PROTOCOL FOR:

Linking early warning signals to the temporal epidemiology of measles in Nigerien cities

Andrew Tredennick (atredenn@gmail.com)

28 September, 2018

Authors

- Andrew Tredennick
- Eamon O'Dea
- Toby Brett
- Pejman Rohani
- John Drake

Background

Theory shows that epidemic transitions can be anticipated by trends in the statistical properties of disease time series (AERO papers). The existence of statistical trends in the data that precede critical transitions, so-called 'early warning signals' (EWS), imply that we may be able to anticipate disease emergence and outbreaks. The end goal is a model-independent detection system, where statistical properties of disease surveilence data can trigger warnings of impending outbreaks without the need to fit mechanistic models of disease transmission (Han and Drake 2017).

However, there is currently a gap between the theoretical work, which has relied on knowing the underlying disease dynamics, and the eventual goal of applying EWS in real-world situations where the underlying disease dynamics may be unknown. Theoretical development of EWS has focused on anticipating when the population becomes supercritical, when $\mathcal{R}_0 > 1$, after which an outbreak is inevitable, perhaps with some bifurcation delay (Dibble et al. 2016). Knowing the value of $\mathcal{R}_0$ through time makes it possible to test the accuracy of EWS that are estimated from state variables alone. Empirical application of EWS does not require knowing the value of $\mathcal{R}_0$ through time, meaning that "tests" require making assumptions about when critical transitions occur. Whether EWS track and/or anticipate underlying dynamics of real disease time series remains unknown, and is a critical knowledge gap that must be filled before EWS can confidently be deployed.

To fill this gap we will fit a mechanistic model to incidence data of measles in Niger to estimate the temporal epidemiology of the disease, yielding the very same parameters that are known in data-free modeling studies. In particular, we are interested in the correlation between EWS and the time-varying repreoductive ratio, known as the effective reprodutive ratio ($\mathcal{R}_E$). If EWS and $\mathcal{R}_E$ are significantly and positively correlated, then we have empirical evidence that EWS are applicable in real-world settings. If EWS and $\mathcal{R}_E$ are negatively correlated or not significantly posivitely correlated, then we have evidence that EWS may not be applicable in certain settings.

Research questions

1. Do model-independent early warning signals *track* and/or *anticipate* the effective reproductive ratio?
2. Do some early warning signals outperform others in terms of their ability to track or anticipate the effective reproductive ratio?
3. Does the degree to which early warning signals track or anticipate the effective reproductive ratio depend on the magnitude or dynamics of the effective reproductive ratio itself? For example, the

temporal correlation between an EWS and $\mathcal{R}_E$ might be weak when $\mathcal{R}_E < 1$, where the dynamics are subcritial and the observed time series is mainly driven by environmental or demographic stochasticity.

Study design

I. Fit a mechanistic SIR model to measles incidence data from 4 cities in Niger

The goal is fit a model to estimate:

- The latent S, E, and I latent states
- Time-varying rate of transmision, β_t
- Time-varying effective reproductive ratio, $\mathcal{R}_{E(t)}$

The mechanistic SEIR model

The model is a discrete-time approximation of a continuous-time SEIR model with limited demography, specified as a set of difference equations,

$$S_{t+\delta t} = n_{S,t} - n_{E,t} \tag{1}$$
$$E_{t+\delta t} = n_{E,t} - n_{I,t} \tag{2}$$
$$I_{t+\delta t} = n_{I,t} + n_{O,t} - n_{R,t}, \tag{3}$$

where $\mathbf{n}_t$ are random variables representing the number of individuals transitioning into or out of each class at each timestep $t \to t + \delta t$. n_S is the number of births, n_E is the number of newly infected individuals that have the disease but are not infectious, n_I is the number of newly infectious indiviuals, n_O is the number of imported infections, and n_R is the number of newly recovered individuals who are no longer infectious and have life-long immunity. The stochastic random variables are specified as follows:

$$n_{S,t} \sim \text{Poisson}(\mu_t N_t \times \delta t) \tag{4}$$
$$n_{E,t} \sim \text{Binomial}(\lambda_{E,t}, S_t) \tag{5}$$
$$n_{I,t} \sim \text{Binomial}(\lambda_{I,t}, E_t) \tag{6}$$
$$n_{O,t} \sim \text{Poisson}(\psi \times \delta t) \tag{7}$$
$$n_{R,t} \sim \text{Binomial}(\lambda_{R,t}, I_t), \tag{8}$$

