Electronic Supplementary Material

Inferring patterns of influenza transmission in swine from multiple streams of surveillance data

Christopher C. Strelioff[1], Dhanasekaran Vijaykrishna[2,3], Steven Riley[4,5,6], Yi Guan[7], J. S. Malik Peiris[7], James O. Lloyd-Smith[1,6]

1 Department of Ecology & Evolutionary Biology, University of California, Los Angeles, CA 90095, USA

2 Laboratory of Virus Evolution, Program in Emerging Infectious Diseases, Duke-NUS Graduate Medical School, 8 College Rd, 169857, Singapore

3 State Key Laboratory of Emerging Infectious Diseases & Department of Microbiology, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong Special Administrative Region, People's Republic of China

4 MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom

5 Department of Community Medicine and School of Public Health, The University of Hong Kong, Hong Kong Special Administrative Region, People's Republic of China

6 Fogarty International Center, National Institutes of Health, Bethesda MD, USA

7 State Key Laboratory of Emerging Infectious Diseases and School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong Special Administrative Region, People's Republic of China

1 Bayesian state-space model

A state-space model is a hierarchical statistical model used to infer time-dependent parameters related to both observed and hidden parameters from available data [1, 2]. We employ Bayesian methods, making this a Bayesian state-space model (BSSM), by sampling from the posterior distribution of model parameters using Markov Chain Monte Carlo (MCMC) and the Python package PyMC in particular [3]. Below we describe the elements of our BSSM using conventional terminology for this type of statistical model when employed in biological and epidemiological research [1]. In other fields, this type of approach can be described as a Kalman filter (for Gaussian response) or, more generally, as a hidden Markov model [2]. Limiting ourselves to the terminology of [1], the framework is made up of data, process and parameter models. Below we describe each of these elements in detail, starting with the data model. Also see figure 1 in the main manuscript for a schematic of all of the relations that make up this statistical model.

1.1 Data model

As described in the main text, there are two types of data available for our joint analysis of swine influenza: monthly counts for virus isolation and seropositivity. We employ the index t to indicate the month of interest. The serological data consist of the number of animals tested $N_s(t)$ as well as the number of positive results $n_s(t)$. For the purposes of this analysis, we use a conventional threshold and define a titer greater than or equal to 1:40 to any test antigen as seropositive. Given the sample size and unknown probability of being seropositive $p_s(t)$, the observed count for a given month has a Binomial distribution: $P(n_s(t)|p_s(t), N_s(t)) = \text{Bin}(p_s(t), N_s(t))$. The likelihood of all available serological data, which we denote $\{n_s(t)\}_{t \in T_s}$, is given by:

$$P(\{n_s(t)\}_{t\in T_s}|\{N_s(t)\}_{t\in T_s},\{p_s(t)\}_t) = \prod_{t\in T_s}\text{Bin}(p_s(t),N_s(t))\ , \tag{S1}$$

where we use T_s to label the subset of months where serology counts are available. The probability of all observed serology data is conditioned on the number of samples taken and the probability of seropositivity for all months (even those without data), denoted $\{p_s(t)\}_t$, that will be specified by part of the process model.

The data model for virus isolation is slightly more complex due to the additional information about the viral strain, which we designate i. Possible strains include classical swine (CS), triple reassortant (TR), Eurasian avian-like (EA), pandemic H1N1 (pH), and seasonal human H1N1 (Hu). The number of animals tested for virus each month is denoted $N_v(t)$ and the number of samples testing positive for strain type i is $n_i(t)$. The complete virus isolation sample for a particular month is designated $\{n_i(t)\}_i$ to indicate the set of counts for all strain types. Given the probability of virus isolation $p_v(t)$ (of any strain) and the probability that the isolate is of type i given a successful isolation, denoted $p_{i|v}(t)$, the probability of isolation of type i is given by the product $p_{i,v}(t) = p_{i|v}(t)\times p_v(t)$. Using these definitions, the number of isolates of type i for a given month t has a multinomial distribution: $P(\{n_i(t)\}_i|p_{i,v}(t),N_v(t)) = \text{Mult}(p_{i,v}(t),N_v(t))$. The likelihood of virus isolation data for all available months, denoted T_v, is given by:

$$P(\{n_i(t)\}_{i,t\in T_v}|\{N_v(t)\}_{t\in T_v},\{p_{i,v}(t)\}_{i,t}) = \prod_{t\in T_V} \text{Mult}(p_{i,v}(t), N_v(t)) \ . \tag{S2}$$

As with the serology data, this expression gives the probability of all isolation data, conditioned on the number of samples taken and the complete set of strain-specific isolation probabilities $\{p_{i,v}(t)\}_{i,t}$. As discussed above, the monthly probability $p_{i,v}(t)$ is made up of two factors: $p_v(t)$ and $p_{i|v}(t)$ that will be part of the process model. Using these elements, the likelihood given in equation (S2) can be calculated.

