

# Forecasting and uncertainty quantification using a hybrid of mechanistic and non-mechanistic models for an age-structured population model

John Lagergren, Amanda Reeder, Franz Hamilton, Ralph C. Smith, Kevin B. Flores

Department of Mathematics  
Center for Research in Scientific Computation  
North Carolina State University

## Abstract

Traditionally, either mechanistic or non-mechanistic modeling techniques have been used for prediction, however it is uncommon for the two to be incorporated together. We compare the forecast accuracy of mechanistic modeling, using Bayesian inference, a non-mechanistic modeling approach based on state space reconstruction, and a hybrid composed of the two for an age-structured population data set. The data come from cannibalistic flour beetles, in which it has been observed that the adults preying on the eggs and pupae results in non-equilibrium population dynamics. Uncertainty quantification methods for the hybrid models are outlined and illustrated on these data. We perform an analysis of the results from Bayesian inference for the mechanistic model and hybrid models to suggest reasons why hybrid modeling may enable more accurate forecasts of multivariate systems than traditional approaches.

## 1 Introduction

Mechanistic modeling strategies for predicting multivariate biological systems involve relying on a dynamical model; i.e., differential equations, to describe the biological mechanisms and interactions that affect the evolution of the system [1, 5, 33]. Applications of this strategy to genetic networks, neuronal networks, and population dynamics has enabled the prediction of complex and emergent behaviors in these systems [1]. A central challenge in utilizing a mechanistic model for prediction is the ability to accurately parameterize it from available time series data, which can often be sparse and noisy in biological settings [2, 3]. Commonly encountered challenges that confound the ability to accurately parameterize a model can be attributed to problems related to some combination of model discrepancy and parameter identifiability [9, 33, 38]. Thus, the development of methodologies to reduce the effects or presence of these challenges may enable the application of the mechanistic modeling strategy to a wider spectrum of intractable data sets arising from complex biological systems.

Model discrepancy is an inherent issue when developing a mathematical model that approximates a biological system [5, 38]. Ideally, a mathematical model is developed to achieve a balance between model complexity and the ability to parameterize the model using available data, with the ultimate goal of maximizing predictive values for out-of-sample data. A general principle is to reduce the mathematical model description to the lowest dimension possible; i.e., with the least number of variables and parameters. Whereas "hold-out" validation approaches are often used to evaluate the ability of the model to predict out-of-sample data [15], to the best of our knowledge, no systematic methodologies exist for minimizing model dimensionality while simultaneously maximizing prediction accuracy. Even in cases of full observability; i.e., when every variable in the model is a longitudinal covariate in the available time series data, accurate parameterization may still be a challenge due to identifiability-related issues [10, 20, 23, 27]. If parameters are not structurally identifiable with respect to an observed set of data, one can attempt to reparameterize the model and then estimate aggregated parameters [23]. One drawback with this technique is that if the primary goal

is to accurately estimate parameters; e.g., to infer the kinetic rates of biological interactions, the aggregated parameters may not be biologically meaningful or interpretable. An alternative approach is to use subset selection techniques to find identifiable combinations of parameters and then fix the non-identifiable parameters to constant values [4, 9]. However, one then encounters the issue of having to justify those fixed values from other sources of experimental data and be able to ensure that model predictions are not sensitive at the values to which the parameters are fixed.

Alternative paradigms exist to forecast time series data without a mechanistic model; we refer to these approaches as “non-mechanistic models”. These include empirical dynamical modeling [39], autoregressive models, e.g., NARX [6], and machine learning, e.g., multi-step ahead prediction [8, 25]. Since these methods do not rely on developing a mechanistic model based on biological knowledge, they do not include parameters that correspond to biologically interpretable quantities; e.g., kinetic rates. Thus, if the goal is to estimate biologically meaningful parameters from time series data, the mechanistic modeling strategy is the only compatible approach. In general, it has been noted that a primary drawback of utilizing non-mechanistic modeling in forecasting is that one forfeits the transferability and theoretical understanding afforded by a validated mechanistic model [18]. While these concerns have been previously noted, our focus here is to present a hybrid strategy that leverages the advantages of both mechanistic and non-mechanistic modeling to maximize predictive accuracy and minimize forecast uncertainty.

We chose to merge two well known methods, state space reconstruction and Bayesian inference, to investigate whether their combination could minimize the drawbacks encountered when utilizing each method separately. The state space reconstruction (SSR) methodology relies on Taken’s theorem of delay embedding and uses time series data to generate a manifold that is one-to-one with the attractor manifold of a dynamical system [7, 13, 17, 19, 21, 22, 28–32, 34, 35, 37, 40]. In theory, since the reconstructed manifold is one-to-one with the attractor of the real system, one can use it to forecast future dynamics using a nearest neighbor approach as described in Section 2.3. A critical limitation to using SSR for prediction is the amount of data needed to accurately reconstruct the attractor manifold. Since no biological knowledge is leveraged, SSR, similar to other non-mechanistic modeling approaches, requires a large amount of time series data to build a purely data-driven representation of the underlying dynamical system. This attribute can be especially limiting in biological scenarios for which data are collected at sparse time points. Bayesian inference methods have been widely applied in modeling of biological systems with this level of data [33]. However, Bayesian inference relies on fitting parameters for a mechanistic model, and therefore is also subject to the previously discussed modeling related issues.

Here we describe a hybrid implementation of SSR and Bayesian inference methodologies in which we reduce model dimensionality by systematically dropping out system variables and replace them with either data or SSR predictions. To validate our methodology, we used a real biological data set consisting of 21 time series of cannibalistic flour beetle (*Tribolium castaneum*) population dynamics [11]. We previously found that combining SSR with mechanistic models enabled more accurate predictions of chaotic systems, including the flour beetle data set. Our goal in this work is to provide a hybrid methodology that produces uncertainty quantification, both for the model predictions and estimated parameters, in addition to forecasts of future time series data. We also provide a deeper investigation of the hybrid approach than in our previous efforts by analyzing uncertainty quantification results. We discuss our analysis and suggest possible reasons why the hybrid approach may yield more accurate predictions.

## 2 Data and Methods

### 2.1 Data

We use longitudinal data of total counts for larvae, pupae, and adults in flour beetle populations. The data came from 7 different experimental conditions in which adult mortality rates were altered, resulting in non-equilibrium dynamics; 3 replicates were performed in each condition for a total of 21 data sets [11]. Data were sampled every other week over an 82 week period for a total of 41 data points per time series. To test our methodology under noisy observation conditions similar to ecological systems, we added normally distributed random observation error to each time series using a coefficient of variation (CV) of 0.2, which is consistent with reported noise levels in survey data [14, 26]. The data from one experiment, shown in Figure 5 as black x’s, exemplify the typical non-equilibrium time series behavior of the beetle system. As denoted

by the vertical dashed line in Figures 5 and 6, each time series is divided into a training set (first 32 time points) and a testing set (last 9 time points). The training set is used for Bayesian inference or SSR and the testing set is used to evaluate the accuracy of the considered models.

