Malaria transmission: Modeling & Inference

Anindya Bhadra

University of Michigan, Ann Arbor

June 8, 2009

Overview

- Malaria data and model
- Inference via Iterated Filtering
- Results

(Joint work with Karina Laneri, Ed Ionides and Mercedes Pascual of the University of Michigan and Menno Bouma of the London School of Hygiene and Tropical Medicine. Data provided by Ramesh Dhiman and Rajpal Yadav, National Malaria Research Institute, India)

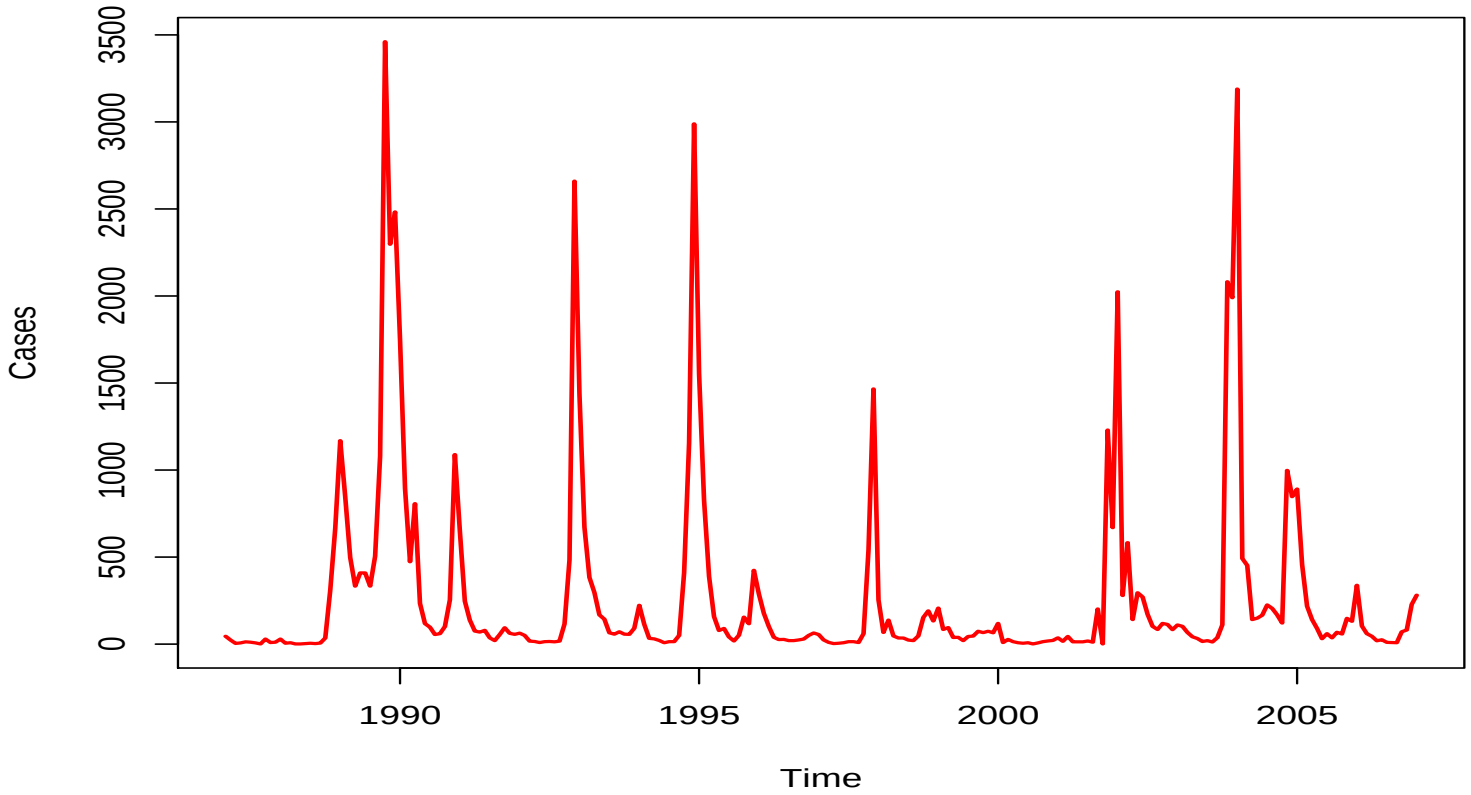
Introduction

- Malaria: 300 - 500 million cases each year, resulting in nearly 1 million deaths
- Common in Sub-Saharan Africa, parts of Asia, central and south America
- Caused by the *Plasmodium* parasite, carried by female *Anopheles* mosquitoes
- Estimated economic impact USD 12 billion/year

