CSCI 5822 Final Project

Rileigh Bandy

May 2, 2021

1 Problem Definition

Real-world physical systems can be extremely complex with many random variables and influential parameters. It quickly becomes intractable to model all of the variables in the system, so we rely on partial models. A partial model trades accuracy for efficiency so it can save resources, but accurate assessments of the error are essential in decisions about the use of such models. Model error is the inconsistency between model output and observations from the detailed system caused by the inaccuracy in the model, such as simplifications, missing dependencies, or empirical relationships. This project will explore reducing model error through model correction in the form of efficient, interpretable modifications of a partial model. Specifically, we will focus on the Bayesian calibration of a modified partial model.

To explore model correction with a popular interaction model, we will use the generalized Lotka-Volterra (GLV) equations, which model the time dynamics of interacting species. We will simulate a detailed system as the GLV equations with n species plus random noise. In the context where only m species are of interest and $m < n$, the modeler can design a partial model involving only those m species. However, the model error can be prohibitively large. We propose a corrected partial model, called the enriched model, and use Bayesian calibration and validation of that enriched model. In this project, we will focus on the calibration of the enriched model.

1.1 Detailed Model

The GLV equations for the detailed model, $\mathcal{D}$, are written as:

$$\frac{d\hat{\mathbf{x}}}{dt} = \mathcal{D}(\hat{\mathbf{x}}) = \text{diag}\left(\hat{\mathbf{x}}\right)\left(\hat{\mathbf{r}} + \hat{A}\hat{\mathbf{x}}\right), \tag{1}$$

where $\hat{\mathbf{x}} \in \mathbb{R}^n_{\geq 0}$ are species concentrations, the vector $\hat{\mathbf{r}} \in \mathbb{R}^n$ represents the intrinsic growth rates, and the matrix $\hat{A} \in \mathbb{R}^{n \times n}$ contains the interaction rates between species. In $\hat{A}$ the ijth entry, a_{ij}, indicates how species j affects the concentration of species i. The GLV equations with n species may have up to 2^n equilibria, out of which at most one is feasible and called the coexistence equilibrium [2]. From eq. (1), this equilibrium is given when all the $\hat{\mathbf{x}}$ are assumed nonzero:

$$\begin{aligned} \hat{\mathbf{r}} + \hat{A}\hat{\mathbf{x}}_{eq} &= 0 \\ \hat{\mathbf{x}}_{eq} &= -\hat{A}^{-1}\hat{\mathbf{r}}, \end{aligned} \tag{2}$$

where $\hat{\mathbf{x}}_{eq}$ denotes the equilibrium concentrations. This equilibrium is feasible if $\hat{\mathbf{x}}_{eq}$ is strictly positive.

1.2 Partial Model

The GLV equations for the partial model, $\mathcal{P}$, are written as:

$$\frac{d\mathbf{x}}{dt} = \mathcal{P}(\mathbf{x}) = \text{diag}\left(\mathbf{x}\right)\left(\mathbf{r} + A\mathbf{x}\right), \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^m_{\geq 0}$ is a subvector of m elements of $\hat{\mathbf{x}}$, and $\mathbf{r}$ and A are the corresponding m-vector of $\hat{\mathbf{r}}$ and $m \times m$ matrix of $\hat{A}$. From eq. (3), the coexistence equilibrium is given as:

$$\mathbf{x}_{eq} = -A^{-1}\mathbf{r}. \tag{4}$$

1.3 Enriched Model

The purpose of the enriched model is to reduced model error without significantly increasing the efficiency with respect to the partial model. To do so, we take the partial model of m species and augment it with a discrepancy model, Δ, which includes important information from the detailed model. We require that the enriched model only involves information from the m species to simulate the intractability of measuring every interaction in the detailed system. The discrepancy model can include proportional, integral, or derivative information from the m species, and it is common to use Occam's Razor in the selection of an optimal discrepancy model.