Finally, the likelihood for all observed data can be written as the product of equations (S1) and (S2), resulting in the form:

$$\begin{aligned}\text{likelihood} = & P(\{n_s(t)\}_{t\in T_s}|\{N_s(t)\}_{t\in T_s},\{p_s(t)\}_t) \\ & \times P(\{n_i(t)\}_{i,t\in T_v}|\{N_v(t)\}_{t\in T_v},\{p_{i,v}(t)\}_{i,t}) \ .\end{aligned} \tag{S3}$$

1.2 Process model

The process model provides a connection between the parameters conditioned on in the data model and the hidden dynamics of influenza exposure on farms and during transport that are the primary focus of our inference problem. We start by considering the monthly probabilities (i) that a sampled pig was exposed during transportation to and holding at the abattoir, $p_t(t)$, and (ii) that a sampled pig was exposed on the farm $p_f(t)$. We define transport exposure to be within the week before sampling, such that infection during transport is likely to lead to virus isolation [4, 5]. Farm exposure reflects transmission

earlier in life, and corresponds to the lifetime hazard of exposure up to one week prior to sampling, as this is the time needed to develop antibodies to influenza following infection. We assume that antibodies developed due to influenza exposure will still be measurable at the time of sampling on average. This is a reasonable assumption, even if titers decline with time, because pigs sent to the abattoir range from four to six months in age.

In the development of the process models, we employ scale parameters that control the variance in statistical relations. For example, exposure of naïve animals during transport may not translate to virus isolation in some cases, due to individual variation among hosts or false negative assay results. Further variation may arise because random sub-samples of the abattoir population obtained each month may be more or less representative of the population of interest. The scale parameters enable variation from all sources to be represented in the model. Each scale parameter is inferred from the data, as described in the next section, to ensure that the degree of variation is appropriate.

Using this general framework, we allow for the possibility that $p_t(t)$ and $p_f(t)$ do not correspond precisely to the proportion of samples that are positive by virological or serological testing. On average, though, we do expect that the probability of seropositive samples will be equal to $p_f(t)$. The scale factor s_S accounts for variation due to unknown factors, as discussed above, resulting in the following expressions for the expectation and variance of seropositivity as a function of farm exposure:

$$\mathbf{E}[p_s(t)|p_f(t), s_S] = p_f(t) \ , \tag{S4}$$

$$\mathbf{Var}[p_s(t)|p_f(t), s_S] = \frac{p_f(t)\big(1 - p_f(t)\big)}{1 + s_S} \ . \tag{S5}$$

The form of equation (S5) shows that as the scale s_S increases, the correspondence between

farm exposure and seropositivity increases (i.e. the variance decreases). This inverse relationship between scale parameter value and variance is found throughout the BSSM.

We assume that the probability of virus isolation depends on two factors: (i) the absence of protective immunity arising from exposure on the farm, and (ii) exposure during transportation and holding before slaughter. We denote the probability corresponding to this combination as $p_{\not f,t}(t) = (1 - p_f(t)) \times p_t(t)$ and write the expectation and variance for the probability of virus isolation as a function of both transport and farm exposure:

$$\mathbf{E}[p_v(t)|p_{\not f,t}(t), s_V] = p_{\not f,t}(t) \tag{S6}$$

$$\mathbf{Var}[p_v(t)|p_{\not f,t}(t), s_V] = \frac{p_{\not f,t}(t)\big(1 - p_{\not f,t}(t)\big)}{1 + s_V} \tag{S7}$$

Again, the variance in the correlation between virus isolation and transport exposure of naïve animals depends inversely on scale parameter s_V. As this scale increases, the correspondence increases as given in equation (S7).

