LUT University

Parameter Estimation and Forecasting for Biased Models

Amadi Miracle, Haario Heikki

This is a Author's accepted manuscript (AAM) version of a publication

published by Springer

in Sriraman, B., Capistrán A., M. (eds) Handbook of Visual, Experimental and Computational Mathematics

DOI: 10.1007/978-3-030-93954-0_36-1

Copyright of the original publication:

© 2023 Springer Nature Switzerland AG

Please cite the publication as follows:

Amadi, M., Haario, H. (2023). Parameter Estimation and Forecasting for Biased Models. In: Sriraman, B., Capistrán A., M. (eds) Handbook of Visual, Experimental and Computational Mathematics. Springer, Cham. DOI: 10.1007/978-3-030-93954-0_36-1

This is a parallel published version of an original publication. This version can differ from the original published article.

Parameter Estimation and Forecasting for Biased Models

Miracle Amadi * and Heikki Haario

Abstract Dynamical models may be biased due to limitations in their ability to represent real-world phenomena. This is often a result of simplifications in the model structure. Such models can be used to answer relevant questions on a simulation basis, but when compared with real data, there is a need to find efficient ways to handle the model bias. Traditionally, parameters of dynamical systems are estimated by directly comparing model simulations to observed data utilizing a least squares approach. However, if the models are biased, this classical approach may not accurately estimate the true states. Here, a data-driven approach is proposed that enables the model to follow the trajectory of the data. More so, following this approach, the model parameters can be adjusted in the process. The simple Ross model for malaria is used as an example, as it does not capture the impact of external factors such as periodically changing weather patterns. A hybrid of state-augmentation and Kalman filter (KF) likelihood methods are applied to estimate both the dynamic and static parameters of the model. Data such as the monthly hospital cases from Nigeria, as well as rainfall in the same region, are utilized as a proxy in order to help the filter to produce an optimal estimate of the evolving state of the system. The idea is that the model bias translates to the parameters, in the sense that the right parameters can be estimated but with larger uncertainties. Markov Chain Monte Carlo (MCMC) is employed as a general tool to diagnose the identifiability of the static parameters. We outline some integral steps to handle model bias using KF likelihood. Examples for hospital planning are discussed: how to anticipate and prepare for potential increases in patient volume during and after periods of rainy seasons.

Miracle Amadi
Lappeenranta-Lahti University of Technology LUT, Yliopistonkatu 34, Lappeenranta, Finland e-mail: miracle.amadi@lut.fi

Heikki Haario
Lappeenranta-Lahti University of Technology LUT, Yliopistonkatu 34, Lappeenranta, Finland e-mail: heikki.haario@lut.fi

* corresponding author

Keywords

State augmentation, Kalman filter likelihood, MCMC Parameter identification, Dynamical models, Hospital planning.

Introduction

Model Bias

The nature of many dynamical models is such that they make overly simplistic assumptions, making it difficult to compare the model solutions to real data. While these models may be relevant in answering certain general questions on a simulation basis, they are often insufficient to effectively identify model parameters using real data. MCMC as a parameter identification diagnostics technique only takes noise into account, but not model bias. Model misspecification is often used with reference to different meanings in the literature Walker (2013); Grünwald and Van Ommen (2017). In (Spanos 2011), model misspecification was distinguished between statistical inadequacy (i.e., when true distribution is not included in the simulation) and substantive inadequacy (i.e., significant opportunities to improve the model). The model bias discussed in this work is the case where the model does not contain all the relevant features necessary to accurately describe the scenarios in the data. The central problem in parameter estimation is defining a proper likelihood function that specifies the plausibility of observed data given model parameter values. Traditionally, parameters of dynamical systems are estimated by comparing model simulations to observed data using a measure such as a sum of squared differences between the model values and observations, assuming that the observations are noisy realizations of the model values. However, applying these techniques directly to a biased model is problematic, as the model may not contain the necessary features needed to fit the real data, rendering direct comparisons of model simulations and observations meaningless unless a technique is used to make the model follow the data to some extent. This work has provided a data-driven approach to handling model bias in dynamic models by making the model follow the data.

Ross Model

The Ross model has played a central role in the development of research on mosquito-borne pathogen transmission and has had a major influence on the development of strategies for malaria control. The model presents the time evolution of the fraction of humans and mosquitoes in the infected classes (i_h, i_m) using two differential equations respectively.

$$\frac{di_h}{dt} = mabi_m(1-i_h) - i_h r \tag{1}$$
$$\frac{di_m}{dt} = aci_h(1-i_m) - \mu i_m,$$

where i_h and i_m denote the fractions of infected humans and mosquitoes, correspondingly, m stands for mosquito-to-human ratio, b and c are the probabilities of transmission during mosquito contact with the host, μ_m denotes the mosquito mortality rate, a is the contact rate and r stands for recovery rate for the humans.

Obviously, the Ross model does not include all the relevant features necessary to accurately describe real-life scenarios. For instance, the mosquito density parameter m is a constant in the model while in reality it is impacted, for instance, by weather conditions Asgarian et al (2021).

The approach

Classical parameter estimation approaches based on directly comparing model simulations and observations using, for example, the least squares approach, may be misleading in estimating the parameters of a model in such a situation. Instead, the model bias is handled using the Kalman filter approach.

Filtering methods can be used for state estimation in dynamical models, where the distribution of the model state evolves with time and is updated sequentially in time as new observations become available (Evensen et al 2009). This can be achieved by applying the Bayes formula sequentially, using the model prediction from the previous time point as the prior. The model is not assumed to be perfectly known, but the uncertainty of it. This is due to stochasticity or bias which is taken into account by a model covariance matrix.