where μ_t is the birth rate at time t, ψ is the rate of imported infections, and λ_E, λ_I, and λ_R are the probabilities of exposure, becoming infectious, and recovery, respectively. These probabilities reflect the processes of tranmission, transition from the latent period to the infectious period, and recovery, which we model as:

$$\lambda_{E,t} = 1 - e^{-\frac{\beta_t I_t \delta t}{N_t}} \tag{9}$$
$$\lambda_{I,t} = 1 - e^{-\eta E_t \delta t} \tag{10}$$
$$\lambda_{R,t} = 1 - e^{-\gamma I_t \delta t}, \tag{11}$$

where β_t is time-varying rate of transmission, η is time-invariant rate from the exposed class to the infectious class, and γ is time-invariant recovery rate. We model rate of transmission as:

$$\beta_t = \beta \left(1 + \sum_{i=1}^{6} q_i \xi_{i_t} \right) \Gamma_t. \tag{12}$$

β is the mean transmission rate, ψ accounts for measles infections from external sources that are not part of the local dynamics, and the term $\sum_{i=1}^{6} q_i \xi_{i_t}$ is a B-spline to model seasonality in transmission. The B-spline bases (ξ_{i_t}) are periodic with a 1 year period. The transmission rate (β_t) is also subject to stochastic process noise at each time step, Γ_t, which we model as a gamma-distributed white (temporally uncorrelated) noise with mean 1 and variance σ^2 (Bretó and Ionides 2011).

We do not include a death process in the model because we expect death rates from the susceptible and infectious classes to be minimal relative to births and we are not interested in robust estimates of the recovered class. Excluding deaths means we can avoid making further assumptions about demographic rates – we are already making assumptions about birth rates (e.g., the rate is the same across cities, but with city-specific population size). We model demographic stochasticity in births and imported infections by drawing time-specific values from Poisson distributions. Transitions in the model are shown in Table 1. In this model, the effective reproductive ratio at time t is: $\mathcal{R}_{E(t)} = \beta_t/\gamma$.

The data are weekly observations of reported cases. To relate our model, which iterates on a daily time step, to the data, we created a variable x that is the cumulative number of individuals that transitioned from the exposed class to the infectious class over seven day periods: $x = \sum_{i=1}^{7} n_{I,i}$. We assume observed case reports ($\mathbf{y}$) for each week w are drawn from a Negative Binomial distribution subject to a constant reporting fraction (ρ) and dispersion parameter τ,

$$y_w \sim \text{Negative Binomial}\left(\rho x_w, \tau\right). \tag{13}$$

Table 1: Transitions in the SEIR model. We show the determinstic transmission rate for clarity, but our model uses the stochastic tranmission rate.

Transition	(ΔS, ΔE, ΔI)	Propensity
birth	$(1, 0, 0)$	$N_t \mu_t$
transmission (deterministic)	$(-1, 1, 0)$	$SI\beta_t/N_t$
transmission (stochastic)	$(-k, k, 0)$	$\frac{S}{k} \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j+1} \tau_{\mathrm{f}}^{-1} \ln(1 + (\beta_t I/N_t)) \tau_{\mathrm{f}}(S-j)$
symptomatic (infectious)	$(0, -1, 1)$	$E\eta$
imported infections	$(0, 0, 1)$	ψ_t
recovery	$(0, 0, -1)$	$I\gamma$

Model fitting

We will fit the model to data via iterated filtering to maximize the likelihood, as implemented in the R package `pomp`. We seek to estimate 12 parameters for each city (Table 2). Note that we set $\gamma = 365/14$ assuming a 2 week infectious period. We assume γ is known to improve estimates of other parameters and because the infectious period of measles is well studied.

Table 2: List of model parameters to be estimated.