For a given month, the desired relations between probabilities related to observed data and probabilities related to unobserved exposures, expressed by equations (S4-S7), can be obtained using the Beta distribution [6], as follows:

$$P\big(p_s(t)|p_f(t), s_S\big) = \text{Beta}\big(p_f(t)s_S, (1 - p_f(t))s_S\big) \ , \tag{S8}$$

$$P\big(p_v(t)|p_{\not f,t}(t), s_V\big) = \text{Beta}\big(p_{\not f,t}(t)s_V, (1 - p_{\not f,t}(t))s_V\big) \ . \tag{S9}$$

The probability of virus isolation and seropositivity for all months considered, $t \in T$ (this includes months with *no* data), can be written as the following products:

$$P\big(\{p_s(t)\}_t|\{p_f(t)\}_t, s_S\big) = \prod_{t\in T}\text{Beta}\big(p_f(t)s_S, (1-p_f(t))s_S\big)\ , \tag{S10}$$

$$P\big(\{p_v(t)\}_t|\{p_{\not f,t}(t)\}_t, s_V\big) = \prod_{t\in T}\text{Beta}\big(p_{\not f,t}(t)s_V, (1-p_{\not f,t}(t))s_V\big)\ . \tag{S11}$$

We can write the probability of seropositivity and virus isolation for all months under consideration, given the complete set of transport and farm exposure probabilities and both scale factors, as the product of equations (S10) and (S11):

$$\begin{aligned}\text{process1} = &P\big(\{p_s(t)\}_t|\{p_f(t)\}_t, s_S\big)\\ &\times P\big(\{p_v(t)\}_t|\{p_{\not f,t}(t)\}_t, s_V\big)\ .\end{aligned} \tag{S12}$$

This grouping, which we call process1, is convenient but arbitrary. Later, this element of the process model will be one factor in the joint distribution over all data and model parameters needed for MCMC sampling of the posterior distribution.

1.2.1 Markov dynamics of process model

In the final major component of the process model, we consider the month-to-month dynamics of (i) transport exposure $p_t(t)$, (ii) farm exposure $p_f(t)$, and (iii) the probability of isolating a particular strain, given successful virus isolation, $p_{i|v}(t)$. We assume these processes obey Markov dynamics, and require that probabilities for consecutive months be the same on average while allowing for gradual change over time. Again, we introduce

scale factors for these associations (now in time) that are inferred as part of the parameter model. For transport exposure this assumption results in the form:

$$\mathbf{E}[p_t(t+1)|p_t(t), s_T] = p_t(t) \ , \tag{S13}$$

$$\mathbf{Var}[p_t(t+1)|p_t(t), s_T] = \frac{p_t(t)(1-p_t(t))}{s_T+1} \ . \tag{S14}$$

For farm exposure we have:

$$\mathbf{E}[p_f(t+1)|p_f(t), s_F] = p_f(t) \tag{S15}$$

$$\mathbf{Var}[p_f(t+1)|p_f(t), s_F] = \frac{p_f(t)(1-p_f(t))}{s_F+1} \ . \tag{S16}$$

And, finally for isolation types we have:

$$\mathbf{E}[p_{i|v}(t+1)|p_{i|v}(t), s_I] = p_{i|v}(t) \ , \tag{S17}$$

$$\mathbf{Var}[p_{i|v}(t+1)|p_{i|v}(t), s_I] = \frac{p_{i|v}(t)(1-p_{i|v}(t))}{s_I+1} \ , \tag{S18}$$

$$\mathbf{Cov}[p_{i|v}(t+1), p_{j|v}(t+1)|p_{i|v}(t), p_{j|v}(t), s_I] = -\frac{p_{i|v}(t)p_{j|v}(t)}{s_I+1}. \tag{S19}$$

The covariance term applies for $i \neq j$ and reflects the fact that an increase in the probability of isolating one strain, decreases the probability of finding a different strain (given a fixed overall probability of virus isolation).

We obtain the desired Markov properties using a Beta distribution to describe the temporal evolution of transport and farm exposure probabilities:

$$P\big(p_t(t+1)|p_t(t), s_T\big) = \mathrm{Beta}\big(p_t(t)s_T, (1-p_t(t))s_T\big) \ , \tag{S20}$$

$$P\big(p_f(t+1)|p_f(t), s_F\big) = \mathrm{Beta}\big(p_f(t)s_F, (1-p_f(t))s_F\big) \ , \tag{S21}$$

and introduce a multinomial generalization of the Beta distribution, called the Dirichlet distribution [6], to reflect the strain-specific dynamics with more than two outcomes:

$$P\big(p_{i|v}(t+1)|p_{i|v}(t), s_I\big) = \mathrm{Dir}\big(\{p_{i|v}(t) \times s_I\}_i\big) \ . \tag{S22}$$

The probabilities obey the relation $\sum_i p_{i|v}(t) = 1$ for all t. This simply means that, given a successful virus isolation, the isolate must be one of the types i under consideration.