## 2.2 Mathematical model

We use the previously validated discrete-time age-structured model

$$L(t) = bA(t-1)e^{-c_{el}L(t-1)-c_{ea}A(t-1)}, \quad (1)$$

$$P(t) = L(t-1)(1-\mu_1), \quad (2)$$

$$A(t) = P(t-1)e^{-c_{pa}A(t-1)} + A(t-1)(1-\mu_a), \quad (3)$$

for flour beetle population dynamics. The total number of larvae, pupae, and adults at time  $t$ , are given by  $L(t)$ ,  $P(t)$ , and  $A(t)$ , respectively. One unit of time is equal to 2 weeks, which matches the time scale of the data. This model quantifies the stage progression of beetles from the larval to pupae stage, and pupae to adult stage. Adult larvae are reproductive and contribute to the recruitment rate in equation (1). The exponential terms in equations (1) and (3), respectively, represent cannibalization of larvae by adults or larvae, and cannibalization of pupae by adults. A more thorough description of the model and parameters can be found in [11]. For reference below, we note that the parameters  $c_{pa}$  and  $\mu_a$  are assumed to be experimentally known; see [11] for further details.

## 2.3 State Space Reconstruction

We used state space reconstruction techniques based on Takens' theorem on delayed embedding to generate a non-mechanistic model prediction of the future system state [7, 13, 17, 19, 21, 22, 28–32, 34, 35, 37, 40]. We summarize the SSR technique here, and refer the reader to the supplemental of [36] for a more in depth description of the practical methodology for SSR. Let the  $i$ -th state variable of the system at time  $t$  be denoted by  $Y_i(t)$ . The method of delayed embedding starts by building a delayed coordinate vector using the observations  $\{Y_i(t)\}_{t=0}^N$ . Using  $L$  delays and a time lag value of  $\tau$ , this vector is given by  $Y_i^L(t) = [Y_i(t), Y_i(t-\tau), Y_i(t-2\tau), \dots, Y_i(t-d\tau)]$ . Here, we let  $\tau = 1$ , which corresponds to the sample rate of the observed data and  $d = 1$  delays. The library of delay vectors based on the training data up to  $N$  time points, which we call  $\Omega$ , can be used to form a prediction of the state at time  $N+P$ ; i.e.,  $Y_i(N+P)$ . We used the method of direct prediction in which the  $K$  nearest neighbors to  $Y_i^L(N)$  in  $\Omega$  are used as a sample space for predictions. We note that the direct prediction method here uses Takens' theorem which, under suitable assumptions, shows that a one-to-one mapping exists between the manifold given by the set of delayed coordinate vectors  $\{Y_i^L(t)\}$ , the “reconstructed attractor”, and the attractor manifold of the multivariate system that generated the time series  $\{Y_i(t)\}$ . Here, we adopt the methodology in [39] for computing the variance of the SSR prediction. If we denote the nearest neighbor sample space as  $\{Y_i^L(G(N, j) + P)\}_{j=1}^r$ , where  $G(N, j)$  denotes the indices for the nearest neighbors of  $Y_i^L(N)$  in  $\Omega$ , then the prediction for  $Y_i(N+P)$  is made by computing a weighted average over the nearest neighbors, given by  $\hat{Y}_i(N+P) = \frac{\sum_{j=1}^r w_j(N)Y_i^L(G(N, j) + P)}{\sum_{j=1}^r w_j(N)}$ . The weights  $w_j(N)$  are used to describe the probability of the  $j$ th element  $Y_i^L(G(N, j) + P)$  in the sample space  $\Omega$ ; i.e., the probability of selecting the  $j$ th element is given by  $p_j(N) = \frac{w_j(N)}{\sum_{k=1}^r w_k(N)}$ , where we assume that the variance of the prediction is given by  $\text{Var}(\hat{Y}_i(N+P)) = \mathbf{E}[(Y_i^L(G(N, j) + P) - \hat{Y}_i(N+P))^2]$ .

## 2.4 Bayesian inference

We performed Bayesian inference using a delayed rejection adaptive metropolis (DRAM) algorithm implemented in MATLAB [16]. Parameter values to initialize the parameter chains were generated using a weighted least squares algorithm (see Section 3.2.3 of [5]) for each data set and model; i.e., full model or hybrid model, separately. We used the following lower and upper bounds for each parameter: the initial conditions are given by  $L_0, A_0 \in [-50, 550]$ ,  $P_0 \in [-200, 500]$  and the model parameters are  $b \in [-5, 30]$ ,  $c_{el}, c_{ea} \in [-0.02, 1]$ , and  $\mu_1 \in [-1, 1]$ . These bounds were set by initially choosing an interval +/-50% around previously estimated parameter values from [12] and then increasing the size of the interval until

the tails of each parameter’s posterior distribution was contained completely within the interval. We used noninformative flat prior distributions defined between the upper and lower bounds for each parameter. We set the chain length to 20,000 with a burn-in length of 20,000. We performed uncertainty quantification; i.e., computation of 95% prediction and credible intervals by sampling from posterior distributions as described in [16, 33].

## 2.5 Hybrid methodology

Our approach for combining SSR and Bayesian inference predictions is to use a partial model for the Bayesian inference methodology by dropping one or more of the variables used in prediction. For example, a hybrid model can be constructed for predicting the  $A$  variable by using a partial model consisting only of equation (3) and either training data or SSR predictions for the remaining  $L$  and  $P$  variables as depicted in Figure 1. We describe the procedure for generating the hybrid prediction model illustrated in Figure 1 and note that the procedure for generating other hybrid models is similar.

An example of a hybrid model uses (3) to describe the  $A$  variable. Since (3) contains terms involving the variable  $P(t)$  describing the number of pupae, we simply replace this variable by the actual number of pupae observed at time  $t$  up to the last training time point at  $t = 32$ . This enables (3) to be used with Bayesian inference to estimate parameters for the  $A$  equation. In this case the only parameter to estimate is the initial condition  $A_0$ , since  $c_{pa}$  and  $\mu_a$  are experimentally known. Simultaneously, SSR is applied to the training data for  $L$  and  $P$  to generate a prediction of those variables after the training data ends. Once the parameter  $A_0$  is estimated for (3), we can generate a prediction with this parameterized equation by continuing to substitute the SSR prediction for  $P$  for the missing time series  $P(t)$  beyond the training data.