Our Goal

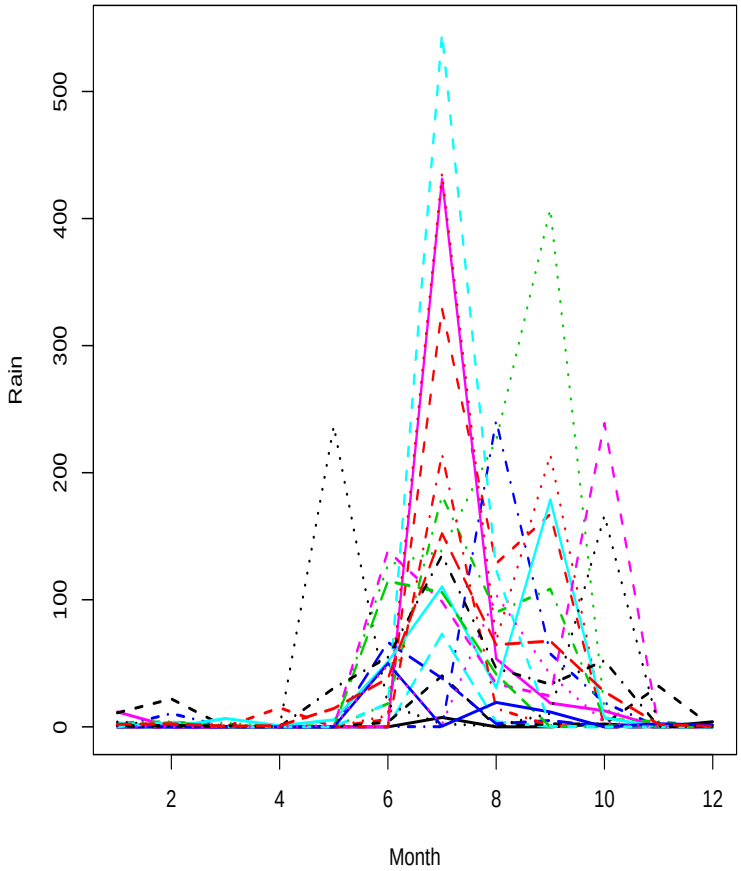
- Fit and compare mechanistic models of the disease transmission via maximum likelihood
- Analyze the role of climate covariates, especially rainfall
- Treat issues of parameter identifiability

The Data

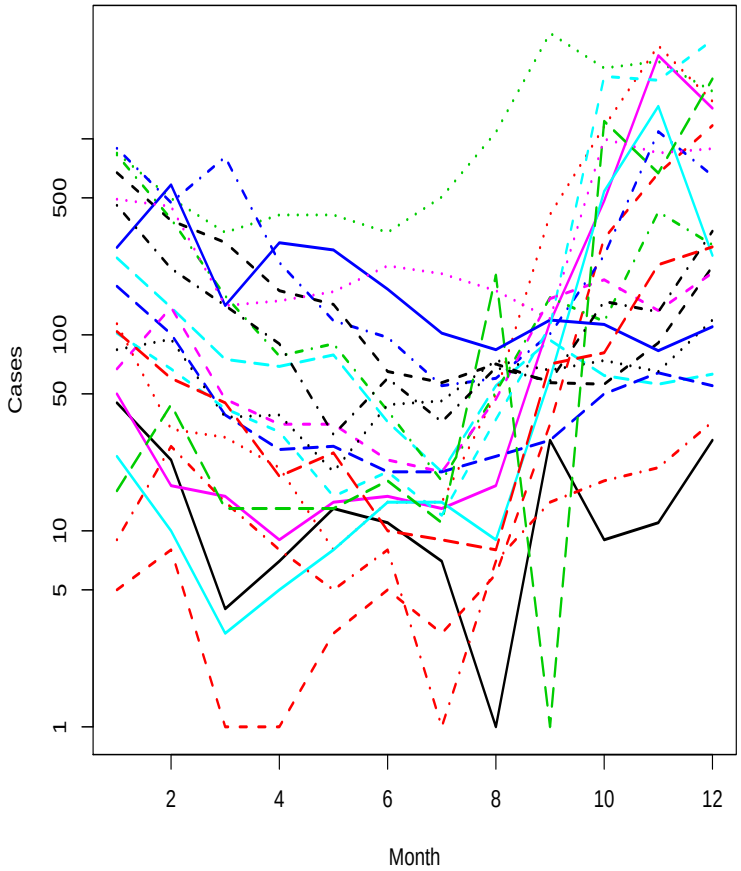


- Monthly clinical case data from Kutch, a district in the state of Gujarat, western India, 1987-2006.