The formulations in equations (S20-S22) give the distribution of each probability parameter each month, where the expected value is the parameter value last month and the variance depends on the appropriate scale factor. Given an initial value for the process at $t = 0$ (a value for the month before data is available) and an appropriate scale factor, the complete set of probabilities can be written using the Markov property:

$$P\big(\{p_t(t)\}_t|p_t(0),s_T\big)=\prod_{t\in 0,T}P\big(p_t(t+1)|p_t(t),s_T\big)\ , \tag{S23}$$

$$P\big(\{p_f(t)\}_t|p_f(0),s_F\big)=\prod_{t\in 0,T}P\big(p_f(t+1)|p_F(t),s_F\big)\ , \tag{S24}$$

$$\begin{aligned}P\big(\{p_{i|v}(t)\}_{i,t}|\{p_{i|v}(0)\}_i,s_I\big)=&\\ &\prod_{t\in 0,T}P\big(\{p_{i|v}(t+1)\}_i|\{p_{i|v}(t)\}_i,s_I\big)\ ,\end{aligned} \tag{S25}$$

Again, we group this part of the process model and assign a useful but arbitrary label process2:

$$\begin{aligned}\text{process2}=&P\big(\{p_t(t)\}_t|p_t(0),s_T\big)\\ &\times P\big(\{p_f(t)\}_t|p_f(0),s_F\big)\\ &\times P\big(\{p_{i|v}(t)\}_{i,t}|\{p_{i|v}(0)\}_i,s_I\big)\ .\end{aligned} \tag{S26}$$

This factor provides the probability of the Markov dynamics for transport exposure, farm exposure and isolation of strain i given successful virus isolation. The product of the likelihood, process1 and process2 provides the joint probability of all observed data and all process parameters, given a set of initial conditions and values for the scale parameters as defined in the next section. Note also that the product of process1 and process2 provides both $\{p_v(t)\}_t$ and $\{p_{i|v}(t)\}_{i,t}$ so that the product $p_{i,v}(t)=p_v(t)\times p_{i|v}(t)$ is defined for all strain types and months, as needed for the likelihood factor.

1.3 Parameter model

The final element of the BSSM is to specify prior distributions for the parameters that influence the process model. These parameters include (i) the starting values of the parameters that obey Markov dynamics and (ii) the scale parameters that control variance in the relation between parameters. The goal here is to provide diffuse, non-informative priors and let the data and constraints of the model determine appropriate values.

The initial conditions for the transport and farm exposure are assigned uniform Beta distributions and the initial condition for the strain-specific isolation frequencies is given by a uniform Dirichlet distribution:

$$P(p_t(0)) = \text{Beta}(1,1) \ , \tag{S27}$$

$$P(p_f(0)) = \text{Beta}(1,1) \ , \tag{S28}$$

$$P(\{p_{i|v}(0)\}_i) = \text{Dir}(\{1\}_i) \ . \tag{S29}$$

These are uniform distributions for the parameters over the simplex of appropriate dimension.

Finally, we choose a Pareto distribution as the prior for all of the scale parameters, allowing for a long-tailed distribution of positive values, greater than one:

$$P(s_*) = \text{Pareto}(1,1) \ . \tag{S30}$$

Combining all of these elements, the probability for all initial conditions and scale parameters is given by:

$$\begin{aligned}\text{parameters} = {} & P(p_t(0))\, P(p_f(0))\, P(\{p_{i|v}(0)\}_i) \\ & \times P(s_S)\, P(s_V)\, P(s_T)\, P(s_F)\, P(s_I)\ . \end{aligned} \tag{S31}$$

1.4 MCMC sampling

The posterior distribution for all parameters in the BSSM developed above can be sampled using Markov Chain Monte Carlo methods. The posterior is proportional to the joint distribution over all data and model parameters. This joint distribution is given by the product of factors making up the data, process and parameter models as follows:

$$\begin{aligned}\text{posterior} \propto {} & \text{likelihood} \times \text{process1} \\ & \times \text{process2} \times \text{parameters}\ , \end{aligned} \tag{S32}$$

where the elements are specified by equations (S3, S12, S26, S31). As is usual in MCMC sampling, we can ignore the constant of proportionality and use the joint distribution defined above to obtain samples from the posterior.