In addition to the state, the model parameters can also be estimated. Two main approaches are available: state augmentation and the Kalman filter likelihood. In the likelihood approach, the probability of a parameter value is evaluated by running the filtering algorithm over a chosen time series data set, while keeping the parameter value fixed. The likelihood is constructed using the residuals between data and model predictions. The method resembles classical parameter estimation by residual sums of squares, but the uncertainty in the model state is "integrated out" using the state estimation technique. While the Kalman filter likelihood applies to the static model parameters, it does not take into account the model bias due to the dynamically changing features of data that are not described by the model. In the state augmentation approach, the time-dependent model parameters are treated as artificial states and assimilated alongside the actual model state (Kitagawa 1998; Ionides et al 2006; Dowd 2011). In this case, the mosquito density is known to periodically follow rainfall patterns, so state augmentation is a natural choice to treat this feature.

Thus, in order to calibrate a biased model by Kalman filtering, several steps are required:

- Estimate the state by filtering.
- Estimate the model parameters:
  - time-dependent parameters by state augmentation
  - static parameters by the Kalman filter likelihood cost function
- Estimate the model error covariance

Since the model under study is non-linear, the extended Kalman filter technique is used here, (Evensen et al 2009). Naturally other filtering methods, such as particle filters could be applied as well. For the second task, the method of choice is Markov chain Monte Carlo (MCMC) sampling for the static parameters.

A problem in the above steps is that they depend on each other: the state and model parameter estimation steps require that the model and measurement error covariances are given. Those covariances again can be estimated if the model parameters are assumed to be known and fixed, see Solonen et al (2014) for a discussion of that scenario. Here, an alternating regression-type iterative approach is used to tackle both estimation steps.

It is demonstrated that reasonable results may be achieved by such a 'gray-box' approach, that is data-driven but supported by a simplistic model, even in a case where the available data is limited and noisy, and the model is strongly biased.

Methods

Kalman filter likelihood

The Kalman filter likelihood (KFL) integrates the uncertainty in the model state using a state estimation technique resembling classical least square parameter estimation, see Petris et al (2009); Särkkä (2013).

The state estimation the Kalman filter algorithm uses observed measurements over a given time series to adjust the state of a system and to predict future states sing Bayes' theorem. In the context of parameter estimation, the likelihood function is a key component, that represents the probability of observing measured data given estimated parameters. For a given static model parameter the likelihood is evaluated by running the Kalman filter algorithm over the whole data set available and collecting the residuals between the data and model predictions. The parameter values are iteratively updated based on the evaluated likelihood function values, until they reach a maximum value, indicating that the estimated parameters provide the best fit to the observed data. This approach results in an estimate of the true state that accounts for the model bias and can lead to improved estimates compared to using the model alone, especially when the model bias is significant.

The extended Kalman filter is one of the popular methods for state estimation in nonlinear dynamical models. EKF is an extension to the Kalman filter for the case where the model is assumed to be linear Särkkä (2013); Kalman (1960); Kalman

and Bucy (1961). Let us now write the state space model as follows:

$$\mathbf{x}_k = \mathscr{M}(\mathbf{x}_{k-1}, \theta) + \mathbf{E}_k \tag{2}$$
$$\mathbf{y}_k = \mathscr{K}(\mathbf{x}_k) + \mathbf{e}_k. \tag{3}$$

Unlike in the standard EKF, the model $\mathscr{M}$ now depends on parameters θ. The model and observation errors are assumed to be zero mean Gaussians: $\mathbf{E}_k \sim \mathbf{N}(0, \mathbf{C}_k^E)$ and $\mathbf{e}_k \sim \mathbf{N}(0, \mathbf{C}_k^e)$.

In the Kalman filtering for the linear case, Equation 15 can be defined as

$$\mathbf{x}_k = \mathbf{M}_k \mathbf{x}_{k-1} + \varepsilon_k \tag{4}$$
$$\mathbf{y}_k = \mathbf{K}_k \mathbf{x}_k + \varepsilon_k. \tag{5}$$

In KF, the prediction and update steps can be written down analytically, since everything is linear and Gaussian. EKF uses the KF formulas by replacing the nonlinear model and observation operators in the covariance computations with linearizations; $\mathbf{M_k} = \frac{\partial \mathscr{M}(\mathbf{x}_{k-1}^{est}, \theta)}{\partial \mathbf{x}_{k-1}^{est}}$ and $\mathbf{K}_k = \frac{\partial \mathscr{K}(\mathbf{x}_k^p)}{\partial \mathbf{x}_k^p}$

In EKF, the predictive distribution for the state at time k is

$$\mathbf{x}_k | \mathbf{y}_{1:k-1}, \theta \sim \mathbf{N}(\mathbf{x}_k^p, \mathbf{C}_k^p), \tag{6}$$

where $\mathbf{x}_k^p = \mathscr{M}(\mathbf{x}_{k-1}^{est}, \theta)$ is the posterior mean $\mathbf{x}_{k-1}^{est}$ from the previous time evolved with the model. The prior covariance $\mathbf{C}_k^p = \mathbf{M}_k^\theta \mathbf{C}_{k-1}^{est} \mathbf{M}_k^{\theta^T} + \mathbf{C}_k^E$ is the covariance of the state estimate evolved with the linearized model $\mathbf{M}_k^\theta$.

In the update step, the prior in Equation 6 is updated with the new observation $\mathbf{y}_k$. In the EKF formulation, the posterior is

$$\mathbf{x}_k | \mathbf{y}_{1:k}, \theta \sim \mathbf{N}(\mathbf{x}_k^{est}, \mathbf{C}_k^{est}), \tag{7}$$

where

$$\mathbf{x}_k^{est} = \mathbf{x}_k^p + \mathbf{G}_k(\mathbf{y}_k - \mathscr{K}(\mathbf{x}_k^p)) \tag{8}$$
$$\mathbf{C}_k^{est} = \mathbf{C}_k^p - \mathbf{G}_k \mathbf{K}_k \mathbf{C}_k^p. \tag{9}$$

Here $\mathbf{G}_k = \mathbf{C}_k^p \mathbf{K}_k^T (\mathbf{K}_k \mathbf{C}_k^p \mathbf{K}_k^T + \mathbf{C}_k^e)^{-1}$ and $\mathbf{K}_k$ is the linearized observation operator. The predictive distribution of measurement $\mathbf{y}_k$, needed in the likelihood evaluation, is given by

$$\mathbf{y}_k | \mathbf{y}_{1:k-1}, \theta \sim \mathbf{N}(\mathscr{K}(\mathbf{x}_k^p), \mathbf{C}_k^y) \tag{10}$$

where $\mathbf{C}_k^y = \mathbf{K}\mathbf{C}_k^p \mathbf{K}^T + \mathbf{C}_k^e$.