For this project, the enriched model, $\mathcal{M}$, we will consider is written as:

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \mathcal{M}(x) \\ &= \mathcal{P}(\mathbf{x}) + \text{diag}(\mathbf{x})\boldsymbol{\delta}_0 \\ &= \mathcal{P}(\mathbf{x}) + \Delta(x), \end{aligned} \tag{5}$$

where $\boldsymbol{\delta}_0 = (\delta_{10}, \delta_{20}, \ldots, \delta_{m0})^T$. The subscripts on each δ_{ij} are chosen so that i indicates that this coefficient appears in the right-hand side of the variable x_i, and j indicates that this coefficient is multiplying the jth derivative of x_i following Morrison's discrepancy operator construction [5]. The model parameters $\boldsymbol{\delta}_0$ will be distributions $\delta_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$, where $\mu_{ij} \in \mathbb{R}$ and $\sigma_{ij}^2 \in \mathbb{R}_{\geq 0}$ are hyperparameters. Following the construction of the model parameters, the vecotorized hyperparameters are $\boldsymbol{\mu}_0 = (\mu_{10}, \mu_{20}, \ldots, \mu_{m0})^T$ and $\boldsymbol{\sigma}_0^2 = (\sigma_{10}^2, \sigma_{20}^2, \ldots, \sigma_{m0}^2)^T$.

1.4 Constraints

Given this project's finite time frame and the infinite number of possible GLV equations model realizations, we need to refine the problem's scope. To reduce computation time, we will fix the detail model's size to $n = 10$. Then, the possible partial models could be from $m = 1$ to $m = 9$. We would like to investigate all of these partial models, but we will select $m = 4$ for an initial partial model to investigate in this project. **Inference Problem**: Before quantifying the enriched model's error or utilizing the enriched model for predictions, the discrepancy model needs to be calibrated to find optimal selections for $\boldsymbol{\delta}_0, \boldsymbol{\mu}_0$, and $\boldsymbol{\sigma}_0^2$.

The interaction rates and growth rates, $\hat{A}$ and $\hat{\mathbf{r}}$, of the GLV equations determine the model's stability. In this project, we will focus on models that converge to the coexistence equilibrium; we will investigate two situations.

Case One: First, we will begin with the strictly competitive GLV equations from [5], where $\hat{A}$ is symmetric, diagonally-dominant, and all of the entries are negative. **Case Two**: Next, we will let the interactions between species be competitive or cooperative, while the interactions within species remain competitive, by making $\hat{A}$ asymmetric and diagonally dominant. Note, the within-species interactions must remain negative and diagonally dominant to ensure negative eigenvalues and convergence to a stable equilibrium. The algorithms to randomly generate the interaction rates and growth rates for the detailed model and partial model are given in appendix A. Since these simulations do not represent a real system, initial species concentrations $\hat{\mathbf{x}}_0$ are arbitrarily selected such that they are nonnegative and held constant between the two cases.

1.5 Observations

Following [5], the datasets from the detailed system are simulated by the detailed model's trajectories of the m species from the partial model, and T equally spaced observations are taken from each trajectory. The observations can be defined as

$$\mathcal{O}^k = \{y_{ij}\}, i = 1, \ldots, m; j = 1, \ldots, T; k = 1, 2, \tag{6}$$

where $\{y_{ij}\}$ is the observation $x_i(t_j)$ given the case k. Random, Gaussian noise, ϵ, with a distribution:

$$p_\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{7}$$

is added to the species concentration generated by the detailed model, y_{ij}^t, to simulate observational error, and the observations become

$$y_{ij} = y_{ij}^t + \epsilon. \tag{8}$$

For this project, observations are generated with $T = 10$ and $\sigma_\epsilon^2 = 0.001$.

2 Markov and Belief Networks

The belief network in vectorized form is illustrated in fig. 1, and the Markov network is given in fig. 2. For this inference problem, the belief network is more appropriate because the arrow directions represent important dependencies in the inference problem, and the Markov network is a complicated mess that is almost a fully connected graph.

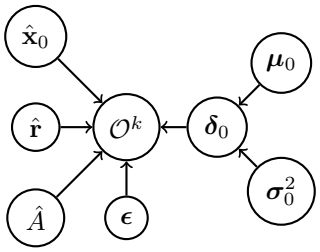


Figure 1: Belief Network of vectorized variables.

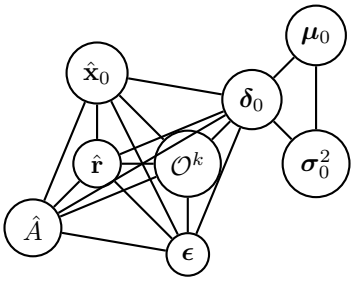
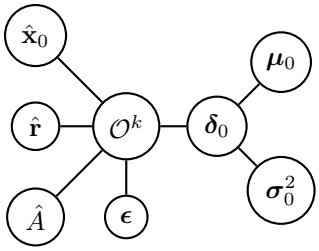


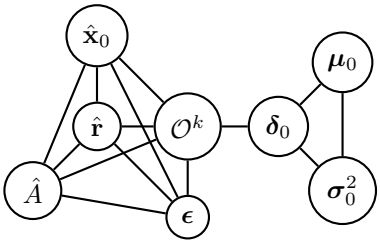
Figure 2: Markov Network of vectorized variables.