The Bayesian state-space model, as detailed in this supplement, was implemented in Python using the PyMC package [3]. The model was burned in for 100 000 updates and an additional 200 000 iterations were collected, thinning by a factor of twenty to produce 10 000 samples from the posterior distribution. Throughout the main manuscript, the posterior mean and regions of 95% high probability density from the MCMC sampling are presented. PyMC implements the Geweke test for convergence of the MCMC sampling [7]. Under this test, comparing twenty sections of the MCMC traces for each parameter,

all estimates were determined to have converged.

References

1 Clark, J. S. & Bjornstad, O. N. 2004 Population time series: process variability, observation errors, missing values, lags, and hidden states. *Ecology*. **85**, 3140–3150. (doi:10.1890/03-0520)

2 West, M., Harrison, P. J. & Migon, H. S. 1985 Dynamic generalized linear models and Bayesian forecasting. *J Am Stat Assoc*. **80**, 73–83.

3 Patil, A., Huard, D. & Fonnesbeck, C. J. 2010 PyMC: Bayesian stochastic modelling in Python. *J Stat Softw*. **35**, 1–81.

4 Van Reeth, K., Labarque, G. & Pensaert, M. 2006 Serological profiles after consecutive experimental infections of pigs with European H1N1, H3N2, and H1N2 swine influenza viruses. *Viral Immunol*. **19**, 373–382. (doi:10.1089/vim.2006.19.373)

5 Vijaykrishna D. *et al.* 2011 Long-term evolution and transmission dynamics of swine influenza A virus. *Nature*. **473**, 519–522. (doi:10.1038/nature10004)

6 Wilks, S. S. 1962 *Mathematical Statistics*. New York: John Wiley & Sons, Inc.

7 Geweke, J. 1992 Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In: *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith) pp. 169 – 193. Oxford University Press.

Scale	Estimate	95% HPD
s_I	374	(244, 490)
s_F	228	(15, 1 004)
s_T	11 410	(216, 39 770)
s_S	51	(4, 186)
s_V	52	(32,72)

Table S1: Posterior mean and 95% HPD for scale parameters inferred using MCMC sampling of the posterior distribution as described in the electronic supplementary material.