Using this approach, there are a total of 7 models corresponding to subsets of the variables for the full model  $\{L, P, A\}$ . For example, we may choose to use equations (1) and (2) to model the  $L$  and  $P$  variables, and then use either training data or SSR predictions as a substitution for the  $A(t)$  time series.

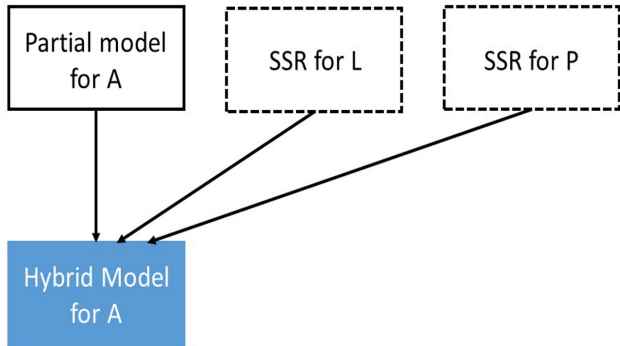


Figure 1: Illustration of the hybrid technique applied to modeling the  $A$  variable in the flour beetle system given by equations (1)-(3). The hybrid model shown here uses a partial model for  $A$ ; i.e., only equation (3), and training data for  $L$  and  $P$  to estimate parameters in the partial model. In this case, the only parameter to estimate for Eq. (3) is the initial condition  $A(0)$ . SSR predictions are used to continue substituting time series for  $L$  and  $P$  after the last time point in the training data.

## 2.6 Hybrid uncertainty quantification

Our methodology for uncertainty quantification in hybrid models is to make a modification to the input of the forward solution when computing prediction and credible intervals with the DRAM algorithm [16]. We will illustrate the method on the hybrid model using an equation for the  $A$  variable and data/SSR for the  $L$  and  $P$  variables, as illustrated in Figure 1, and note that computations are similar for other hybrid model choices. We note that, in this scenario, a posterior distribution is obtained for the parameter  $A_0$ , and the equation  $A(t) = P(t - 1)e^{-c_{pa}A(t-1)} + A(t - 1)(1 - \mu_a)$  is used to model the  $A$  variable. The method for computing prediction and credible intervals for times  $t \leq 32$  is unchanged from the DRAM algorithm outlined in [16], since we assume that the replacement of  $P(t)$  with training data at times  $t \leq 32$  are an exact model for the pupae population. At times  $t \geq 33$ , the trajectory of the forward solution for  $A$  is

affected by the uncertainty in the SSR prediction for  $P$ . Adopting the notation from Section 2.3, the SSR uncertainty is given by the probability distribution generated by the sample space of nearest neighbors  $\Omega$ , where the trajectory of each neighbor is selected with probability  $p_j$ . Thus, prediction and credible intervals are obtained for the hybrid model for  $t \geq 33$  by sampling from the joint density of the posterior distribution for  $A_0$  and the sample space of SSR nearest neighbor trajectories.

## 2.7 Prediction accuracy

The standardized root mean square error (SRMSE) is used to quantify prediction accuracy for the testing data; e.g., the last 9 time points in Figures 5 and 6). The SRMSE quantifies the mean of the squared error over all the training data, and thus represents an accuracy score at each of the 9 time points predicted, aggregated over the entire 21 time series in the experimental data set. We note that the SRMSE normalizes the prediction score with respect to the standard deviation of the training data. Thus,  $\text{SRMSE} < 1$  implies that the prediction is better than using the mean of the training data, “naive prediction”, to forecast future time series. It is expected that, in the long-term, all forecasts will eventually converge to  $\text{SRMSE} = 1$ .

## 3 Results

We performed parameter estimation and uncertainty quantification for the full model in equations (1)-(3) and every hybrid model option corresponding to each subset of the variables  $\{L, P, A\}$ . For clarity, we will refer to the full model, which does not use SSR, as “choice 7”. The hybrid models will be referred to as choices 1-6. For example, choice 1 corresponds to modeling the  $A$  variable with Eq. (3) and using data or SSR predictions for the  $L$  and  $P$  variables. Table 1 indicates the model used for predicting  $L$ ,  $P$ , and  $A$  (either data/SSR or one of equations (1)-(3)), corresponding to choices 1 through 7.

Choice	$L$	$P$	$A$	Parameters
1	data/SSR	data/SSR	Eq. (3)	$A_0$
2	data/SSR	Eq. (2)	data/SSR	$P_0, \mu_1$
3	data/SSR	Eq. (2)	Eq. (3)	$P_0, A_0, \mu_1$
4	Eq. (1)	data/SSR	data/SSR	$L_0, b, c_{el}, c_{ea}$
5	Eq. (1)	data/SSR	Eq. (3)	$L_0, A_0, b, c_{el}, c_{ea}$
6	Eq. (1)	Eq. (2)	data/SSR	$L_0, P_0, b, c_{el}, c_{ea}, \mu_1$
7	Eq. (1)	Eq. (2)	Eq. (3)	$L_0, P_0, A_0, b, c_{el}, c_{ea}, \mu_1$

Table 1: Model used to predict the  $L$ ,  $P$ , and  $A$  variables for choices 1-7. Either SSR or one of equations (1)-(3) if used for each variable. Choice 7 corresponds to the full LPA model, while choices 1-6 represent one of 6 possible hybrid models. The corresponding estimated parameters are also indicated for each hybrid model.

### 3.1 Forecast accuracy

We evaluated the forecast accuracy of the full model (choice 7), each hybrid model (choices 1-6), and SSR for each of the  $L$ ,  $P$ , and  $A$  variables using the SMRSE to quantify accuracy as illustrated in Figures 2 - 4. We found that the full model outperformed the SSR method for the  $L$  variable for up to 6 time points (12 weeks) of prediction; Figure 2. Comparisons between the full model and the SSR method were less clear for the  $P$  and  $A$  variables. For example, in the short term, at the first time point of prediction, SSR outperformed the full model for the  $P$  variable, but the full model outperformed SSR at later time points; e.g. forecast horizon at time points 3-6 as shown in Figure 3. We observed that the hybrid method, corresponding to at least one of the hybrid models from choices 1-6, was able to outperform both the full model and the SSR method for

each of the  $L$ ,  $P$ , and  $A$  variables in the first 2-3 time points of prediction and was comparable to the full model in subsequent time points of the forecast horizon. These results are similar to what we previously observed when combining SSR methods with weighted least squares techniques for parameter estimation. Focusing on the comparison between the hybrid models and the SSR method, we found that hybrid models were able to outperform SSR and stay below a mean SRMSE of 0.8 up to 5 time points in the forecast horizon (10 weeks of prediction). These results indicate that the hybrid models are the most accurate choice for predicting future time series as compared to the full model or SSR alone.