The Data



- Monthly rain (mm)



- Monthly cases (log scale)

Challenges with malaria modeling

- Malaria is a complex disease with incomplete immunity, mosquito -human interaction and an intial symptom of non-specific fever.
- Previously developed methodology, e.g. the one for measles, have been unable to model malaria successfully.
- Mechanistic inclusion of climate covariates (e.g. rainfall) have not been successful. Lags? Integrals? Threshold effects?
- MCMC and Stochastic EM are known not to work.

Partially Observed Markov Processes (POMP)

- $\mathbf{x_t}$ is an unobserved vector valued stochastic process (discrete or continuous time).
  - It is assumed to be Markovian.
  - reasonable if all important dynamic processes are modeled as part of the system.
- $\mathbf{y_t}$ is a vector of the available observations (discrete time), assumed to be conditionally independent given $\mathbf{x_t}$ (a standard measurement model, which can be relaxed).
- θ is a vector of unknown parameters.

Disease Transmission Model

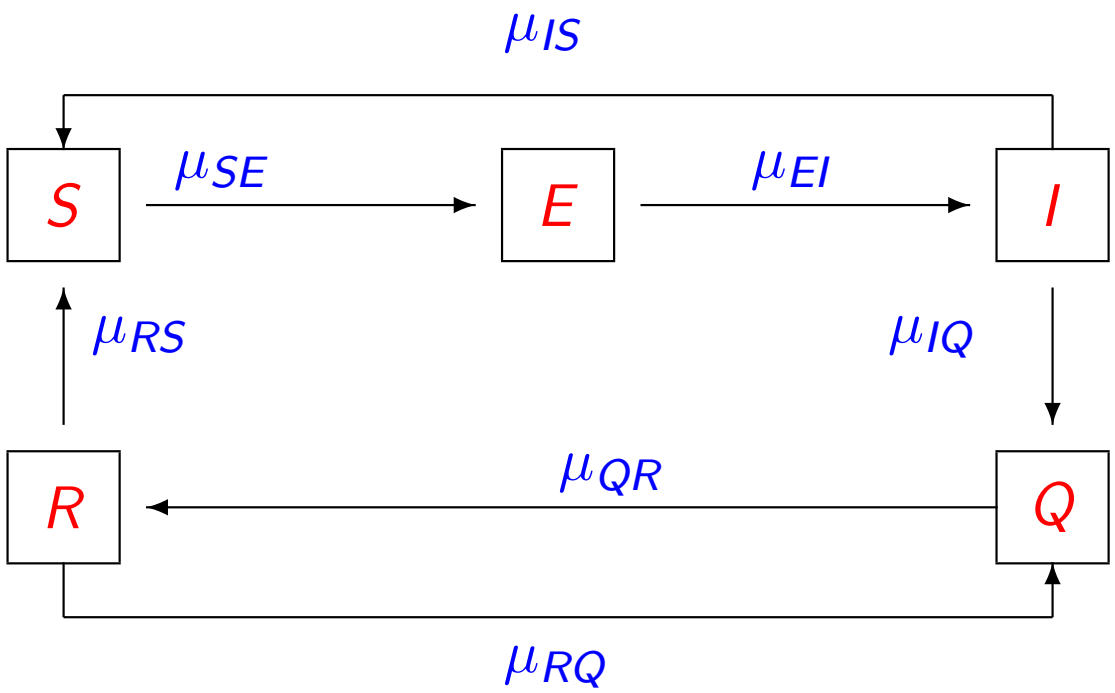


Figure: Flow diagram for the SEIQR model with superinfection

Model Equations

- State Model

$$\begin{aligned}
\frac{dS}{dt} &= \mu_{BS}N - \mu_{SE}S + \mu_{IS}I + \mu_{RS}R - \delta S \\
\frac{dE}{dt} &= \mu_{SE}S - \mu_{EI}E - \delta E \\
\frac{dI}{dt} &= \mu_{EI}E - \mu_{IS}I - \mu_{IQ}I - \delta I \\
\frac{dQ}{dt} &= \mu_{IQ}I + \mu_{RQ}R - \mu_{QR}Q - \delta Q \\
\frac{dR}{dt} &= -\mu_{RS}R + \mu_{QR}I - \mu_{RQ}R - \delta R
\end{aligned}$$