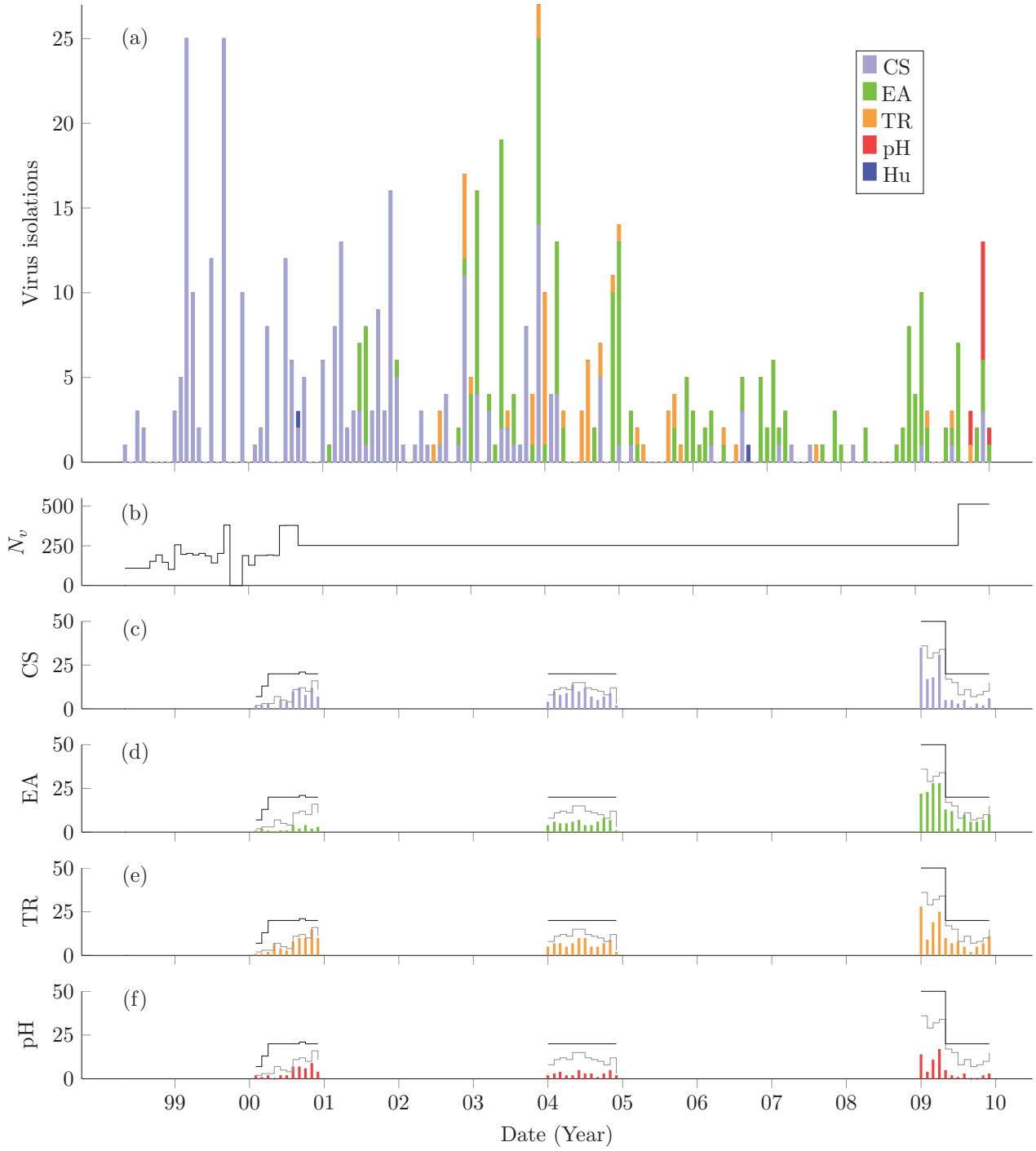

Figure S1: [Color online] (a) Raw virus isolation counts, by strain type and (b) number of attempted virus isolations each month. (c-f) Raw counts of the number of individual pigs that have HI tests showing titer greater than 1:40 to the specific test antigen. An individual pig can be 'positive' to one or more strain types by this test. The gray line shows the number of positives to *any* type and the black line shows the total number of samples taken.

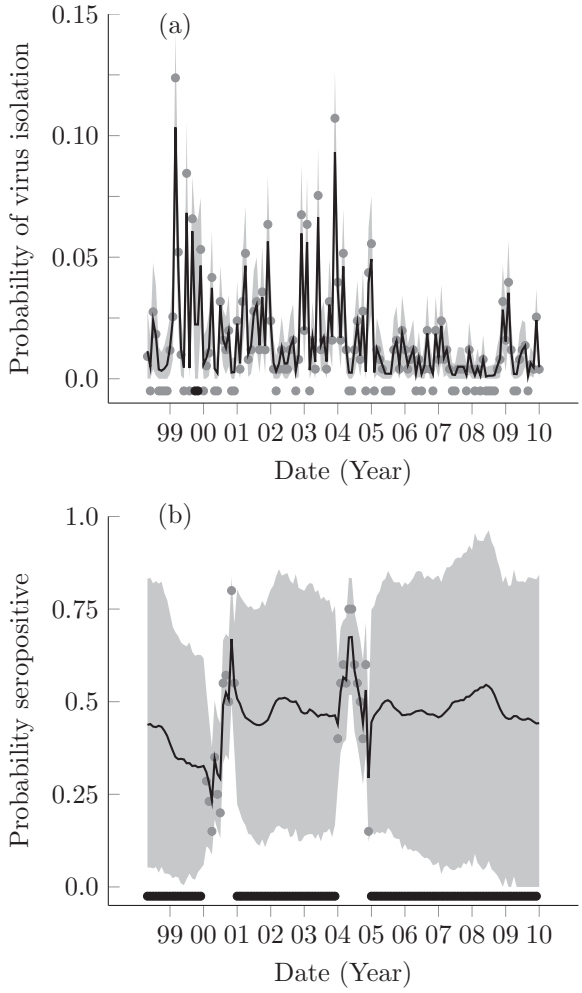

Figure S2: A fit of the Bayesian state-space model without 2009 serology data (compare with figure 2a and 2c in main manuscript). The lack of serology data after 2004 results in an increasing uncertainty in the estimation of $p_s(t)$, reflected by the gray region of 95% HPD. This fit does not have an increasing posterior mean for the probability of seropositivity (panel b, solid line) but the uncertainty does not rule out this possibility.