Next, let us construct the expression for Kalman filter likelihood. The combined likelihood of observing $\mathbf{y}_{1:n}$, given parameters θ, can be written as

$$p(\mathbf{y}_{1:n}|\theta) = p(\mathbf{y}_1|\theta)\prod_{k=2}^{n} p(\mathbf{y}_k|\mathbf{y}_{1:k-1},\theta) \tag{11}$$

$$= \prod_{k=1}^{n} \exp\left(-\frac{1}{2}\mathbf{r}_k^T(\mathbf{C}_k^y)^{-1}\mathbf{r}_k\right)(2\pi)^{-d/2}|\mathbf{C}_k^y|^{-1/2}$$

$$\propto \exp\left(-\frac{1}{2}\sum_{k=1}^{n}\mathbf{r}_k^T(\mathbf{C}_k^y)^{-1}\mathbf{r}_k + \log|\mathbf{C}_k^y)|\right) \tag{12}$$

where $\mathbf{r}_k = \mathbf{y}_k - \mathscr{K}(x_k^p)$ gives the residual between the model prediction and the data at time step k, and $|\cdot|$ denotes the determinant. Note that the normalization factors of the likelihood terms, often omitted as constants, here depend on θ implicitly through the covariances $\mathbf{C}_k^p$, and the term $\log|\mathbf{C}_k^y|$ therefore needs to be included.

The above likelihood resembles the traditional "least square" type of Gaussian likelihoods. The difference is that the model state is allowed to change between time steps, and the residuals are weighted by the term $\mathbf{C}_k^y = \mathbf{K}\mathbf{C}_k^p\mathbf{K}^T + \mathbf{C}_k^e$ instead of the measurement error covariance $\mathbf{C}_k^e$. In fact, it can be easily verified that making the model uncertainty negligible by setting $\mathbf{C}_k^E$ close to zero reduces the likelihood to that of the classical least-squares Gaussian likelihood. It should be noted that while the above formulas are, in principle, written for stochastic systems, they can also be used for deterministic models. Consequently, adding the prediction error covariances to the sum of squares terms essentially gives more weight to situations that are predictable, and down-weights the terms where the model prediction is uncertain Hakkarainen et al (2012).

The Kalman filter likelihood approach is a computationally demanding approach since it involves many repeated filter sweeps, the number of which depends on the parameter estimation method utilized.

State augmentation

State augmentation (SA) is a technique to estimate both the state of a system and model parameters. It involves adding additional variables to the state vector that represent unknown parameters. The parameters are then estimated using the Kalman filter along with the original state vector. In practice this means that model parameters are updated together with the state, whenever new observations become available. An additional issue is that the model error covariance has to be specified for the parameters. This is somewhat ad-hoc but can be tuned based on how much the parameters are assumed to change between the measurement points. As such parameters are not assumed to be fixed, they can be used to account for time-dependent features lacking in the model.

Consider the state space model given below, in which the parameter vector is modeled as an additional dynamical variable:

$$\mathbf{x}_k = \mathscr{M}(\mathbf{x}_{k-1}, \theta_k) + \mathbf{E}_k \tag{13}$$

$$\theta_k = \theta_{k-1} + \mathbf{E}_k^{\theta} \tag{14}$$

$$\mathbf{y}_k = \mathscr{K}(\mathbf{x}_k) + \mathbf{e}_k. \tag{15}$$

In SA, the combined vector $s_k = [\mathbf{x}_k, \theta_k]^T$ is treated as the state vector that is updated at each time step k. The model for the combined vector can be written as

$$s_k = \tilde{\mathscr{M}}(s_{k-1}) + \mathbf{E}_{k,\theta}, \tag{16}$$

where $\tilde{\mathscr{M}}(s_{k-1}) = [\mathscr{M}(\mathbf{x}_{k-1}, \theta_k) + \mathbf{E}_k]^T$ and $\mathbf{E}_{k,\theta}$ is the error of the augmented model $\tilde{\mathscr{M}}(s_{k-1})$, here assumed to be zero mean Gaussian with covariance matrix $\mathbf{C}_{k,\theta}^E$.

In EKF, $\tilde{\mathscr{M}}(s_{k-1})$ needs to be linearized with respect to s_k, which results in the following Jacobian matrix:

$$\frac{\partial \tilde{\mathscr{M}}(s_{k-1})}{\partial (s_{k-1})} = \begin{bmatrix} \frac{\partial \mathscr{M}(s_{k-1})}{\partial \mathbf{x}_{k-1}} & \frac{\partial \mathscr{M}(s_{k-1})}{\partial \theta_{k-1}} \\ \frac{\partial \theta_{k-1}}{\partial \mathbf{x}_{k-1}} & \frac{\partial \theta_{k-1}}{\partial \theta_{k-1}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathscr{M}(s_{k-1})}{\partial \mathbf{x}_{k-1}} & \frac{\partial \mathscr{M}(s_{k-1})}{\partial \theta_{k-1}} \\ 0 & 1 \end{bmatrix}.$$

In the above matrix, in addition to the linearization with respect to the actual states (top left term), which is needed in the standard states-only EKF as well, the derivative with respect to the parameters is given (the top right term). The matrix can then be used in EKF formulas. Just as in EKF, the model error covariance matrix $\mathbf{C}_{k,\theta}^E$ also needs to be defined. In SA, this must be defined in the joint space of the state and the parameters. and modeled for simplicity as

$$\mathbf{C}_{k,\theta}^E = \begin{bmatrix} \mathbf{C}_k^E & 0 \\ 0 & \mathbf{C}_k^{\theta} \end{bmatrix}.$$

See Hakkarainen et al (2012) for a detailed interpretation of the model error covariance $\mathbf{C}_k^E$ and parameter error covariance $\mathbf{C}_k^{\theta}$. It can easily be shown that when $\mathbf{C}_{k,\theta}^E$ is set to be small the method reacts slowly on new data and the parameter values take small steps. On the other hand, if $\mathbf{C}_{k,\theta}^E$ is set large, the method allows larger deviations. Some theoretical discussion about the effect of $\mathbf{C}_{k,\theta}^E$ is given in Hakkarainen et al (2012). The procedure described in section is applicable for finding an optimal $\mathbf{C}_{k,\theta}^E$ value.