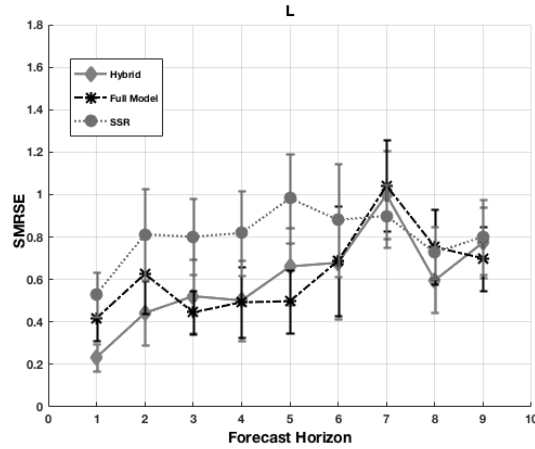


Figure 2: Forecast errors (SRMSE) for each forecasting method for the larvae population data. Points are the mean SRMSE over 21 data sets; bars are the standard errors. 1 time step = 2 weeks in the forecast horizon. The solid, dotted, and dashed lines are the forecast error for the hybrid model (choice 5), full model (choice 7), and SSR predictions, respectively.

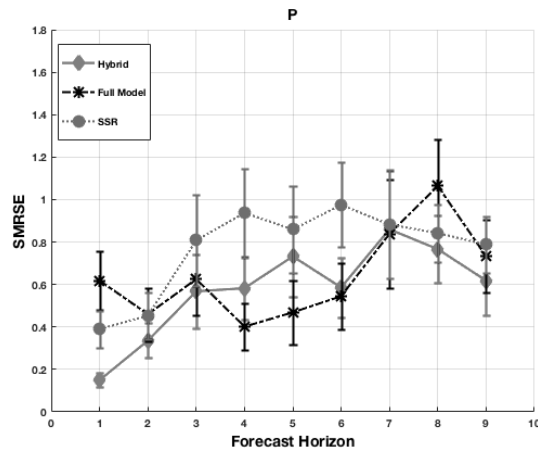


Figure 3: Forecast errors (SRMSE) for each forecasting method for the pupae population data. Points are the mean SRMSE over 21 data sets; bars are the standard errors. 1 time step = 2 weeks in the forecast horizon. The solid, dotted, and dashed are the forecast error for the hybrid model (choices 2 and 3), full model (choice 7), and SSR predictions, respectively.

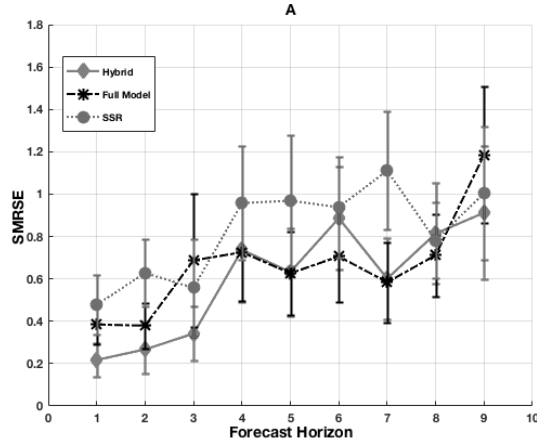


Figure 4: Forecast errors mean (SRMSE) for each forecasting method for the adult population data. Points are the SRMSE over 21 data sets; bars are the standard errors. 1 time step = 2 weeks in the forecast horizon. The solid, dotted, and dashed lines are the forecast error for the hybrid model (choice 3), full model (choice 7), and SSR predictions, respectively.

### 3.2 Uncertainty quantification

Here we illustrate the methodology for uncertainty quantification with hybrid models on the flour beetle data set and model. Figure 5 illustrates the full model fit to experimental data for one time series; only the  $A$  variable is shown in order to compare to one of the hybrid model that does not use the  $L$  or  $P$  variables. This instance is representative of a scenario in which the full model prediction is approximately equivalent to the mean of the training data, and thus performs no better than the naive prediction. In contrast, Figure 6 shows the hybrid model for the same experimental data; SSR predictions are also shown for comparison. In addition to the error between the testing data interval being lower, the hybrid model model also has narrower 95% credible intervals than the full model, indicating higher confidence in the predicted population size at any given time point.

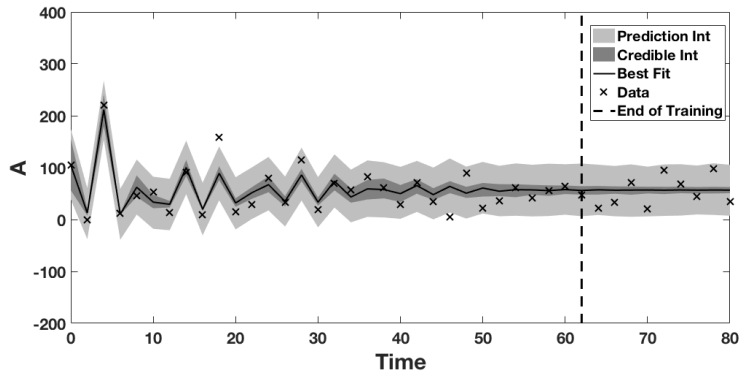


Figure 5: Full model prediction with uncertainty quantification for the  $A$  variable. Data (black x's) are from experiment 1 for which  $c_{pa}$  was experimentally set to zero. The vertical dashed line separates the training data used for parameter estimation from the testing data used for evaluating forecast accuracy. The 95% credible and prediction intervals are shown as dark and light grey, respectively. The black line represents the mean of the credible interval (best fit).