- Observation Model

$$\begin{aligned}
M_n &= \rho \int_{t_{n-1}}^{t_n} dN_{EI}(s) \\
Y_n|M_n &\sim Negbin(mean = M_n, var = M_n + \sigma^2_{obs}M_n^2)
\end{aligned}$$

Modeling the force of infection $\mu_{SE}(t)$

$$\mu_{SE}(t) = \int_0^\infty f(s)p(t-s)\,ds$$

- $f(t)$ is the effective human-human transmission rate
- $p(t)$ is a delay distribution describing the vector survival

$$f(t) = \frac{I(t)+qQ(t)}{N(t)} \exp\left\{\sum_{i=1}^{k}\beta_i s_i(t) + \beta_t t + \beta_c C(t)\right\}\frac{d\Gamma}{dt}$$

$$\int_{t_0}^{t_0+\Delta} \frac{d\Gamma}{dt}\,dt \sim \text{Gamma}(\Delta/\sigma^2, \sigma^2)$$

Coupled system of Levy SDEs

- We consider this a special case of POMP, so our general statistical framework applies.
- Our inference method is based on numerical solution to Levy SDEs which are similar to more standard Gaussian SDEs.
- Theoretical possibility of an infinite jump in in a Levy process does not appear to be a problem in practice.

Plug-and-play inference

- Statistical methods for pomps are **plug-and-play** if they require simulation from the dynamic model but not explicit likelihood ratios.
- Bayesian plug-and-play:
  1. Artificial parameter evolution (Liu and West, 2001)
  2. Approximate Bayesian computation ("Sequential Monte Carlo without likelihoods," Sisson et al, *PNAS*, 2007)
- Non-Bayesian plug-and-play:
  3. Simulation-based prediction rules (Kendall et al, *Ecology*, 1999)
  4. Maximum likelihood via iterated filtering (Ionides et al, *PNAS*, 2006)

Plug-and-play is a VERY USEFUL PROPERTY for investigating scientific models.

The cost of plug-and-play

- Approximate Bayesian methods and simulated moment methods lead to a loss of statistical efficiency.
- In contrast, iterated filtering enables (almost) exact likelihood-based inference.
- Improvements in numerical efficiency may be possible when analytic properties are available (at the expense of plug-and-play). But many interesting dynamic models are analytically intractable.

Iterated Filtering

- **Filtering**: the conditional distribution of the unobserved state vector x_n given the observations up to that time, $y_1, y_2, \ldots, y_n$.
- **Iterated filtering**: an algorithm which uses a sequence of solutions to the filtering problem to maximize the likelihood function over unknown model parameters.
  (Ionides, Bretó & King. *PNAS*, 2006)
- If the filter is plug-and-play (e.g. using standard sequential Monte Carlo methods) this is inherited by iterated filtering.

Key ideas of iterated filtering

- Bayesian inference for time-varying parameters becomes a solveable filtering problem. Set $\theta = \theta_n$ to be a random walk with

$$E[\theta_n|\theta_{n-1}] = \theta_{n-1}$$
$$Var(\theta_n|\theta_{n-1}) = \sigma^2$$

- The limit $\sigma \longrightarrow 0$ can be used to maximize the likelihood for fixed parameters.

Theoretical basis for iterated filtering

Theorem 1. (Ionides, Bretó & King, *PNAS*, 2006)
Suppose θ_0, C *and* $y_{1:N}$ *are fixed and define*

$$\begin{aligned} \theta_n^F &= E[\theta_n | y_{1:n}] \\ V_n^P &= Var(\theta_n | y_{1:n-1}) \end{aligned}$$

Assuming sufficient regularity conditions for a Taylor series expansion, $\lim_{\sigma \to 0} \sum_{n=1}^{N} (V_n^P)^{-1} (\theta_n^F - \theta_{n-1}^F) = \nabla \ell(\theta)$

Theoretical basis for iterated filtering

Theorem 2. (Ionides, Bhadra & King, (arXiv:0902.0347v1))
Let $\tilde{\theta}^F_{n,m}$ *and* $\tilde{V}^P_n$ *be the sequential Monte Carlo estimates of* $\theta^F_{n,m}$ *and* V^P_n *respectively with number of particles* J_m. *If* $\tau_m \to 0$ *and* $\tau_m J_m \to \infty$, *then under suitable regularity assumptions*

$$\lim_{m\to\infty} E\left[\sum_{n=1}^{N} (\tilde{V}^P_{n,m})^{-1}(\tilde{\theta}^F_{n,m} - \tilde{\theta}^F_{n-1,m})\right] = \nabla\ell(\theta)$$

$$\lim_{m\to\infty} {\tau_m}^2 J_m Var\left(\sum_{n=1}^{N} (\tilde{V}^P_{n,m})^{-1}(\tilde{\theta}^F_{n,m} - \tilde{\theta}^F_{n-1,m})\right) < \infty$$

