Journal of Chemical Kinetics*, *47*(4), 246–276.

Subramanian, A., & Mahadevan, S. (2019). Error estimation in coupled multi-physics models. *Journal of Computational Physics*, *395*, 19–37.