Given that one of the model parameters, the mosquito density (denoted as *mab* based on the reparametrization done in Amadi and Haario (2022)), is known to periodically follow rainfall patterns and is thus time-dependent, the state augmentation approach becomes a useful tool for estimating this parameter. Thus, there are three states, two of which are the actual Ross model compartments and the last is the value of the mosquito density.

MCMC parameter identification

In addition to finding a parameter value that makes a model to fit available data, one should study the reliability of the fit: how well the parameters of a model can be distinctly identified by the data. When model parameters are poorly identifiable, one has little confidence that the estimated values are close to the true values. The fit of a model to data may be equally good for many combinations of parameters, with values far away from each other. This may be due to one large parameter region, or possibly due to several global minima. Also, an optimizer may converge to local minima. (Hines et al 2014).

Parameter identifiability is usually diagnosed using the Markov chain Monte Carlo (MCMC) approach. This method moves us away from the idea of accepting a single best fit to the data and instead identifies all regions of the parameter space that are in agreement with the observations. Browning et al (2020). Thus, by providing the *posterior distribution* of parameter values consistent with the available data, the approach yields both the estimates of model parameters and their uncertainties.

The MCMC approach involves drawing candidate parameter values from a predefined proposal distribution. These candidate points are either accepted or rejected based on the likelihood values so that the closeness of the model output to the data is taken into account. Adaptive MCMC is particularly employed in this study since at the initial instant it may not be possible to clearly identify a well-working proposal distribution Haario et al (2006). Adaptive MCMC is an improved version of the Metropolis algorithm that updates the proposal covariance during the MCMC run, by using the information of the previously sampled points.

The structure of the posterior distribution will indicate whether the model parameters are uniquely constrained by data. A good practice in visualizing how well the chain is mixing is to make a plot of the autocorrelation functions of the parameter chain, where one can observe the extent samples that are k steps apart correlate with each other. Since in MCMC sampling the next points depend on the previous points, it is expected that subsequent points correlate more with each other than points that are farther away. Thus, poor mixing is indicated if far-away samples still have a high autocorrelation. Again, if the parameter values are in best agreement with the data and are confined to a small region of the parameter space, then the model parameters are well identified.

In the example presented here, the static parameters ac, r, μ are identified by MCMC, using the filter-likelihood. Note that using a re-parametrization the dimension of the unknown parameter vector is reduced to four instead of the original six-parameter Ross model, following a previous study where the structural identifiability of Ross model parameters Amadi and Haario (2022) was checked. This approach yields better identifiability and algorithm convergence.

Estimating the model error covariance for biased model

The state and model parameter estimation procedures described earlier require that the model error covariance $\mathbf{C}^E_{k,\theta}$ is given, in addition to the measurement error covariance. A simple parameterization $\mathbf{C}^E_{k,\theta} = \sigma^2 \mathscr{I}$ is used here, where σ^2 is the variance parameter. Recall that the likelihood in Equation 12 is only an approximation of the true likelihood in Equation 11, and the accuracy of this approximation depends on how well the EKF assumptions (linear model used in error propagation, model error assumed independent of the state) are met Solonen and Järvinen (2013); Solonen et al (2014). Checking the effect of specifying non-optimal values for the model error covariance on the quality of parameter estimation has been studied in Hakkarainen et al (2012). As discussed in more detail in Hakkarainen et al (2012), the ability to estimate the variance parameters from the measurements together with the unknown parameters θ is one of the advantages of the likelihood approach, compared to the state augmentation.

In this work, the variance parameters for the case of real data are estimated using the procedure described in Algorithm 1. The algorithm gives an iterative way to get a suitable model error covariance with respect to the bias in the model under study.

Algorithm 1 Model error covariance in the case of model bias

1. Simulate a synthetic case with similar behavior as the real data. Add noise and bias.
2. Obtain initial $\mathbf{C}^E_{k,\theta}$ by hand-tuning using data from step 1.
3. Fix $\mathbf{C}^E_{k,\theta}$ from step 2, perform MCMC to obtain $\hat{\theta}$ for the real data.
4. Fixing $\hat{\theta}$ from step 3, optimize the parameters of $\mathbf{C}^E_{k,\theta}$ by minimizing the filter likelihood
5. Repeat steps 3 and 4 a given number of times.
6. Using the last $\mathbf{C}^E_{k,\theta}$ perform MCMC and obtain the final posterior $\hat{\theta}$.