Theoretical basis for iterated filtering

Theorem 3. (Ionides, Bhadra & King, (arXiv:0902.0347v1))
Let $\tau_m \to 0$ *and* $\tau_m J_m \to \infty$, $\sum_m a_m = \infty$, $a_m \to 0$ *and* $\sum_m a_m^2 J_m^{-1} \tau_m^2 < \infty$. *define a recursion by,*

$$\hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^{N} (\tilde{V}_{n,m}^P)^{-1} (\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F)$$

Then under suitable regularity assumptions for stochastic approximation (Kushner and Clark, 1978), $\lim_{m\to\infty} \hat{\theta}_m = \hat{\theta}$, *the MLE, with probability 1.*

Debate over the role of climate factors

- (Hay et al., Nature, 2002) think the epidemic cycles are generated by the disease dynamics itself, without external forcing.
- (Zhou et al., PNAS, 2004, Pascual et al., Proc. Royal Society Interface, 2007) think external drivers such as temperature and rainfall play a role.
- Previous analyses have been based on more limited data from Eastern African highlands, which are desert fringes.
- The data from India also come from a desert region and help us answer these scientific questions.

Role of rainfall

- Recall the force of infection term-

$$\mu_{SE}(t) = \int_0^\infty f(s)p(t-s)\,ds$$

$$f(t) = \frac{I(t)+qQ(t)}{N(t)} \exp\left\{\sum_{i=1}^{k}\beta_i s_i(t) + \beta_t t + \beta_c C(t)\right\}\frac{d\Gamma}{dt}$$

- For us, $C(t) = max(R(t) - 200, 0)$, where $R(t)$ is the accumulated rainfall at time t over past 6 months.
- This introduces a threshold effect of rainfall, i.e. rainfall over a certain threshold is conducive to malaria transmission.
- Observed correlation in the data: correlation between Rainfall (summed over May to August) and Cases (summed over September to December) = 0.78

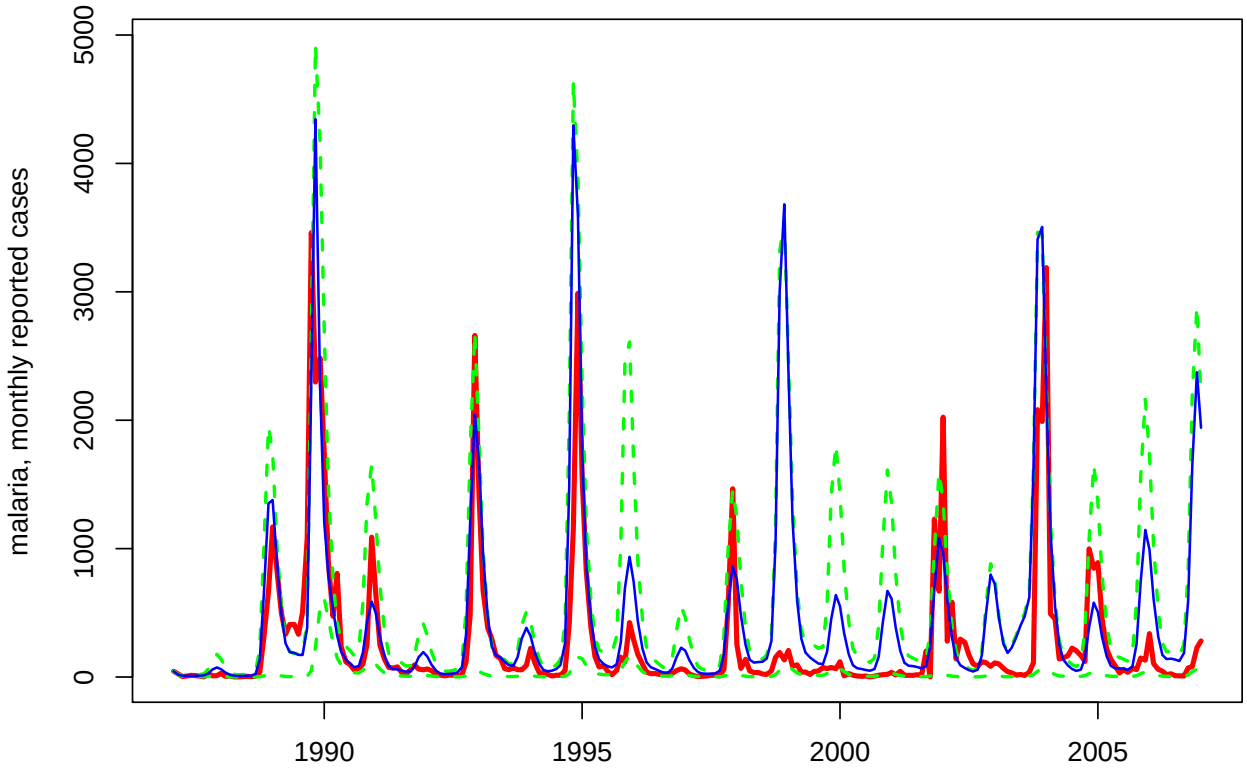
Application of iterated filtering to the malaria model : Model comparison