3 Graph Skeleton, Moralized Graph, and the Triangulated Graph

Note, the moralized graph is already triangulated.



(a) Graph Skeleton



(b) Moralized and Triangulated Graph

Figure 3: Drawing of the graph skeleton, the moralized graph, and the triangulated graph.

4 Approximate Parameter Inference

4.1 Prior Distribution

The prior distribution is given by

$$\pi_{prior} = p\left(\boldsymbol{\delta}_0 \mid \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2\right) p\left(\boldsymbol{\mu}_0\right) p\left(\boldsymbol{\sigma}_0^2\right), \tag{9}$$

where

$$\begin{aligned} p\left(\boldsymbol{\delta}_0 \mid \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2\right) &= \mathcal{N}\left(\boldsymbol{\delta}_0 \mid \boldsymbol{\mu}_0, \operatorname{diag}\left(\boldsymbol{\sigma}_0^2\right)\right) \\ &= \prod_{i=1}^{m} \mathcal{N}\left(\delta_{i0} \mid \mu_{i0}, \sigma_{i0}^2\right). \end{aligned} \tag{10}$$

Since the model parameters, $\boldsymbol{\delta}_0$, are defined as univariate Gaussian distributions in the discrepancy model. The hyperparameters, $\boldsymbol{\sigma}_0^2$, represent variance, which must be nonnegative. From previous work and [5], we can infer that the expectation of $\boldsymbol{\sigma}_0^2$ is relatively small, say 0.1, so the exponential distribution with $\lambda = 0.1$ defined as

$$p\left(\boldsymbol{\sigma}_0^2\right) = \prod_{i=1}^{m} \lambda \exp\left\{\lambda \sigma_{0i}^2\right\} \tag{11}$$

would be the maximum entropy distribution. The prior distribution of $\boldsymbol{\mu}_0^2$ is case dependent. **Case One**: When the interaction matrix is strictly competitive, the fact that all of the entries must be negative provides an upper bound because $\delta_{ij} < 0$. The diagonal dominance of $\hat{A}$ means $a_{ii} < \delta_{ij}$ and provides a lower bound. Following the maximum entropy distribution,

$$p\left(\boldsymbol{\mu}_0\right) = \prod_{i=1}^{m} \mathcal{U}\left(a_{ii}, 0\right). \tag{12}$$

Case Two: When the species interactions are competitive or cooperative, the upper bound becomes $-a_{ii} > \delta_{ij}$, and the the maximum entropy distribution is

$$p\left(\boldsymbol{\mu}_0\right) = \prod_{i=1}^{m} \mathcal{U}\left(a_{ii}, -a_{ii}\right). \tag{13}$$

4.2 Likelihood

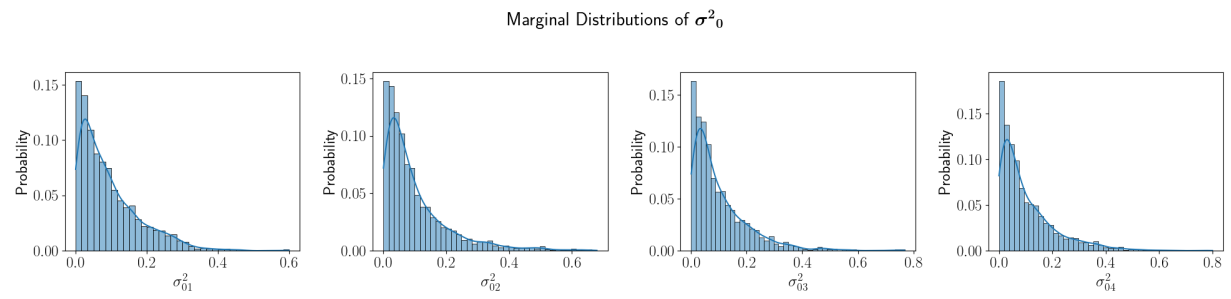
The likelihood distribution is given by

$$\begin{aligned} \pi_{like} &= p\left(\mathcal{O}^k \mid \boldsymbol{\delta}_0, \hat{A}, \hat{\mathbf{r}}, \hat{\mathbf{x}^0}, \boldsymbol{\epsilon}\right) \\ &= \mathcal{N}\left(\mathcal{O}^k \mid \boldsymbol{\delta}_0, \Sigma_\epsilon\right) \\ &\propto \prod_{i=1}^{Tm} \exp\left\{-\frac{1}{2} \frac{\left(\mathcal{O}_i^k - Y(\boldsymbol{\delta}_0)_i\right)^2}{\sigma_\epsilon^2}\right\}. \end{aligned} \tag{14}$$