The idea of the above algorithm is to learn about the model error covariance for the real data case using values by a synthetic case where bias and noise are added. In the example case given here, the simulations employ the parameter values for high transmission case in Wallace et al (2014). The synthetic situation is created so that it resembles the real data and can be used to get initial values for the estimation. This is basically the idea of steps 1 to 3 in the algorithm. However, the model error covariance is then double-checked in step 4 by following the procedure described in Solonen et al (2014). In practice, it is assumed that the model is wrong to an unknown extent, and the observed data is noisy. For the sake of getting a suitable model error covariance $\mathbf{C}^E_k$ value, the simulation is run with different model error values. These values are used to construct the model error covariance matrix (see Hakkarainen et al (2012); Solonen et al (2014); Solonen and Järvinen (2013). The model error value is selected that gives the lowest root mean square error (RMSE) value. The size of the model error with the lowest RMSE value is reached as a compromise between the erroneous model and the noisy data. The root

mean square prediction (RMSP) is calculated alongside, which as expected generally gives higher values as compared to RMSE. In principle, the larger the model error, the more easily the data is trusted rather than the model. The Kalman filter estimates associated with the model error that yields the lowest RMSE are then considered for parameter estimation. So both the model error covariance parameter and the parameters of the model are estimated. It was established in Solonen et al (2014) that with a large enough number of accurate enough observations, the model error calibration has very little effect on the accuracy of the filter, and the other way round. Thus, correctly choosing the model error covariance is especially important in this case because a large number of accurate observations is not available. Basically, the process of double-checking the model error covariance to achieve a good level of optimality can be achieved in a few iterations, after which the final posterior parameter values can be obtained (step 6).

Data

The Ross model describes the time-dependent evolution of the fraction of infected humans and mosquitoes in the population. But the data for the study consists of records for hospital cases from Delta Nigeria. Additionally, the common habit of people in that region is that the majority of them obtain antimalarial drugs from the pharmacy Isiguzo et al (2014). Thus, in this case, the data is very noisy and is only used as a relative proxy for the fraction of infected humans, that was not directly observed. More so, the data on hospital cases were scaled here such that it starts with the initial condition given in Wallace et al (2014) for the case of high transmission. For the mosquito density compartment, the rainfall data was used as a proxy, given that rainfall influences mosquito density. Thus, the rainfall data were scaled to the value of the mosquito density given in Wallace et al (2014) for the case of high transmission. Indeed, there is usually some delay after the rainfall peak before the peak in mosquito density Odongo-Aginya et al (2005).

Results

Kalman filter likelihood: no bias

It can be easily checked that making the model uncertainty terms negligible (by setting $\mathbf{C}_k^E$ close to zero) reduces the Kalman filter likelihood to that of the classical Gaussian likelihood. With a larger model error covariance, the KFL and Gaussian cost function values become evidently different. This is illustrated using the classical Ross model with synthetic data generated by adding Gaussian noise. An example was created where the KF likelihood is employed with a substantial model error

covariance ($\mathbf{C}_k^E = 0.001$ used in the main diagonal of the covariance matrix) but no bias. It can be seen in (Figure 1 (a)) that the model solutions with both cost functions are different as expected. The chains generated by MCMC using both cost functions (standard Gaussian likelihood corresponding to negligible model error followed by KF likelihood) reveal the impact of the model prediction uncertainty term in the KFL formula (see Figure 1 (b)-(d)). The expected values of the parameters remain close, but the uncertainty increases with the KF likelihood.

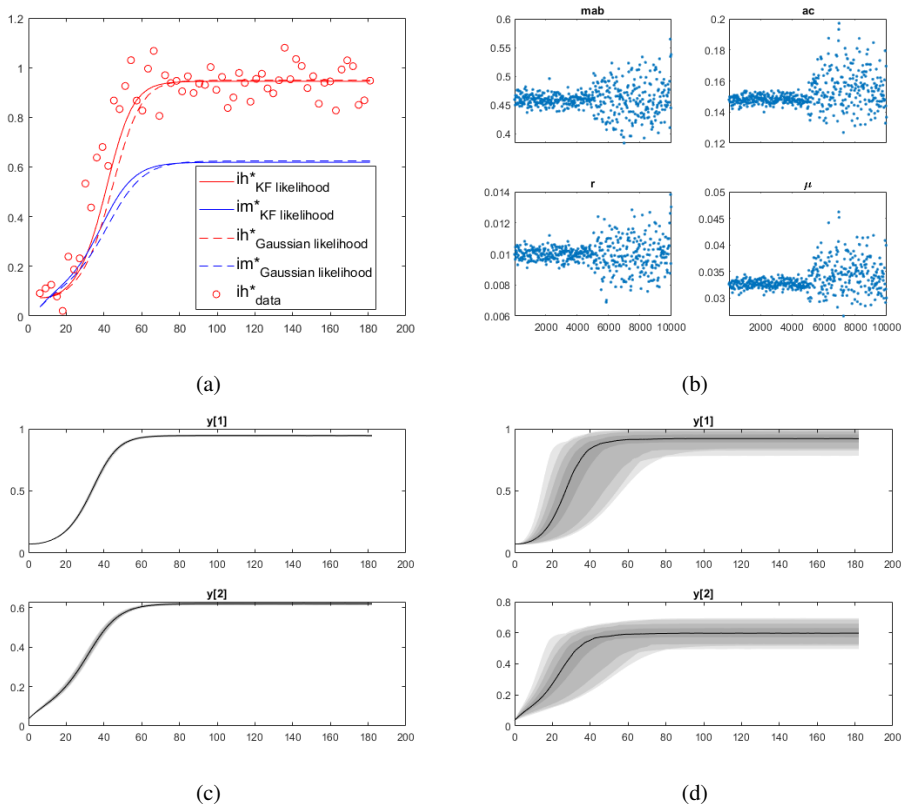


Fig. 1: **(a) Model solutions using the classical Ross model with synthetic data, for both Gaussian and Kalman filter likelihoods (b) MCMC chains with Gaussian likelihood (first 5000 chains) followed by KF likelihood (last 5000 chains) (c) Posterior predictive distribution with Gaussian likelihood (d) Posterior predictive distribution with Kalman filter likelihood.**

State, model and covariance parameter estimation: synthetic data, biased model

Using filter likelihood, the biased model can be taken into account during parameter estimation so that the right parameters for the true model can be obtained, but with larger uncertainties. This is first demonstrated using the Ross model with synthetic data. The Ross model is first simulated and then a sinusoidal component is added to it, together with Gaussian noise. The reason for the sinusoidal component is to create a situation that resembles the real data with seasonality. In this case, the model error covariance is hand-tuned such that the model follows the data. The results are given Figure 2. It can be seen that the model follows the data in a reasonable way (Figure 2 (c)). Furthermore, although the uncertainty has now grown larger as compared to the previous case of Figure 1 (b) the parameter values of the original Ross models are still contained in the posterior distributions of the parameters (see Figure 2 (b)).