Parameter	Description
β	Mean rate of transmission
ρ	Constant reporting fraction
q_i for $i \in \{1,2,3,4,5,6\}$	B-spline coefficients for seasonality
ψ	External infections
σ^2	Variance of gamma-distributed environmental noise
S_0	Initial conditions for susceptible class
I_0	Initial conditions for infected class

Model fitting will proceed in three steps, independently for each city:

~~Generate initial parameter sets from a large Latin hypercube search space of 5000 parameter sets. Run 10 replicate particle filters at each of these parameter sets, with 2000 particles each. Retain the 500 parameter sets with the highest likelihood.~~

1. Conduct a semi-global parameter search using MIF, starting from 1000 parameter sets from a Latin hypercube. After convergence, assessed every 50 MIF iterations, we will retain the 500 estimates with the highest likelihood for the next MIF step. At this stage, we will use the following `pomp` parameters, `pomp::mif2(Np = 2000, Nmif = 50, cooling.fraction.50 = 1, cooling.type = "geometric")`, and initial parameter conditions allowed to take random walks with a standard deviation of 0.02. For states (i.e., S_0, I_0, R_0), we will set the initial conditions random walk standard deviation to 0.1 and the random walk will stop after 1 year (52 weeks) of the simulated dynamics.

2. Conduct a local parameter search using the retained final 500 parameter sets from above (the 500 with the highest likelihood). This parameter set will serve as starting conditions for 500 instances of iterated filtering using the following `pomp` parameters, `pomp::mif2(Np = 2000, Nmif = 50, cooling.fraction.50 = 0.95, cooling.type = "geometric")`, and initial parameter conditions allowed to take random walks with a standard deviation of 0.02. For states (i.e., S_0, I_0, R_0), we will set the initial conditions random walk standard deviation to 0.1 and the random walk will stop after 1 year (52 weeks) of the simulated dynamics. We will consider the parameter set with the highest likelihood at this stage to be the MLEs ($\hat{\theta}$).

3. To complete our inference from the model, we will perform particle MCMC using $\hat{\theta}$ as the initial conditions. The MCMC analysis allows us to estimate the uncertainty around all parameters and to generate estimates of the latent states and the $\mathcal{R}_E$ for the observed data. That is, we will end up with a time series of $\mathcal{R}_E$, including uncertainty, that corresponds to the observation time series.

    - Can we just get this from `pfilter(save.states)`?
    - Instead of the above, we will implement a random walk for β and use the filtering distribution at each time t to estimate the time-varying rate of transmission and reproductive ratio.

~~Conduct a local parameter search using the the 50 parameter sets from the global search with the highest likelihoods. MIF parameters will be the same as above, except the cooling factor will be `cooling.fraction.50` = 0.9. We will consider the parameter set with the highest likelihood at this stage to be the MLEs ($\hat{\theta}$).~~

II. Calculate a suite of EWS from the observed data

We will consider 10 candidate EWS (Table 3). The EWS will be calculated using the function `spaero::get_stats()` in R for each of the case report time series shown in Figure 1 (below). We will use a "backward looking" bandwidth of 52 weeks. "Backward looking" means that the rolling window over which the EWS are calculated covers the 35 weeks prior to the focal time point t, rather than the window being centered on t.

Table 3: List of candidate early warning signals and their estimating equations. Note that b denotes the bandwidth. See Brett et al. (2018) for details.

EWS	Estimator	Theoretical Correlation with $\mathcal{R}_E(t)$
Mean	$\mu_t = \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{X_s}{2b-1}$	Positive
Variance	$\sigma_t^2 = \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{(X_s-\mu_s)^2}{2b-1}$	Positive
Coefficient of variation	$CV_t = \frac{\sigma_t}{\mu_t}$	Null
Index of dispersion	$ID_t = \frac{\sigma_t^2}{\mu_t}$	Positive
Skewness	$S_t = \frac{1}{\sigma_t^3} \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{(X_s-\mu_s)^3}{2b-1}$	Positive
Kurtosis	$K_t = \frac{1}{\sigma_t^4} \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{(X_s-\mu_s)^4}{2b-1}$	Positive
Autocovariance	$\text{ACov}_t = \sum_{s=t-(b-1)\delta}^{t+(b-1)\delta} \frac{(X_s-\mu_s)(X_{s-\delta}-\mu_{s-\delta})}{2b-1}$	Positive
Autocorrelation	$\text{AC}_t = \frac{\text{ACov}_t}{\sigma_t \sigma_{t-\delta}}$	Positive
Decay time	$\overline{\tau}_t = -\delta / \ln[\text{AC}_t(\delta)]$	Positive
First differenced variance	$\Delta\sigma_t^2 = \sigma_t^2 - \sigma_{t-\delta}^2$	Positive