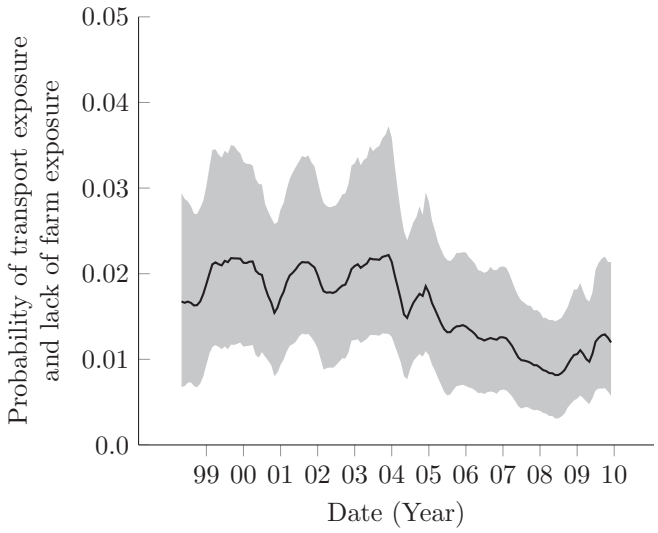

Figure S3: The probability of exposure during transport for swine that have no prior immunity. This probability is estimated by the product of probabilities corresponding to a lack of previous exposure on the farm $(1 - p_f(t))$ and exposure during transport $p_t(t)$. This probability is equal, on average, to the probability of virus isolation with the variance described by scale factor s_V.