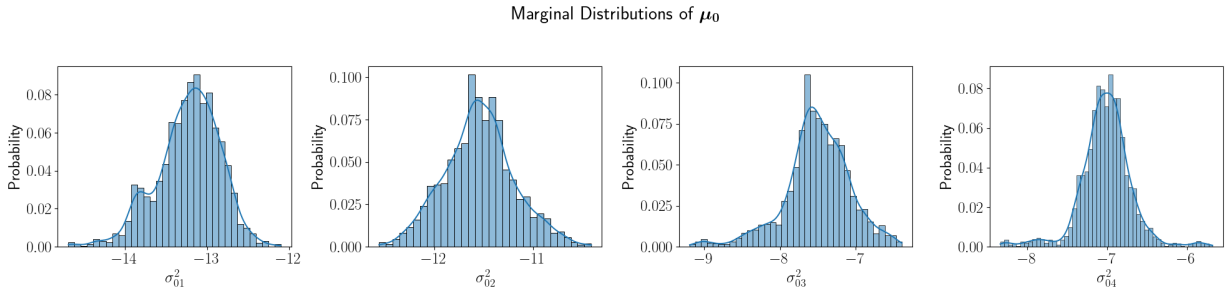
The likelihood distribution is Gaussian since the independent, identically distributed additive noise is Gaussian in eq. (7). Since this is a simulation where we know the exact distribution of the noise, the likelihood distribution should be correct.

4.3 Implement Approximate Inference

We chose to implement Markov chain Monte Carlo (MCMC) with Gibbs Sampling. All of the code can be found in this public repository [1].



(a)



(b)

Figure 4: Hyperparameter marginal distributions for case one.

4.4 Marginal Probabilities

Since the marginal distribution of the model parameters $\boldsymbol{\delta}_0$ is a distribution over distributions, the best way the plot it is not immediately clear and will be omitted. The marginal distributions of the sampled hyperparameters for case one are given in fig. 4, and the marginal distributions of the sampled hyperparameters for case two are given in fig. 5. Since the parameters are not the quantity of interest of the enriched model, these visualizations are not very useful for analysis of the enriched model. Instead, we should use the generated parameter distributions to create a distribution of species concentration trajectories that we can compare to observations from the detailed system and the partial model. See appendix B for species concentration trajectories.

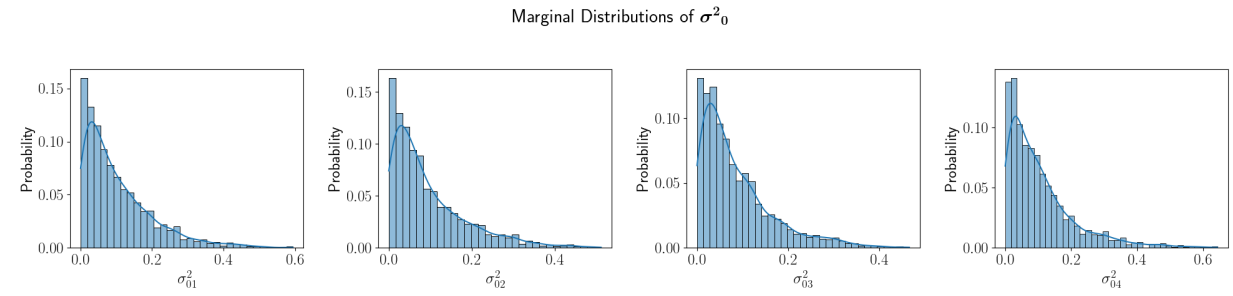
4.5 Sampling Algorithm Assumptions

Gibbs sampling assumes we can traverse the full posterior distribution, where every state can be visited infinitely often. This assumption fails when there are regions not connected by a probable Gibbs path, which can happen in a high dimensional Gaussian distribution like our posterior. The best way to circumvent this issue is by using parallel chains. Parallel chains can be as efficient if not more efficient than standard Gibbs sampling assuming the chains' starting positions are well selected.