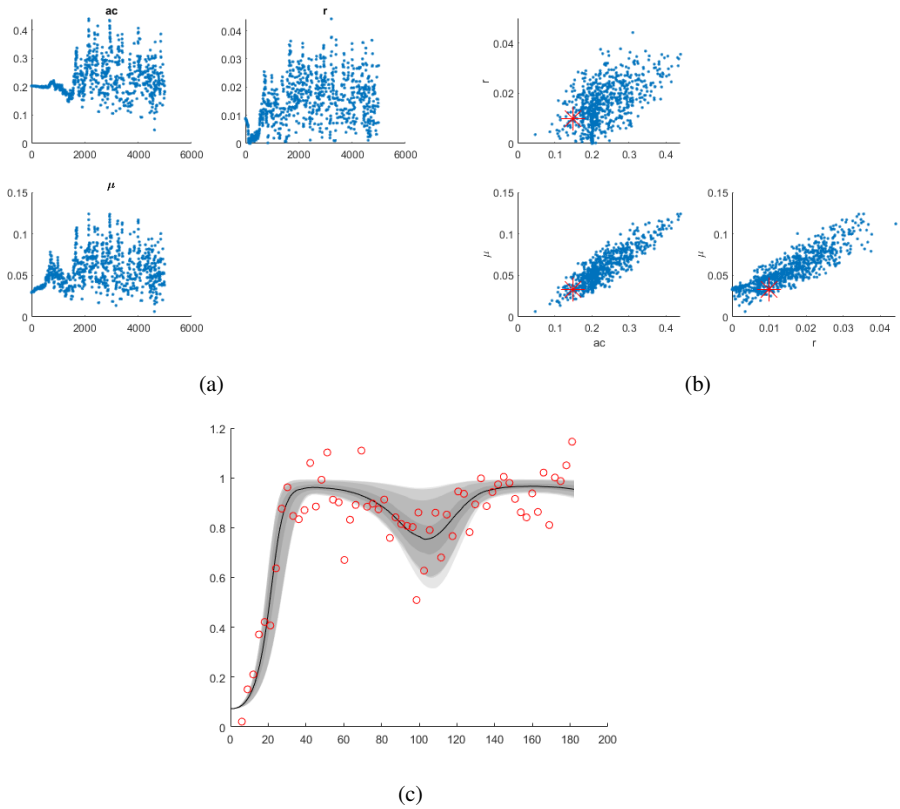


Fig. 2: **Results for Ross model with bias using a synthetic data: (a) Chain plots (b) Pairwise distribution plots: true parameter values (red star) (c) Posterior predictive distributions for the fractions of infected humans with data (red circles).**

State, model, and covariance parameter estimation: real data, biased model

The case with real data is now discussed, following the iterative procedure discussed in Algorithm 1. Basically, the exercise with the synthetic case yields initial parameters and model error covariance values for the case with real data (steps 1 to 3 in Algorithm 1). Then, the parameters for the real case are estimated and fixed. Next, the hand-tuned model error covariance is double-checked using the procedure in Solonen et al (2014) (see section). For this step in the algorithm (step 4), 20 different model error levels are used as shown in Figure 3, where the X-axis presents the index of each of the model errors considered. It can be seen that there is a sig-

nificant difference in the RMSE values for different model error variances. This is an indicator that the model error calibration has a significant effect on the accuracy of the filter, and is also meaningful as there seems to be an optimum value for $\mathbf{C}^E_{k,\theta}$ that yields the most accurate predictions. Using the optimal model error covariance value from the above-described procedure, the final posterior parameter estimates are obtained (see Figure 4. The iteration can be done a desired number of times, but here only used 2 iterations were used.