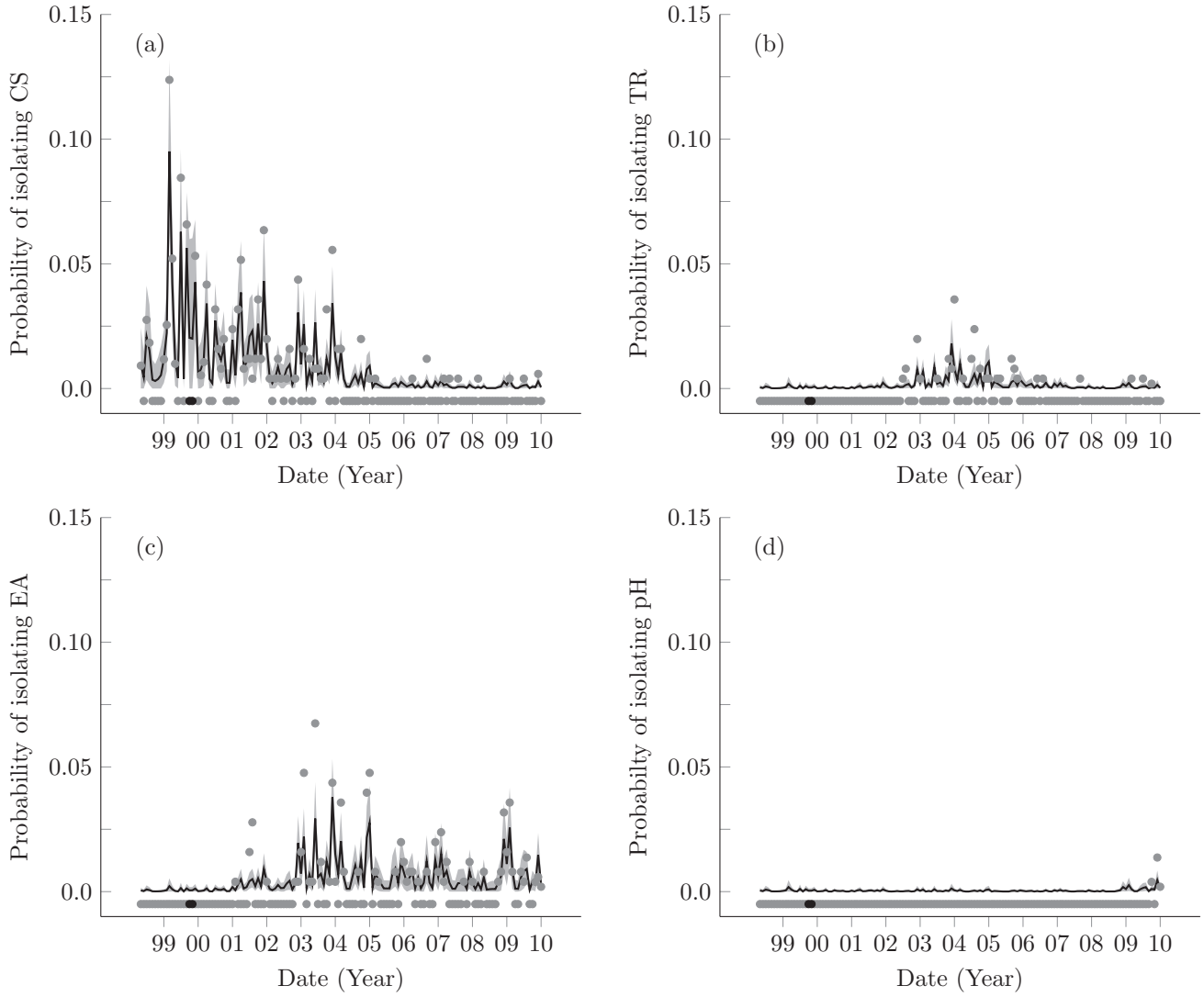

Figure S4: (a) The probability of isolating the classical swine (b) triple reassortant (c) Eurasian avian-like and (d) pandemic H1N1 viral strains. The probability of isolating human seasonal H1N1 is included in the model but is not plotted here because there are only two isolations. In all panels, the black lines show the posterior mean and gray shading indicates the region of 95% HPD. Gray dots show the maximum likelihood estimates for the probability of isolating the focus strain (estimates equal to zero are show below the zero-line for visual clarity). Black dots indicate missing data and are also plotted below the zero-line. Note that the sum of the plotted strain-specific virus isolation rates $p_{i,v}(t)$ equals the total isolation probability $\sum_i p_{i,v}(t) = p_v(t)$ (See main text figure 2a).

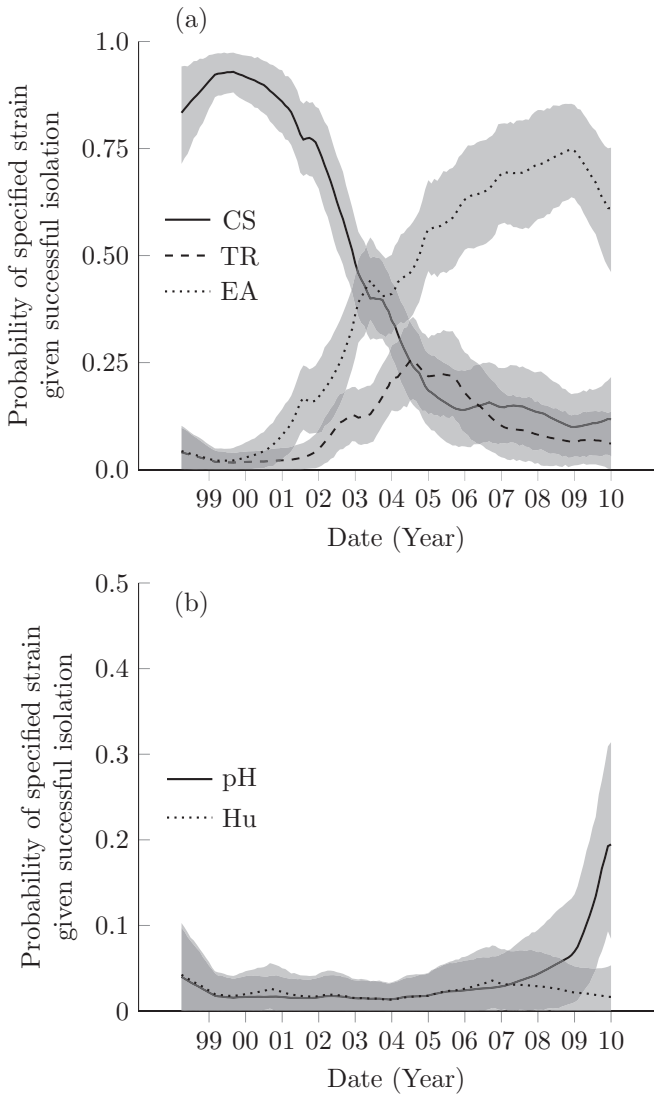

Figure S5: The probability of isolating strain type $i \in \{\mathrm{CS}, \mathrm{EA}, \mathrm{TR}, \mathrm{Hu}, \mathrm{pH}\}$, given that an isolation has occurred. (a) The three dominant strains are plotted, showing a clear switch from classic swine to Eurasian avian-like. (b) Pandemic and human seasonal H1N1 are plotted separately due to the lower probability of isolation (note the difference in scale between panels a and b).