III. Calculate the correlation between EWS lags and $(\mathcal{R}_E)$

Answering our Research questions requires calculating the correlation between each EWS and $\mathcal{R}_E$ through time. We are interested in the correlation between EWS and $\mathcal{R}_E$ because these correlations tell us whether a particular EWS can anticipate a change in $\mathcal{R}_E$. We will calculate the temporal correlation of $[\text{EWS}(t), \mathcal{R}_E(t)]$ for all EWS with Spearman's ρ using the R function `cor(..., method = "spearman")`. Pearson's correlation coefficient is not suitable for testing nonlinear associations and Kendall's τ does not take into account the distance between discordant pairs, so we will use Spearman's rank correlation.[1] As shown in Table 2, we expect all EWS to positively covary with $\mathcal{R}_E$ over time, a prediction we can test by calculating the correlations and their significance using the function `cor.test(..., method = "spearman")`.

There often exists a lag between when the dynamical system goes critical and when an outbreak or epidemic occurs, so-called bifucation delay (Dibble et al. 2016). Therefore, EWS that correlate well with underlying dynamics (i.e., $\mathcal{R}_E(t)$) should be useful for anticipating outbreaks.

Data sources

We will use weekly measles case report data from four Nigerien cities, collected over an 11 year period (1995-2005) (Figure 1). These data are ideal for stress testing EWS because each city has different population sizes, has different dynamics in terms of size of outbreaks and length of inter-epidemic periods, and each time series has different amounts of "noise" (though, the difference in variability is probably just reflective of population size differences). The data come from [*somewhere/someone*], and used here with permission from [*somewhere/someone*].

Analysis and expected results

The results will start with a figure of example time series of $\mathcal{R}_E$ and an EWS (e.g., variance), as well as a scatterplot of their correlation (all in one multipanel plot, e.g., Figure 1).

[1] Note that Spearman's ρ does not handle ties very well, so we may have to fall back on Kendall's τ is there are many ties and we do not want to rely on approximate p-values.

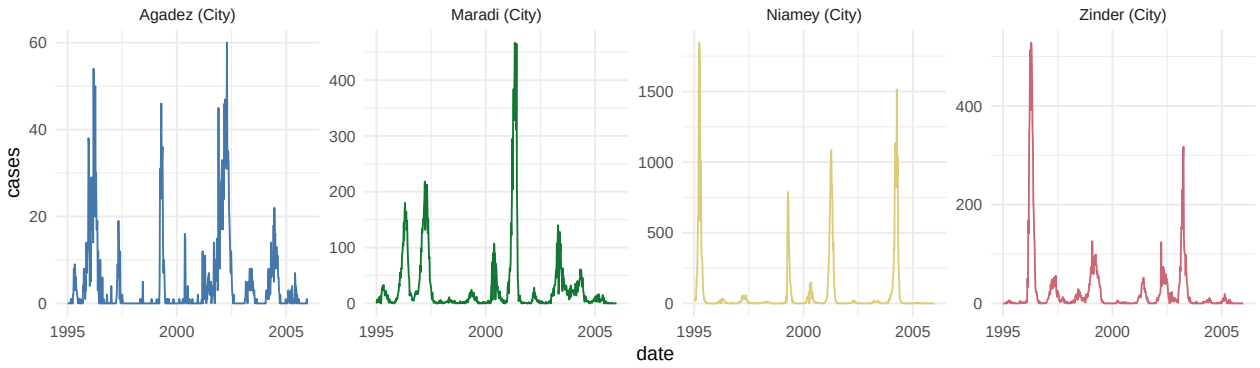

Figure 1: Time series of weekly measles case reports from four cities in Niger.