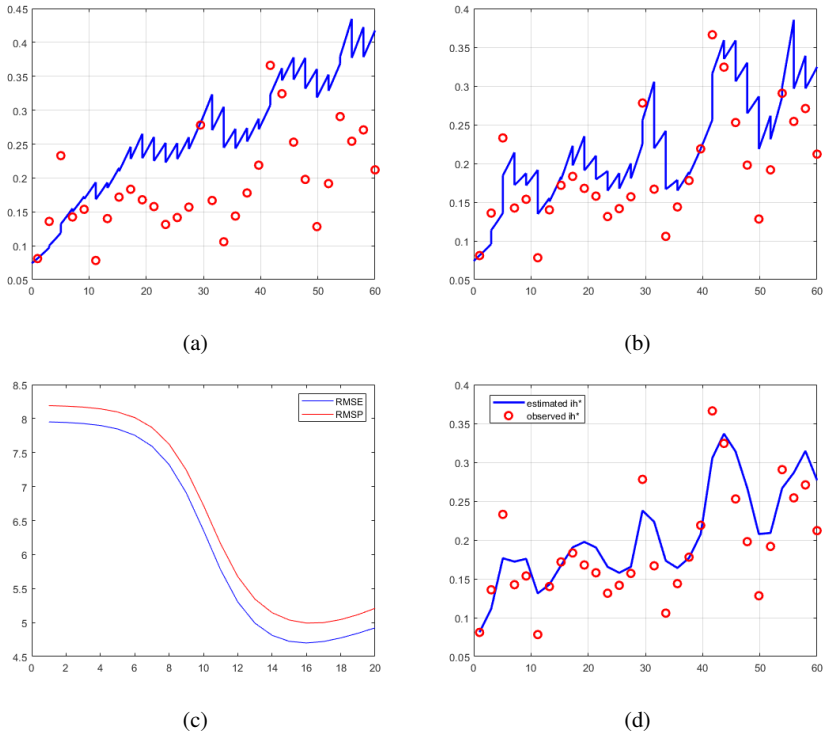
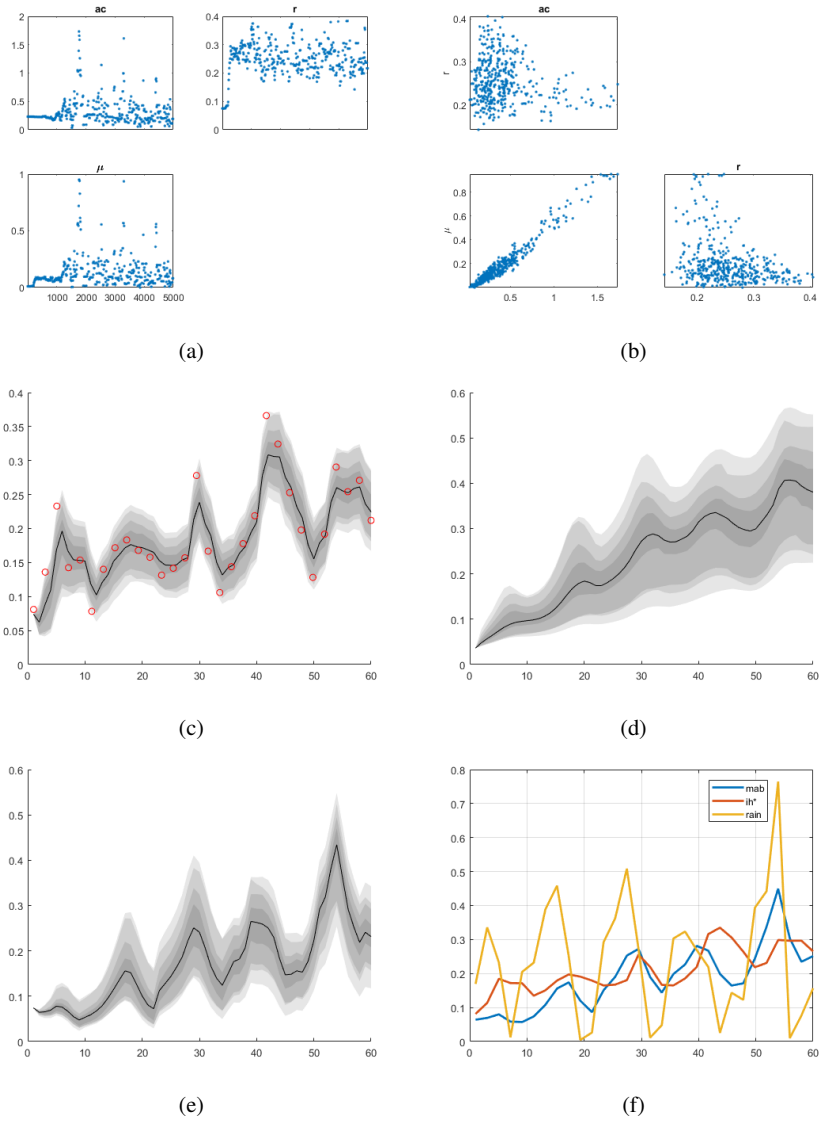


Fig. 3: **(a)-(b) Kalman filter prediction showing jumps towards the data (red circles) as model error increases (ranging from 0.00005 to 0.1) (c) Root mean square error (blue line) and root mean square prediction (red line) for different model error values. (d) Kalman filter estimate (using the model error value of 0.01, which corresponds to the lowest RMSE value) versus observed data (red circles).**

The final results are given in Figure 4. It can be seen that the parameters are well estimated but with quite large uncertainties. Also, the model trajectory follows the data reasonably well. The *mab* parameter thus arises as a compromise between the rain data and the data on infected humans (see Figure 4 (f). Apparently, there is some

delay after the peaks in rain and mosquito density before a peak in infected humans, which is realistic. Naturally, this delay can be estimated in more detail if sufficient data become available. The amount of time allowed after a peak in mosquito density to reach a peak in malaria incidence would depend on what the indicators of malaria in the area are (see section).



(a) (b)

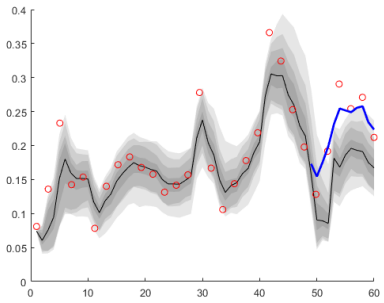
(c) (d)

(e) (f)

Fig. 4: **Results for Ross model with bias using real data: (a) Chain plots (b) Pairwise distribution plots (c) Posterior predictive distribution for the fractions of infected humans with data (red circles) (d) Posterior predictive distribution for the fractions of infected mosquitoes (e) Posterior predictive distribution for the mosquito density (f) Plot of the estimated mosquito density parameter *mab*, the estimated fraction of infected human population i_h^*, and rainfall data**

Applications of the approach

The presented approach can be used for various applications which utilize dynamical, biased models. The example presented here is, in fact, rather extreme: seasonal data and a totally non-seasonal model. In this case, the approach could be useful for hospital planning as it allows to anticipate and prepare for potential increases in patient volume for periods of rainy seasons. More so, it can help healthcare providers and policymakers allocate resources, such as medicines, hospital beds, and personnel, more efficiently. Early detection of an increase in malaria cases can help healthcare providers and policymakers take action such as spraying insecticides and distributing bed nets, to prevent a potential outbreak. For example, at a certain time in Figure 5, expecting a peak of a rainy season, the number of hospital cases can be predicted. This is demonstrated by making predictions of the malaria hospital cases using some assumed rainfall data for the next months. In order to verify how well the approach can extend to new situations, the calibration is performed with the first four years' data and then the prediction for malaria hospital cases for the fifth year using the rainfall data taken as the mean of the previous four years (see Figure 5). It can be seen that although the tails were not captured properly as in the reference case Figure 4 (c) where the fifth year also was used for calibration, the approach could fairly well predict the trajectory of the hospital cases. The results for such predictions naturally depend on the quality of data and the amount of model bias. The prediction accuracy can be expected to increase with more informative and less noisy data.