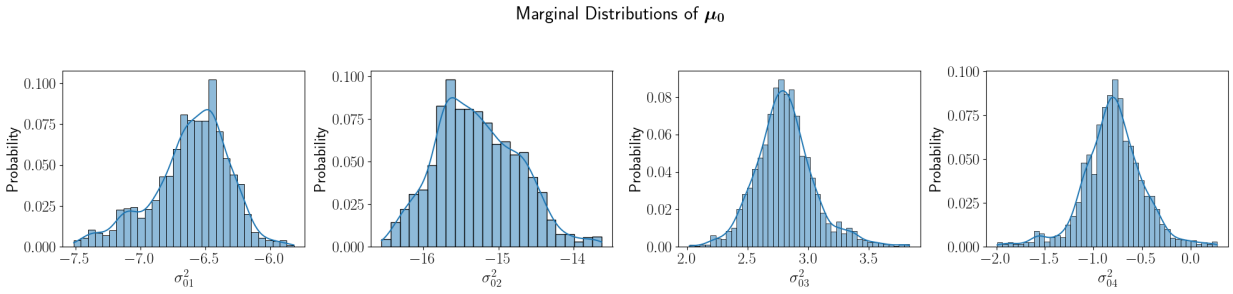
Furthermore, there are alternative MCMC algorithms that are more efficient than Gibbs sampling. A Hamiltonian Monte Carlo (HMC) method is probably better suited for this problem because of the large number of random walk samples we have to discard in Gibbs sampling to generate "independent" samples. Efficiency measurements comparing HMC to random walk MCMC can be found in [3, 4, 7].

4.6 Predictions

Before using our enriched model for extrapolative predictions, we want to be confident that the enriched model is an accurate model of the detailed system, which would involve validation tests and a posterior predictive check detailed in [6]. Since model validation is outside the scope of this class, we will assume the enriched model is valid. Then, we could use the distributions of the model parameters and hyperparameters to



(a)



(b)

Figure 5: Hyperparameter marginal distributions for case two.

predict species concentration trajectories for unobserved initial concentrations, growth rates, and interaction matrices. In general, we could use the enriched model in place of the detailed system to save computation time and resources. However, we should only use the enriched model for GLV equations that converge to the coexistence equilibrium.

References

[1] R. Bandy, *Mcmc with gibbs sampling*. https://github.com/rbandy/MCMC_with_Gibbs/tree/master, 2021.

[2] G. Barabás, M. J. Michalska-Smith, and S. Allesina, *The effect of intra-and interspecific competition on coexistence in multispecies communities*, The American Naturalist, 188 (2016), pp. E1–E12.

[3] M. Girolami and B. Calderhead, *Riemann manifold langevin and hamiltonian monte carlo methods*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (2011), pp. 123–214.

[4] M. D. Hoffman and A. Gelman, *The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.*, J. Mach. Learn. Res., 15 (2014), pp. 1593–1623.

[5] R. E. Morrison, *Data-driven corrections of partial lotka–volterra models*, Entropy, 22 (2020), p. 1313.

[6] R. E. Morrison, T. A. Oliver, and R. D. Moser, *Representing model inadequacy: A stochastic operator approach*, SIAM/ASA J. Uncertain., 6 (2018), pp. 457–496.

[7] Y. Zhang and C. Sutton, *Semi-separable hamiltonian monte carlo for inference in bayesian hierarchical models*, arXiv preprint arXiv:1406.3843, (2014).

A Generating GLV Model Realizations

This section outlines how to generate the two types of interaction matrices and the their corresponding growth rates. Note, algorithm 1 and algorithm 3 are taken from [5]. The strictly competitive $\hat{A}$ and corresponding $\hat{\mathbf{r}}$ are created in algorithm 1. The competitive and cooperative $\hat{A}$ and corresponding $\hat{\mathbf{r}}$ are created in algorithm 2, and the partial model's interaction rates and growth rates, A and $\mathbf{r}$, are defined in algorithm 3. For this project, $\sigma_B^2 = \sigma_C^2 = 1.0$.

Algorithm 1: Generating a strictly competitive detailed model.

Initialize n
Sample $B_{ij} \sim log\mathcal{N}(0, \sigma_B^2)$, $1 \leq i < j \leq n$
Set $B_{ji} = B_{ij}$
Sample $C_{ii} \sim log\mathcal{N}(0, \sigma_C^2) + \sum_{k \neq i} B_{ki}$, $1 \leq i <\leq n$
Set interaction matrix $\hat{A} = -(B + C)$
Set growth rate vector $\hat{\mathbf{r}} = \max\{C\}\mathbf{1}_n$
Return $D = \{\hat{A}, \hat{\mathbf{r}}\}$

Algorithm 2: Generating a competitive/cooperative detailed model.