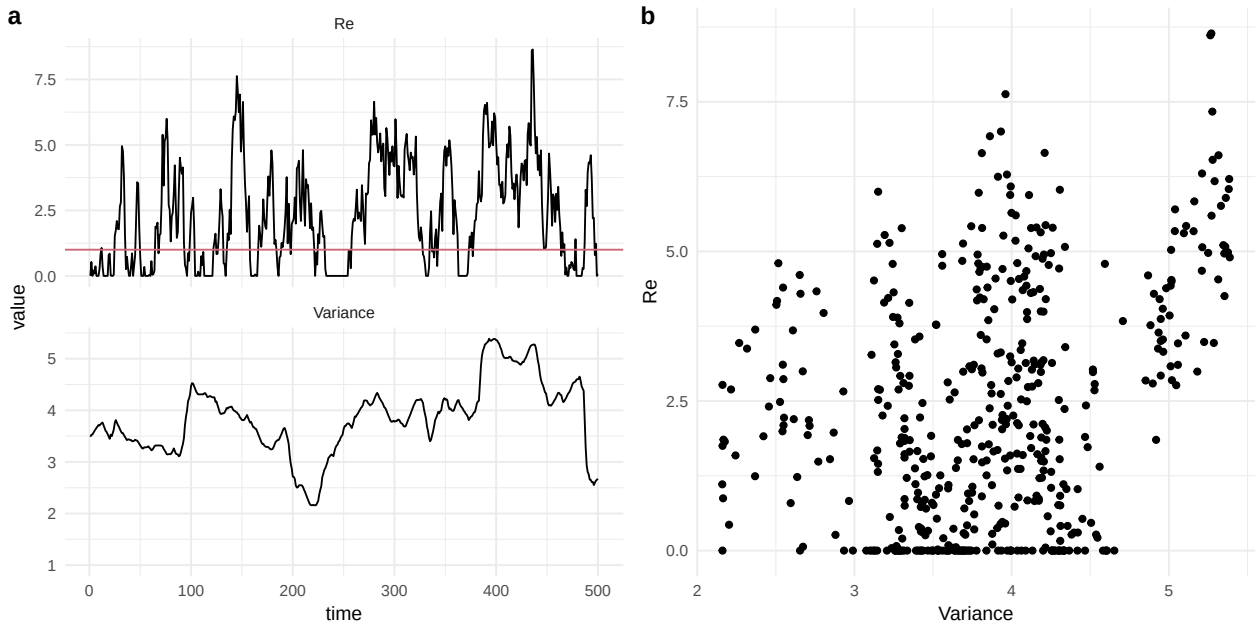

Figure 2: Time series of $\mathcal{R}_E$ and the variance of case reports (a). Scatterplot showing the correation between $\mathcal{R}_E$ and the variance of case reports (b).

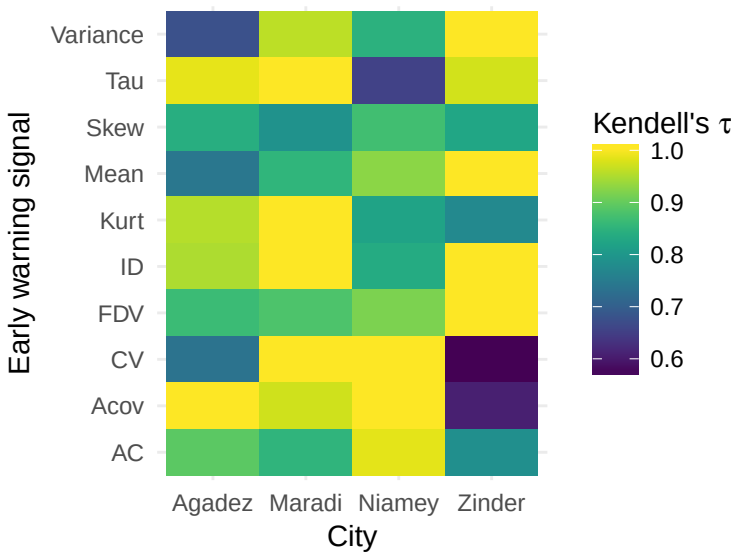

Figure 3: Heatmap of temporal correlations of EWS and R_E through time for each city.

The main figure will show the correlation between each EWS and $\mathcal{R}_E$ for each city (Figure 2). Figure 2 answers question one (see Research questions): do EWS track and/or anticipate $\mathcal{R}_E$? If, in general, correlations are significant and positive, as expected by theory, then the answer is "yes." If, in general, correlations are weak (not significant), null, or negative, then the answer is "no." We can also indicate which correlations are significant, either in this figure or in a supplementary table.

The second question, "Do some EWS outperform others?", will be answered by simply ranking each EWS by correlation strength (also visible in Figure 2).