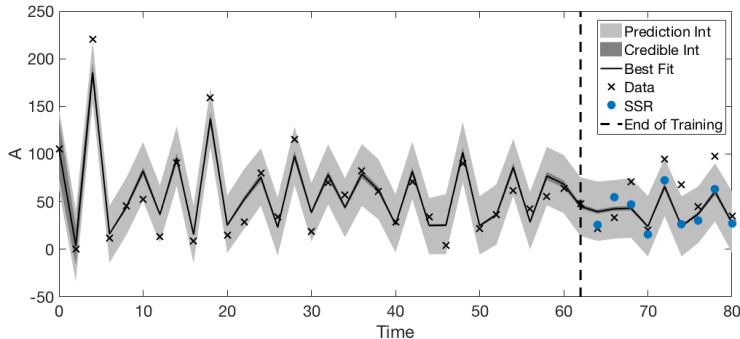


Figure 6: Hybrid model prediction (choice 1) with uncertainty quantification for the  $A$  variable. Data (black  $x$ 's) are from experiment 1 for which  $c_{pa}$  was experimentally set to zero. The vertical dashed line separates the training data used for parameter estimation from the testing data used for evaluating forecast accuracy. The 95% credible and prediction intervals are shown as dark and light grey, respectively. The black line represents the mean of the credible interval (best fit).

### 3.3 Practical identifiability analysis

The results from Bayesian inference on a model and time series data can be used to produce other outputs besides credible and prediction intervals. Two outputs in particular, parameter correlation plots and the fisher information matrix (FIM), provide key information about the practical identifiability of parameters in the model; i.e., the ability to estimate parameters with reasonably low uncertainty levels from the available data in the presence of noise. Thereby, analysis of correlation plots and the FIM provides insight into difficulties encountered in the parameter estimation task. We analyzed parameter correlations and the FIM for the full model and hybrid models to test the hypothesis that hybrid models simplify the parameter estimation task by maximizing the level of data information content with respect to the set of estimated parameters. For example, even though estimation for the full model (choice 7) uses 3 time series, corresponding to the  $L$ ,  $P$  and  $A$  variables, and the estimation for the hybrid choice 1 only uses a single time series, corresponding to the  $A$  variable, it is not immediately obvious that the gain in the amount of data afforded by using the full model justifies the additional number of parameters that need to be estimated. Moreover, it is unclear how the presence of 20% observation noise or the structure of the model may independently or synergistically affect the balance between model dimension and the amount of data used for parameter estimation.

In Figures 7 and 8 we illustrate the parameter correlation information that is output from applying Bayesian inference to the full model and one of the hybrid models on one of the 21 experiments in the data set. Briefly, parameter correlation information can be seen graphically by plotting the values obtained in the MCMC chain used to build the joint posterior distribution of the estimated parameters. For example, if there are  $p$  parameters to be estimated, then a  $p$ -dimensional vector  $(\hat{\theta}_{1,k}, \dots, \hat{\theta}_{p,k})$  of estimated parameters is obtained at step  $k$  in the chain. If the number of chain iterations used to construct the posterior distribution is equal to  $M$ , then a parameter pair correlation plot for the  $i$ -th vs.  $j$ -th parameter is made by plotting  $M$  pairs of points  $(\hat{\theta}_{i,k}, \hat{\theta}_{j,k})$  for  $k = 1, \dots, M$ . Figure 7 shows that correlations exist when using the full model between the pairs  $(A_0, P_0)$ ,  $(b, P_0)$ ,  $(b, L_0)$ ,  $(c_{el}, L_0)$ ,  $(c_{el}, b)$ ,  $(c_{ea}, P_0)$ ,  $(c_{el}, b)$ ,  $(\mu_1, L_0)$ ,  $(\mu_1, b)$ , and  $(\mu_1, c_{el})$ . In contrast, Figure 8 shows that no correlations exist for the hybrid model among the three possible parameter pairs when using choice 3; i.e., data/model for the  $L$  variable, and (2) and (3) for the  $P$  and  $A$  variables, respectively.

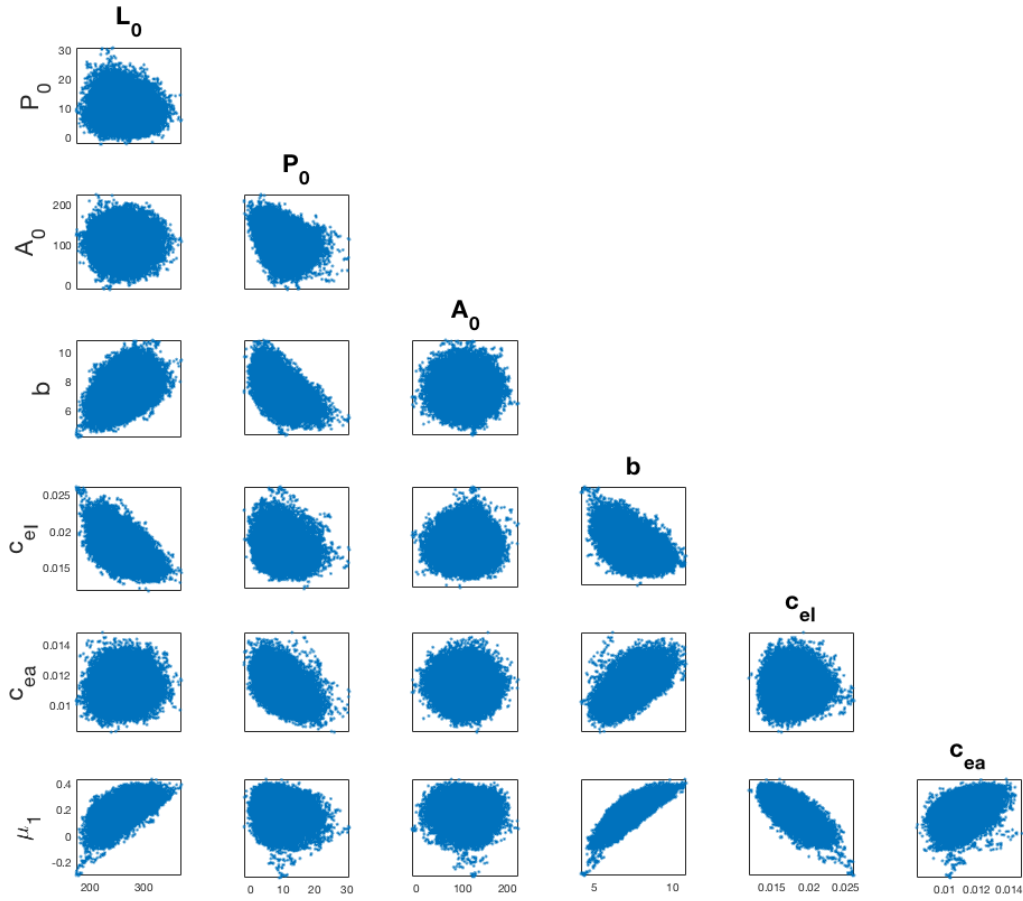


Figure 7: Parameter correlation plots for the full model (choice 7). The data from experiment 1 were used for parameter estimation.