Initialize n
Sample $B_{ij} \sim log\mathcal{N}(0, \sigma_B^2)$, $1 \leq i \leq n$, $1 \leq j \leq n$
Set $B_{ii} = 0$
Sample $C_{ii} \sim log\mathcal{N}(0, \sigma_C^2) + \sum_{k \neq i} B_{ki}$, $1 \leq i <\leq n$
Sample $flip \sim$ randint(range $= \{-1, 1\}$, size $= (n, n)$)
Set $flip_{ii} = 1$
Set interaction matrix $\hat{A} = -(B + C) * flip$
Set growth rate vector $\hat{\mathbf{r}} = \max\{C\}\mathbf{1}_n$
Return $D = \{\hat{A}, \hat{\mathbf{r}}\}$

Algorithm 3: Subsampling the partial model.

Initialize $m < n, D$
Set A as submatrix $A = \hat{A}_{1:m,1:m}$
Set $\mathbf{r}$ as subvector $\mathbf{r} = \hat{\mathbf{r}}_{1:m}$
Return $P = \{A, \mathbf{r}\}$

B Quantity of Interest Results

The species concentrations over time are the quantity of interest (QoI) in this problem. After generating distributions for $\boldsymbol{\delta}_0, \boldsymbol{\mu}_0$, and $\boldsymbol{\sigma}_0^2$, we can push $\boldsymbol{\delta}_0$ through the enriched model and generate a distribution of species concentrations over time. Since the QoI is a distribution, we can compute confidence intervals. In figs. 6 and 7 the enriched model's median value, 50% confidence interval, and 95% confidence interval correspond to the dark blue plot, blue shading, and light blue shading. Visually, these results indicate the enriched model is capturing the observations, but further calibration and validation tests need to be conducted before we can confidently use the enriched model in place of the detailed system.

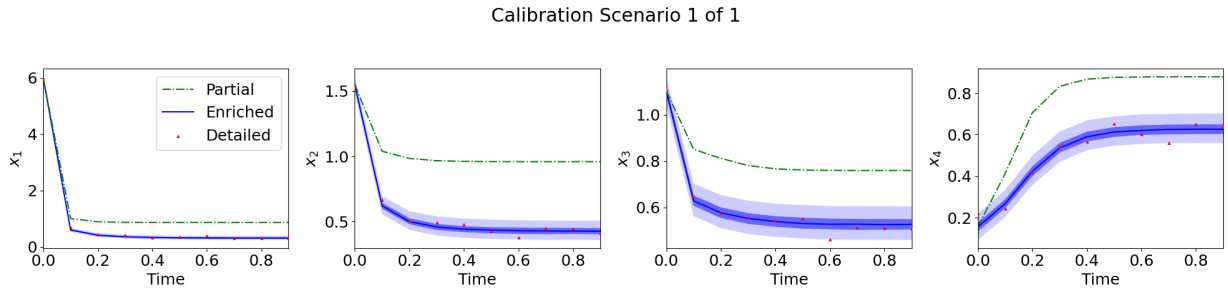


Figure 6: Observations from the detailed system compared to enriched and partial models for case one.

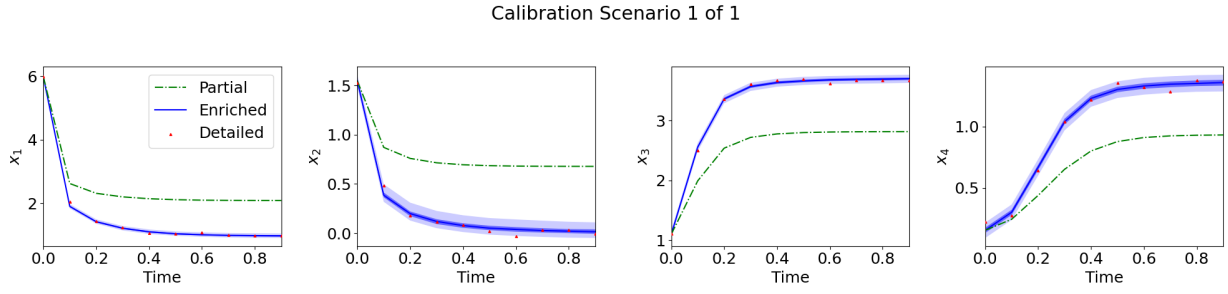


Figure 7: Observations from the detailed system compared to enriched and partial models for case two.