(a)

Fig. 5: **Predictive plot for the hospital cases for the fifth year, using four years data for calibration: data (red circles); mean KF estimate when the fifth year was used for calibration (blue line).**

Discussion

This work has shown how a data-driven approach using the Kalman filter likelihood can enable forecasting even when the model is biased. The Ross model used in this study is an example of a strongly biased model, which cannot capture the trajectory of real data without the aid of filtering techniques. Therefore, incorporating parameter estimation with data assimilation using state augmentation and Kalman filter likelihood together enables us to account for uncertainty in the model and improve the accuracy of the state estimates. The implementation of state augmentation with Kalman filtering depends on the specifics of the data assimilation system, such as the type of model being used and the quality of the observations. It is worth noting that these methods contain tuning parameters, and correctly specifying the model error covariance values for both the model states and parameters is critical for achieving good performance. This work discusses an iterative way of estimating the model error covariance values together with model parameters. However, effective parametrization of the model error covariance for more complex situations is still an open question and there is room for improvement.

In order to tailor the model complexity to the information content of the data, one option is to reduce the model complexity according to the available data, corresponding to a decrease in the dimension of the ODE system. This is based on a practical requirement that parameters should be expressed as functions of the known quantities of the system Wieland et al (2021). This can help to reduce uncertainties in the model. In this work, considering that the presence of mosquito dynamics gives an additional degree of freedom, a reduced Ross model which has only the infectious human compartment can be proposed. Separating the vector dynamics from the host dynamics typically requires different time scales Rocha et al (2013).

The observation error is one of the important tuning parameters. It is kept fixed in this work. However, this parameter can be estimated in an iterative manner in the data assimilation system. Generally, this approach can be used for various applications, such as weather forecasting, and control systems. Utilizing the data under study, the approach has proven to extend to new scenarios fairly well. It can however be argued that the performance of the approach can be improved when better quality data becomes available.

Conclusion

This work has proposed a data-driven approach using the Kalman filter likelihood, which can enable a meaningful comparison between the model simulations and the observations, even when the model is biased. As the method is data-driven, better quality data could further enhance its performance and improve forecasting accuracy. The approach presented in this research has broad applications in situations where the model is known to be biased.

References

Amadi M, Haario H (2022) Parameter identification and forecast with a biased model. In: Progress in Industrial Mathematics at ECMI 2021, Springer, pp 227–232

Asgarian TS, Moosa-Kazemi SH, Sedaghat MM (2021) Impact of meteorological parameters on mosquito population abundance and distribution in a former malaria endemic area, central iran. Heliyon 7(12):e08,477

Browning AP, Warne DJ, Burrage K, Baker RE, Simpson MJ (2020) Identifiability analysis for stochastic differential equation models in systems biology. Journal of the Royal Society Interface 17(173):20200,652

Dowd M (2011) Estimating parameters for a stochastic dynamic marine ecological system. Environmetrics 22(4):501–515

Evensen G, et al (2009) Data assimilation: the ensemble Kalman filter, vol 2. Springer

Grünwald P, Van Ommen T (2017) Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. Bayesian Anal 12(4):1069–1103

Haario H, Laine M, Mira A, Saksman E (2006) Dram: efficient adaptive mcmc. Statistics and computing 16:339–354

Hakkarainen J, Ilin A, Solonen A, Laine M, Haario H, Tamminen J, Oja E, Järvinen H (2012) On closure parameter estimation in chaotic systems. Nonlinear processes in Geophysics 19(1):127–143

Hines KE, Middendorf TR, Aldrich RW (2014) Determination of parameter identifiability in nonlinear biophysical models: A bayesian approach. Journal of General Physiology 143(3):401–416

Ionides EL, Bretó C, King AA (2006) Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences 103(49):18,438–18,443

Isiguzo C, Anyanti J, Ujuju C, Nwokolo E, De La Cruz A, Schatzkin E, Modrek S, Montagu D, Liu J (2014) Presumptive treatment of malaria from formal and informal drug vendors in nigeria. PloS one 9(10):e110,361

Kalman RE (1960) A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82(1):35–45

Kalman RE, Bucy RS (1961) New results in linear filtering and prediction theory. Journal of Basic Engineering 83(3):95–108

Kitagawa G (1998) A self-organizing state-space model. Journal of the American Statistical Association 93:1203–1215

Odongo-Aginya E, Ssegwanyi G, Kategere P, Vuzi P (2005) Relationship between malaria infection intensity and rainfall pattern in entebbe peninsula, uganda. African health sciences 5(3):238–245

Petris G, Petrone S, Campagnoli P (2009) Dynamic linear models with R. Springer-Verlag, New York, USA

Rocha F, Aguiar M, Souza M, Stollenwerk N (2013) Time-scale separation and centre manifold analysis describing vector-borne disease dynamics. International Journal of Computer Mathematics 90(10):2105–2125

Särkkä S (2013) Bayesian filtering and smoothing. 3, Cambridge university press, Cambridge, UK

Solonen A, Järvinen H (2013) An approach for tuning ensemble prediction systems. Tellus A: Dynamic Meteorology and Oceanography 65(1):20,594

Solonen A, Hakkarainen J, Ilin A, Abbas M, Bibov A (2014) Estimating model error covariance matrix parameters in extended kalman filtering. Nonlinear Processes in Geophysics 21(5):919–927

Spanos A (2011) Foundational issues in statistical modeling: Statistical model specification and validation. Rationality, Markets and Morals 2(47):146–178

Walker SG (2013) Bayesian inference with misspecified models. Journal of statistical planning and inference 143(10):1621–1633

Wallace DI, Southworth BS, Shi X, Chipman JW, Githeko AK (2014) A comparison of five malaria transmission models: benchmark tests and implications for disease control. Malaria Journal 13(1):1–16

Wieland FG, Hauber AL, Rosenblatt M, Tönsing C, Timmer J (2021) On structural and practical identifiability. Current Opinion in Systems Biology 25:60–69