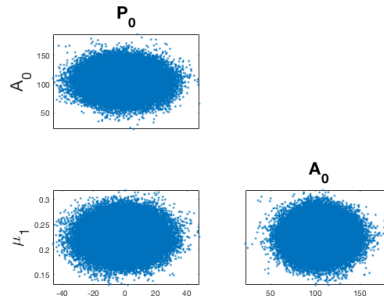


Figure 8: Parameter correlation plots for a hybrid model (choice 3). The data from experiment 1 were used for parameter estimation.

To investigate the presence of parameter correlations among all 21 time series and all model choices, we computed linear the correlation coefficients ( $\rho$ ) between all parameter pairs for any given model as plotted in Figure 9. We found that, with the exception of one pair ( $b$  vs.  $c_{ea}$ ), on average the correlation coefficients were lower for any hybrid model (choices 1-6) compared to the full model (choice 7).

Given the presence of parameter correlations, we next analyzed the rank deficiency of the fisher infor-

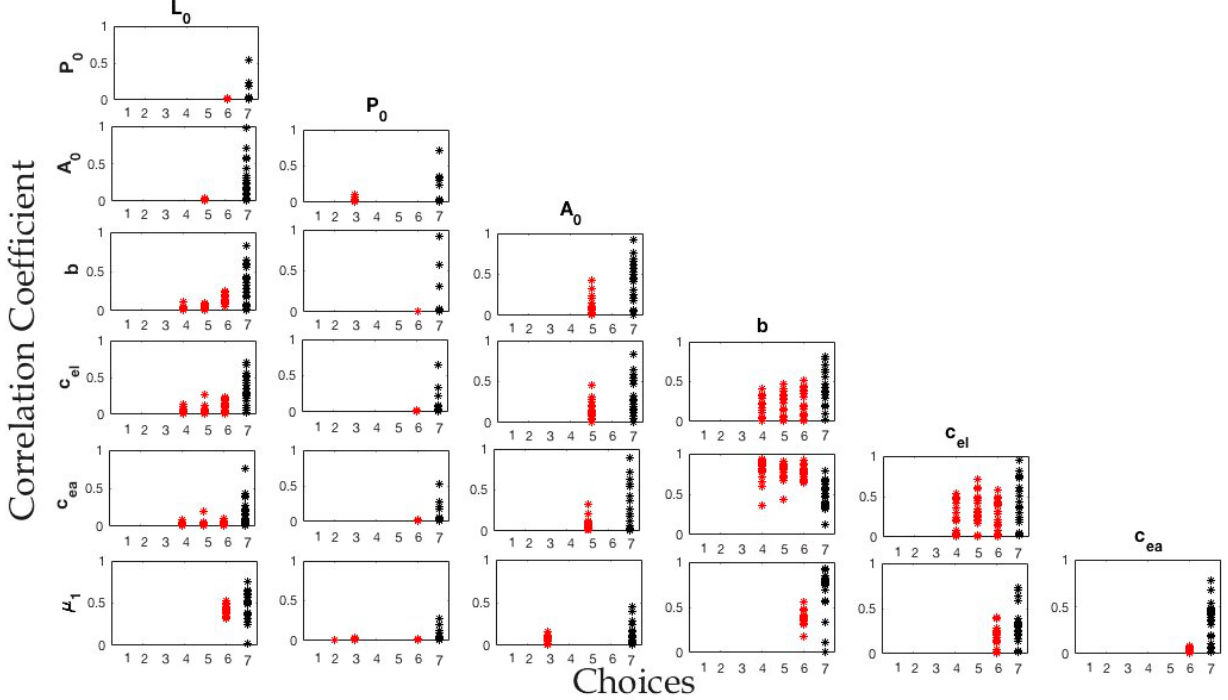


Figure 9: Correlation coefficients among all pairs of estimated parameters. Computations were performed for each model choice where choice 7 corresponds to the full model and choices 1-6 correspond to hybrid models. Coefficients were computed only for pairs of parameters that existed among the hybrid models. Each star within each subplot represents a correlation coefficient for a single time series, with 21 total possible time series.

mation matrix (FIM). The FIM has been previously used in subset selection algorithms that seek to predict which subsets of parameters are identifiable for a given model and available set of data. Importantly, these FIM based methods take into account the sensitivity of the model output with respect to parameters and combines this information with the effect of parameter correlations. For example, insensitive parameters are more difficult to identify from data since a large change in the parameter doesn't affect the model output, and in particular doesn't affect how well the model fits the data. If the number of estimated parameters is given by  $p$ , then the FIM is a  $p \times p$  matrix, and the rank of the FIM can be used to estimate the number of parameters that are practically identifiable [24]. The FIM is equal to  $\sum_{i=1}^N \chi^T(t_i)\chi(t_i)$ , where the matrix  $\chi(t_i)$  contains sensitivities of the model with respect to parameters at time point  $t_i$  in the training data. The  $k, j$ -th entry of  $\chi(t_i)$  is given by  $\{\frac{\partial y_k(t_i)}{\partial \theta_j}\}$  where  $y_k$  is the  $k$ -th observable; e.g.,  $L$ ,  $P$ , or  $A$ , and  $\theta_j$  is the  $j$ -th model parameter. We used the parameters estimated from bayesian inference to compute the FIM for each of the 21 time series and model choices 1-7.

We determined the rank of the FIM by first computing the singular value decomposition (SVD); i.e.,  $\text{FIM} = USV^T$ , where  $S$  is a diagonal matrix of the singular values of the FIM listed in decreasing order, and  $U$  and  $V$  are orthogonal matrices containing left and right singular vectors. Since  $S$  is a diagonal matrix, it can be viewed as a list  $\{s_1, \dots, s_p\}$ . We used the location in the list, if any, where the ratio  $\frac{s_m}{s_{m+1}} > 10^{10}$  to indicate that the rank of the FIM was equal to  $m$ . Thus, the rank deficiency is given by  $p - m$ , and the number of parameters that are not practically identifiable increases as a function rank deficiency. We found that the full model (choice 7) had the largest average rank deficiency over all the 21 time series compared to any of the hybrid models. Among the hybrid models, choice 6 had more rank deficiency than choices 4 and 5, and each of choices 4 and 5 had more rank deficiency than choices 1, 2, and 3. Choices 1 and 2 had no rank deficiency.