model	log-likelihood
SEIR	-1275
SEIR with rainfall	-1268
SEIQR with superinfection	-1274
SARIMA $(1,0,1)X(1,0,1)_{12}$	-1330

Table: Table of log-likelihoods of the fitted model

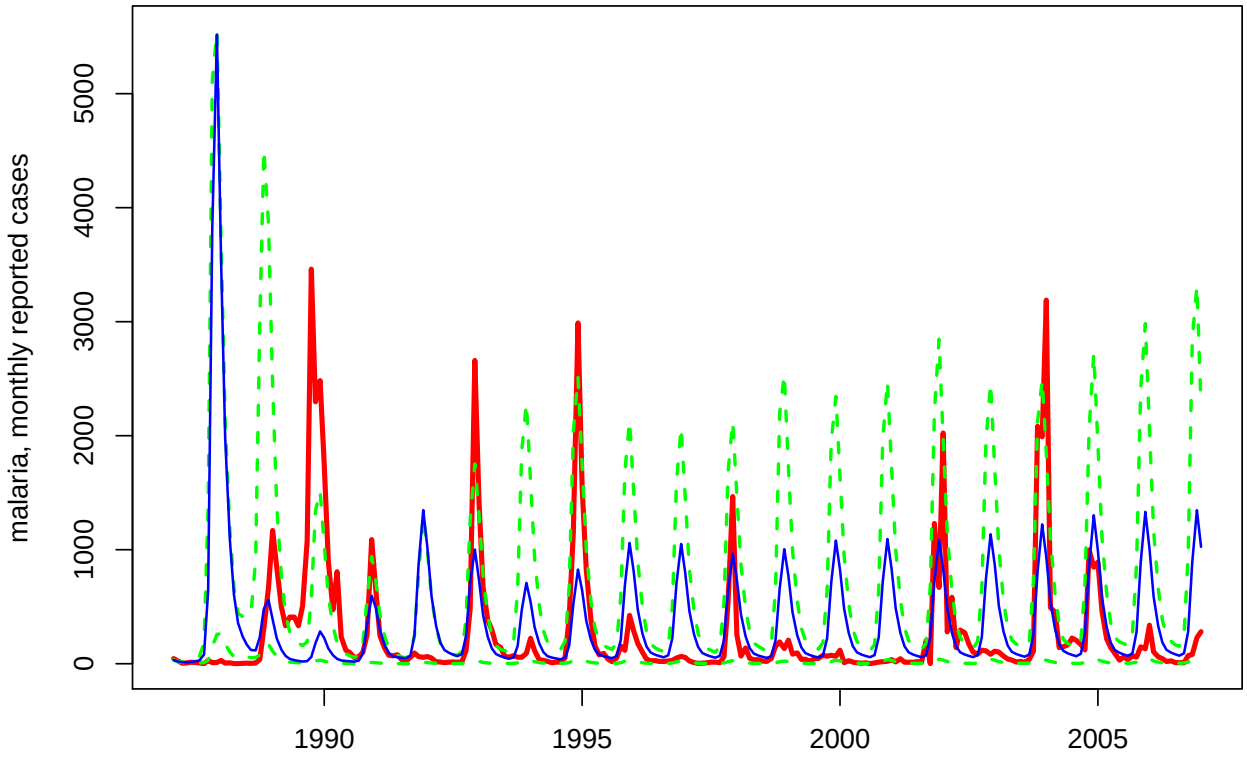
- The full model seems to be unidentifiable. From now we deal with just the first two.
- Including rainfall properly in the model improves the likelihood.