The third question, "Does the degree to which early warning signals track or anticipate the effective reproductive ratio depend on the magnitude or dynamics of the effective reproductive ratio itself?", will be answered by conducting a *post hoc* statistical analysis. First, the $\mathcal{R}_E(t)$ time series will be broken up into clusters according to whether $\mathcal{R}_E(t) > 1$ or not. Then, we will run the correlation analysis seperately for each cluster (e.g., a single time series might have 10 sections where R < 1, and 10 sections where R > 1, yielding a sample size of 20 time series). Lastly, we will conduct an ANOVA to determine if the threshold value of $\mathcal{R}_E$ determines the correlation between EWS$_t$ and $\mathcal{R}_E(t)$. One expectation is that the correlation will be low/weak when $\mathcal{R}_E \lessapprox 1$ but high/strong when $\mathcal{R}_E \gtrapprox 1$. An example figure is shown below (Figure 3), in which whether $\mathcal{R}_E(t)$ is above/below the threshold of 1 does impact its correlation with each EWS.

Checklist

- ~~Create cleaned dataset of just the four focal cities from Niger~~
- ~~Write R code to define the `pomp` model object (SDE)~~
- Simulate time series from the `pomp` object with parameters that give us irregular dynamics like the observed data
- Do test analysis of steps II-III using the simulated data
- Perform iterated filter parameter search for MLEs
  - ~~Get set up on UGA cluster~~
  - ~~Get set up on MIDAS cluster~~
- ~~Calculate EWS on the observed data~~
  - Work with Eamon on bandwidth, etc.
- Caclulate correlations with real data and empirically-derived $\mathcal{R}_E$ and lags
- Write the paper!

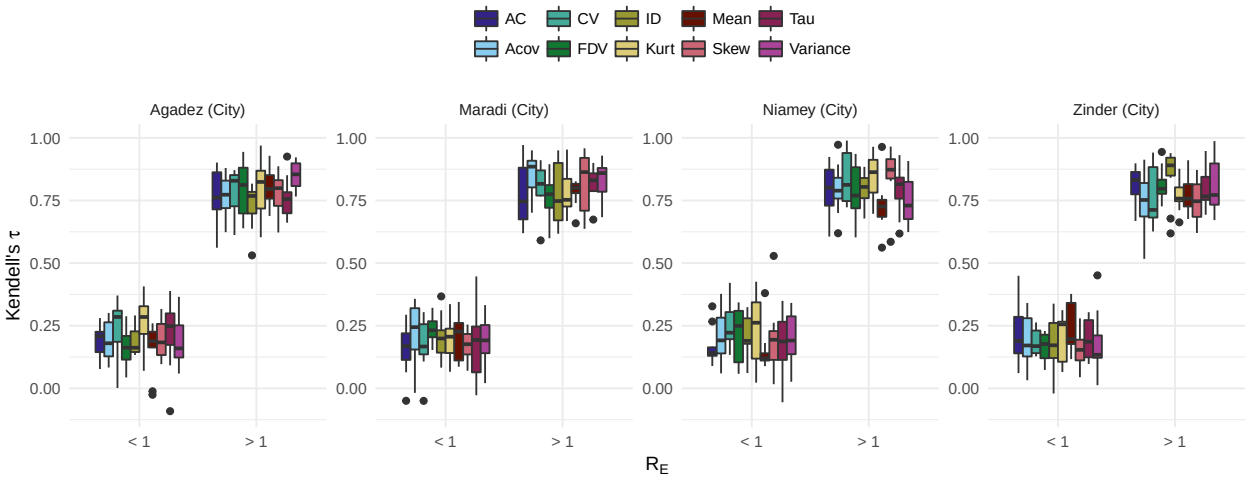

Figure 4: Boxplots of temporal correlations between EWS and R_E through time, split apart by the value of R_E.

Changelog

- **09-10-2018**
- We are going to forego the proposed initial search via particle filtering because doing so over a large enough parameter space to be meaningful is too computationally expensive. Instead, after conferring with Pej, the new plan is to start with MIF from 1,000 starting parameter sets from a Latin Hypercube design. Then we'll gradually take subsets of those.
- Imported infections are now stochastic: `psi[t] ~ rpois(psi)`.
- Likihood is now poisson. Initial fitting kept showing very little overdispersion, so we're going to use Poisson to speed up computation.
- **09-12-2018**
- Set $\gamma = 365/14$ assuming it is a known parameter.
- **09-14-2018**
- Remove R from the model because we don't need it – we have population size as a covariate. This should speed things up.
- Added some text on using the filtering distribution for final inference.
- **09-27-2018**
- Updates model to SEIR and better notation after consulting with Toby.