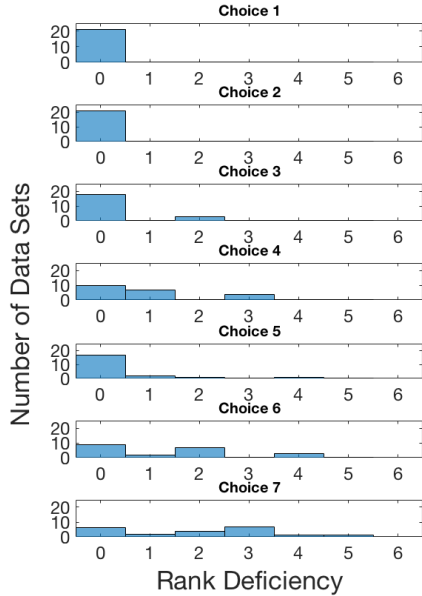


Figure 10: Rank deficiency for each model choice. The full model = choice 7, hybrid models = choices 1-6. The y-axis indicates the number of time series (out of 21 total) that a particular choice resulted in a certain level of rank deficiency given by the x-axis.

## 4 Discussion and future work

We illustrated a hybrid methodology that may be used for minimizing the dimensionality of parameter inference from time series data while also maximizing predictive accuracy of the resulting parameterized model that is broadly applicable to multivariate systems. Since this balance is often the goal when developing predictive models of biological systems for which data are typically sparse by having either low frequency or few time points relative to the number of estimated parameters, we hypothesize that the methodology illustrated here may enable the application of the mechanistic modeling paradigm to a wider range of biological scenarios for which data limitations or high dimensionality of the system inhibits accurate parameterization. An important feature of the combination of Bayesian inference with SSR is that uncertainty quantification; i.e., the computation of 95% prediction and credible intervals, is readily obtained by jointly sampling from the posterior distribution of the estimated parameters and the SSR sample space defined by the nearest neighbor prediction. In practice, the ability to ascertain uncertainty in predictions is a necessary feature for applying the hybrid methodology to real world scenarios. For example, it would be clearer to make an ecological management decision based on population densities forecasted by a hybrid model if one can also associate a level of confidence with those predictions.

In addition to illustrating uncertainty quantification for hybrid models applied to a real biological data set, our intent in combining Bayesian inference with SSR in this work was to also use the results from Bayesian inference to investigate the mechanisms by which the hybrid technique may alleviate some of the practical identifiability issues that commonly arise in difficult parameter estimation problems. We analyzed the correlation plots and found that for every pair of parameters, with the exception of  $b$  vs.  $c_{ea}$ , the average correlation among all 21 time series was lower for any choice of hybrid model compared to the full model. We note that both  $b$  and  $c_{ea}$  are parameters located in the equation for the  $L$  variable, see (1), suggesting that estimating parameters for this variable may be the source of difficulty in general for the flour beetle system. Under this hypothesis it is expected that hybrid models containing the  $L$  variable will have the same or increased level of parameter correlations as the full model, since the hybrid models will use less data for bayesian inference by removing either or both of the equations and corresponding time series for  $P$  and  $A$ .

Our analysis of the rank deficiency of the fisher information matrix (FIM) corroborates the finding that the equation for the  $L$  variable is problematic for parameter estimation. For clarity, we show the different model choices and the corresponding parameters contained within each of them in Figure 11. The equation

for the  $L$  variable contains four parameters, while the  $P$  variable contains two parameters, and the  $A$  variable only contains one parameter. The rank deficiency histograms in Figure 10 indicate that the source of parameter non-identifiability stems from the  $L$  variable. For example, the only hybrid model choice with one modeled variable (choices 1, 2, or 4) that has any rank deficiency is choice 4, which corresponds in partly to using an equation for the  $L$  variable and data/SSR for the  $P$  and  $A$  variables. This deficiency can be mitigated by including the equation for the  $A$  variable in the hybrid model, since it only contains one parameter, and we see that choice 5 ( $L$  and  $A$  variables modeled with equations) has lower average rank deficiency among the entire experimental data set than choice 7. We observed that the same mitigating effect is not present for choice 6 ( $L$  and  $P$  variables modeled with equations). These findings are in agreement with the forecast accuracy results in Figures 2, 3, and 4; the hybrid models choices used in these plots are those with the best forecast accuracy and these choices also have the lowest rank deficiency. Together with our parameter correlation analysis, these results suggest that one likely mechanism by which hybrid modeling increases forecast accuracy is by eliminating variables for which parameters may not be practically identifiable from the available data and replacing their inaccurate estimation with non-mechanistic model based forecasting.

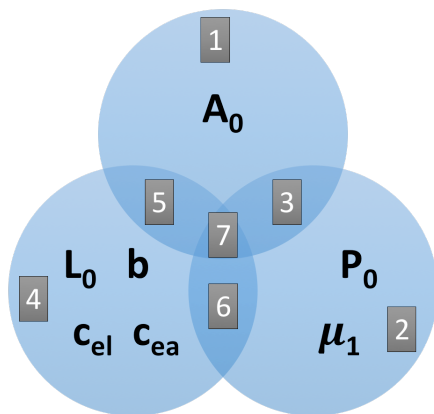


Figure 11: Venn diagram illustration of the model parameters estimated for each choice. Grey boxes contain the model choice number, where choice 7 corresponds the full model and choices 1-6 represent the hybrid models.

In future work, we will consider other non-mechanistic models as a substitution for state space reconstruction (SSR) methods. We outlined a general strategy for circumventing mechanistic modeling parameter estimation issues; however, the reliance of this strategy on SSR remains to be tested. Machine learning methods, such as neural networks, have successfully been used in forecasting longitudinal data within a multistep ahead prediction framework [8, 25]. Autoregressive models are also commonly used in statistical forecasting from longitudinal data and has some similarities to SSR; e.g., predictions of the future state are based on a non-mechanistic model of the recent history [6]. Additionally, completely non-mechanistic yet semi-parametric hybrid approaches may be considered by substituting an autoregressive or machine learning model for the mechanistic model component in the hybrid and combining these methods with SSR.

## 5 Acknowledgements

The research was partially supported by grants RTG/DMS-1246991 and DMS-1514929 from the National Science Foundation.

## References

- [1] Mathematical and Experimental Modeling of Physical and Biological Processes, January 2009.
- [2] Kaska Adoteye, H. T. Banks, and Kevin B. Flores. Optimal design of non-equilibrium experiments for genetic network interrogation. *Applied Mathematics Letters*, 40:84–89, February 2015.