Simulations from the model with rainfall



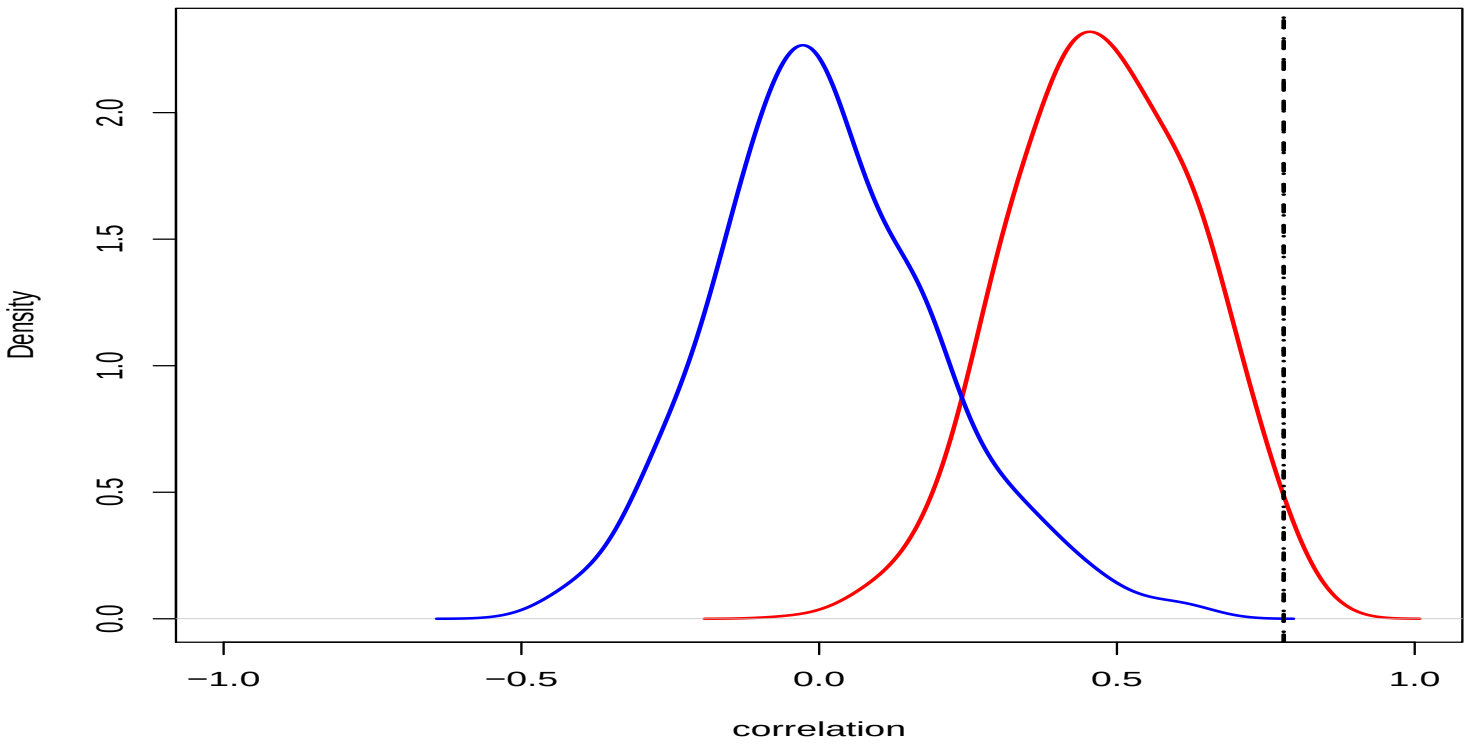
- **Cases**, **deterministic skeleton (ODE) from fit with rainfall as a covariate**, **10th and 90th percentiles of SDE model**.
- Simulations from the initial values for a long interval of 20 years look remarkably similar to the data