- [3] H. T. Banks, J. E. Banks, Kathryn Link, J. A. Rosenheim, Chelsea Ross, and K. A. Tillman. Model comparison tests to determine data information content. Applied Mathematics Letters, 43:10–18, May 2015.
- [4] H. T. Banks, Robert Baraldi, Karissa Cross, Kevin Flores, Christina McChesney, Laura Poag, and Emma Thorpe. Uncertainty quantification in modeling HIV viral mechanics. Mathematical biosciences and engineering: MBE, 12(5):937–964, October 2015.
- [5] H. T. Banks, Shuhua Hu, and W. Clayton Thompson. Modeling and Inverse Problems in the Presence of Uncertainty. CRC Press, April 2014.
- [6] S. A. Billings. Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains. John Wiley & Sons, September 2013.
- [7] Martin Casdagli. Nonlinear prediction of chaotic time series. Physica D: Nonlinear Phenomena, 35(3):335–356, 1989.
- [8] Haibin Cheng, Pang Ning Tan, Jing Gao, and Jerry Scripps. Multistep-ahead time series prediction. Lecture Notes in Computer Science: Advances in Knowledge Discovery and Data Mining, 3918(765-774), 2006.
- [9] A. Cintrón-Arias, H. T. Banks, A. Capaldi, and A. L. Lloyd. A sensitivity matrix based methodology for inverse problem formulation. Journal of Inverse and Ill-posed Problems, 17(6):545–564, 2009.
- [10] C. Cobelli and J. J. DiStefano. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. American Journal of Physiology - Regulatory, Integrative and Comparative Physiology, 239(1):R7–R24, July 1980.
- [11] R. F. Constantino, R. A. Desharnais, J. M. Cushing, and B. Dennis. Chaotic dynamics in an insect population. Science, 276:1881–1882, 1997.
- [12] Brian Dennis, Robert A. Desharnais, J. M. Cushing, and R. F. Costantino. Nonlinear Demographic Dynamics: Mathematical Models, Statistical Methods, and Biological Experiments. Ecological Monographs, 65(3):261–282, February 1995.
- [13] J Farmer and J Sidorowich. Predicting chaotic time series. Phys. Rev. Lett., 59:845–848, 1987.
- [14] Chris R. I. C. Francis, Rosemary J. Hurst, and James A. Renwick. Quantifying annual variation in catchability for commercial and research fishing. Fishery Bulletin, 101(2):293–304, 2003.
- [15] Seymour Geisser. Predictive Inference. CRC Press, June 1993. Google-Books-ID: wfdlBZ\_iwZoC.
- [16] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. Statistics and Computing, 16(4):339–354.
- [17] F. Hamilton, T. Berry, and T. Sauer. Ensemble kalman filtering without a model. Physical Review X, 6:011021, 2016.
- [18] Florian Hartig and Carsten F. Dormann. Does model-free forecasting really outperform the true model? Proceedings of the National Academy of Sciences, 110(42):E3975–E3975, October 2013.
- [19] Chih-Hao Hsieh, Sarah M Glaser, Andrew J Lucas, and George Sugihara. Distinguishing random environmental fluctuations from ecological catastrophes for the north pacific ocean. Nature, 435(7040):336–340, 2005.
- [20] Joseph DiStefano III. Dynamic Systems Biology Modeling and Simulation. Academic Press, January 2015. Google-Books-ID: nWoYAgAAQBAJ.
- [21] J Jimenez, JA Moreno, and GJ Ruggeri. Forecasting on chaotic time series: A local optimal linear-reconstruction method. Phys. Rev. A, 45(6):3553, 1992.

- [22] D Kugiumtzis, OC Lingjærde, and N Christophersen. Regularized local linear prediction of chaotic time series. Physica D: Nonlinear Phenomena, 112(3):344–360, 1998.
- [23] Nicolette Meshkat, Christine Er-zhen Kuo, and Joseph DiStefano Iii. On Finding and Using Identifiable Parameter Combinations in Nonlinear Dynamic Systems Biology Models and COMBOS: A Novel Web Implementation. PLOS ONE, 9(10):e110261, October 2014.
- [24] Mette S. Olufsen and Johnny T. Ottesen. A practical approach to parameter estimation applied to model predicting heart rate regulation. Journal of Mathematical Biology, 67(1):39–68, July 2013.
- [25] A. G. Parlos, O. T. Rais, and A. F. Atiya. Multi-step-ahead prediction using dynamic recurrent neural networks. Neural Networks, 13(7):765–786, September 2000.
- [26] Charles Perretti, Stephan Munch, and George Sugihara. Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. Proceedings of the National Academy of Sciences, 110:5253–5257, 2013.
- [27] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics, 25(15):1923–1929, August 2009.
- [28] S Regonda, B Rajagopalan, U Lall, M Clark, and Y-I Moon. Local polynomial method for ensemble forecast of time series. Nonlin. Proc. in Geophys., 12:397–406, 2005.
- [29] T Sauer. Time series prediction by using delay coordinate embedding. In Time Series Prediction: Forecasting the Future and Understanding the Past, pages 175–193. Addison Wesley, 1994.
- [30] Björn Schelter, Matthias Winterhalder, and Jens Timmer. Handbook of time series analysis: recent theoretical developments and applications. John Wiley and Sons, 2006.
- [31] Christian G Schroer, Tim Sauer, Edward Ott, and James A Yorke. Predicting chaos most of the time from embeddings with self-intersections. Phys. Rev. Lett., 80(7):1410, 1998.
- [32] Leonard A Smith. Identification and prediction of low dimensional dynamics. Physica D: Nonlinear Phenomena, 58(1):50–76, 1992.
- [33] Ralph C Smith. Uncertainty quantification: theory, implementation, and applications. SIAM, Philadelphia, 2014. OCLC: 875327904.
- [34] Christopher C Strelhoff and Alfred W Hübler. Medium-term prediction of chaos. Phys. Rev. Lett., 96(4):044101, 2006.
- [35] George Sugihara. Nonlinear forecasting for the classification of natural time series. Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences, 348(1688):477–495, 1994.
- [36] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting Causality in Complex Ecosystems. Science, 338(6106):496–500, October 2012.
- [37] George Sugihara and Robert M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature, 344(6268):734–741, 04 1990.
- [38] H Voss, J Timmer, and J Kurths. Nonlinear dynamical system identification from uncertain and indirect measurements. Int J Bif Chaos, 14:1905–1924, 2002.
- [39] Hao Ye, Richard J. Beamish, Sarah M. Glaser, Sue C. H. Grant, Chih-hao Hsieh, Laura J. Richards, Jon T. Schnute, and George Sugihara. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. Proceedings of the National Academy of Sciences, 112(13):E1569–E1576, March 2015.
- [40] G Yuan, M Lozier, L Pratt, C Jones, and K Helfrich. Estimating the predicability of an oceanic time series using linear and nonlinear methods. J. Geophys. Res., 109:C08002, 2004.