Simulations from the model without rainfall



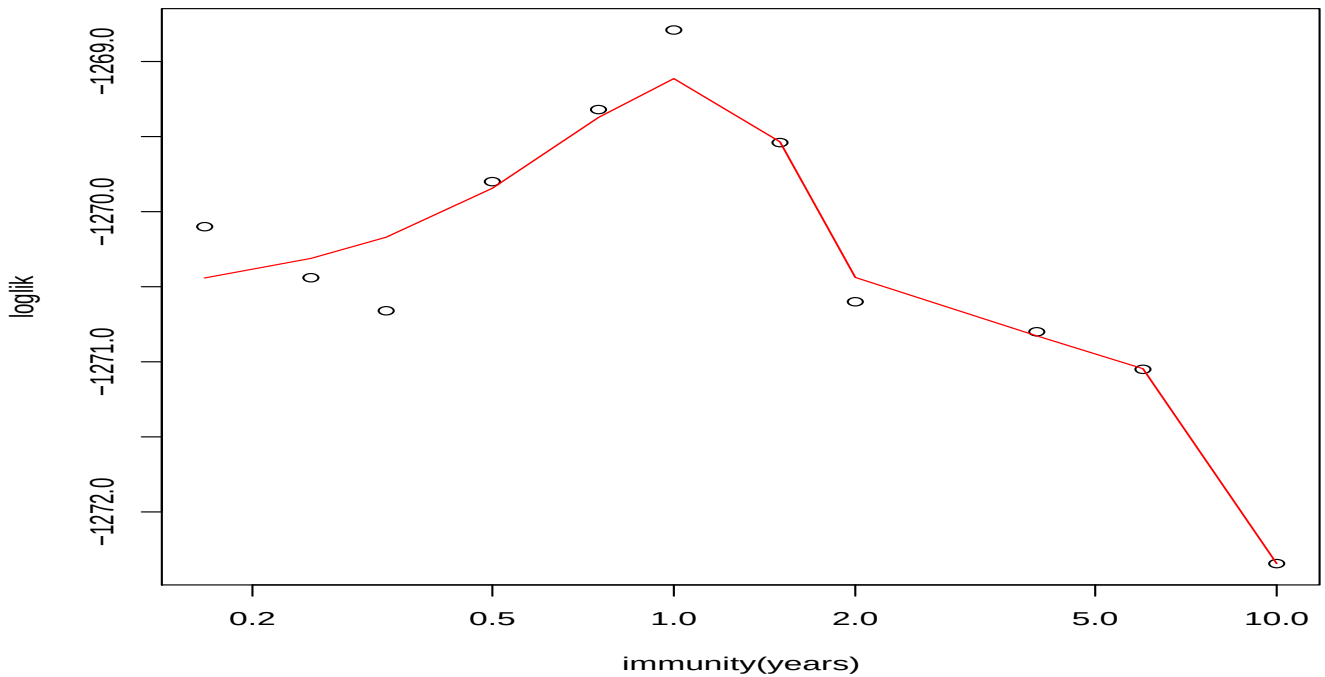
- **Cases**, **deterministic skeleton (ODE) from fitted model without rainfall**, **10th and 90th percentiles of SDE model**.
- Simulations, in particular the timing of the peaks, look very different from the data

Capturing the observed correlation in the simulations



- Density plot of correlation between Rainfall (summed over May to August) and Cases (summed over September to December) from the model with rainfall and without rainfall
- Broken black line is the observed correlation in the data

Duration of immunity



- The profile plot suggests an average duration of immunity of 1 year.
- Possibly too short to generate interannual variability

Conclusion

- Plug-and-play statistical methodology permits likelihood-based analysis of stochastic dynamic models. This enables inference on a very flexible class of models that was impossible so far.
- General-purpose statistical software for partially observed Markov processes is available in the **pomp** package for R (on CRAN).
- This technique allows us to investigate the role of climate covariates on malaria while taking into account intrinsic disease dynamics.

References

1. Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the USA*, 103 : 18438-18443.
2. Ionides, E. L. Bhadra A., and King, A. A. (2009). Iterated Filtering, arXiv:0902.0347v1.
3. Liu, J. and West, M. (2001). Combining parameter and state estimation in simulation-based filtering. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 197-224. Springer, New York.
4. Kendall, B. E., Briggs, C. J., Murdoch, W. W., Turchin, P., Ellner, S. P., McCauley, E., Nisbet, R. M., and Wood, S. N. (1999). Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology*, 80 : 1789-1805.
5. Pascual, M., B. Cazelles, M.J. Bouma, L.F. Chaves, and K. Koelle. (2008). Shifting patterns: malaria dynamics and climate variability in an East African highland. *Proc. R. Soc. London B*, 275(1631): 123-132.
6. Hay, S. I. et al. (2002) *Nature* 415, 905-909.

Thank you!

These slides are available at
`http://www.sitemaker.umich.edu/anindya`