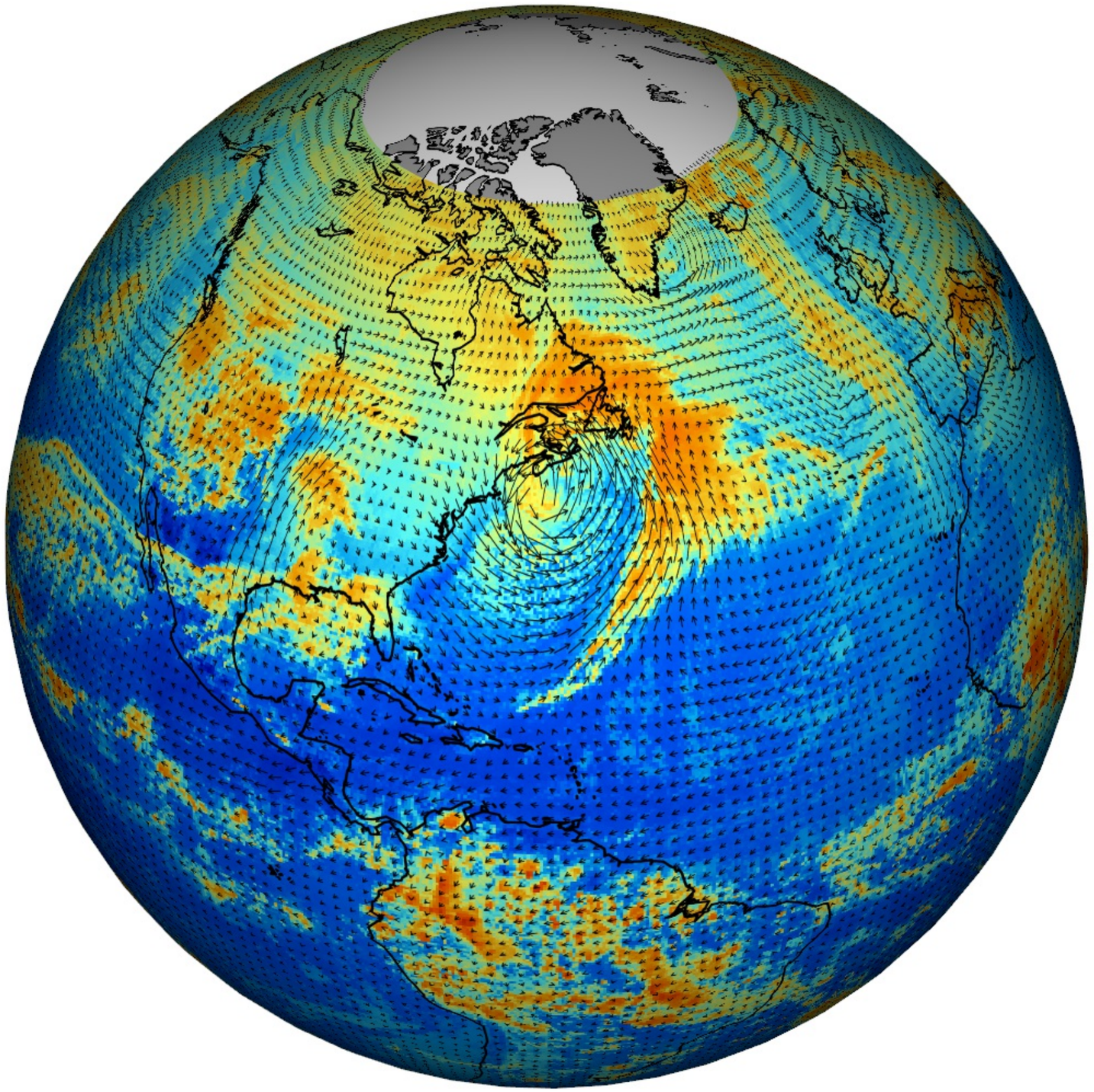
Forecasting

Forecasting is a big problem



General, overlapping approaches

Machine learning

Statistical models

Nonlinear forecasting

"Mechanistic" modeling

DREAM challenges

Inference of gene regulatory networks

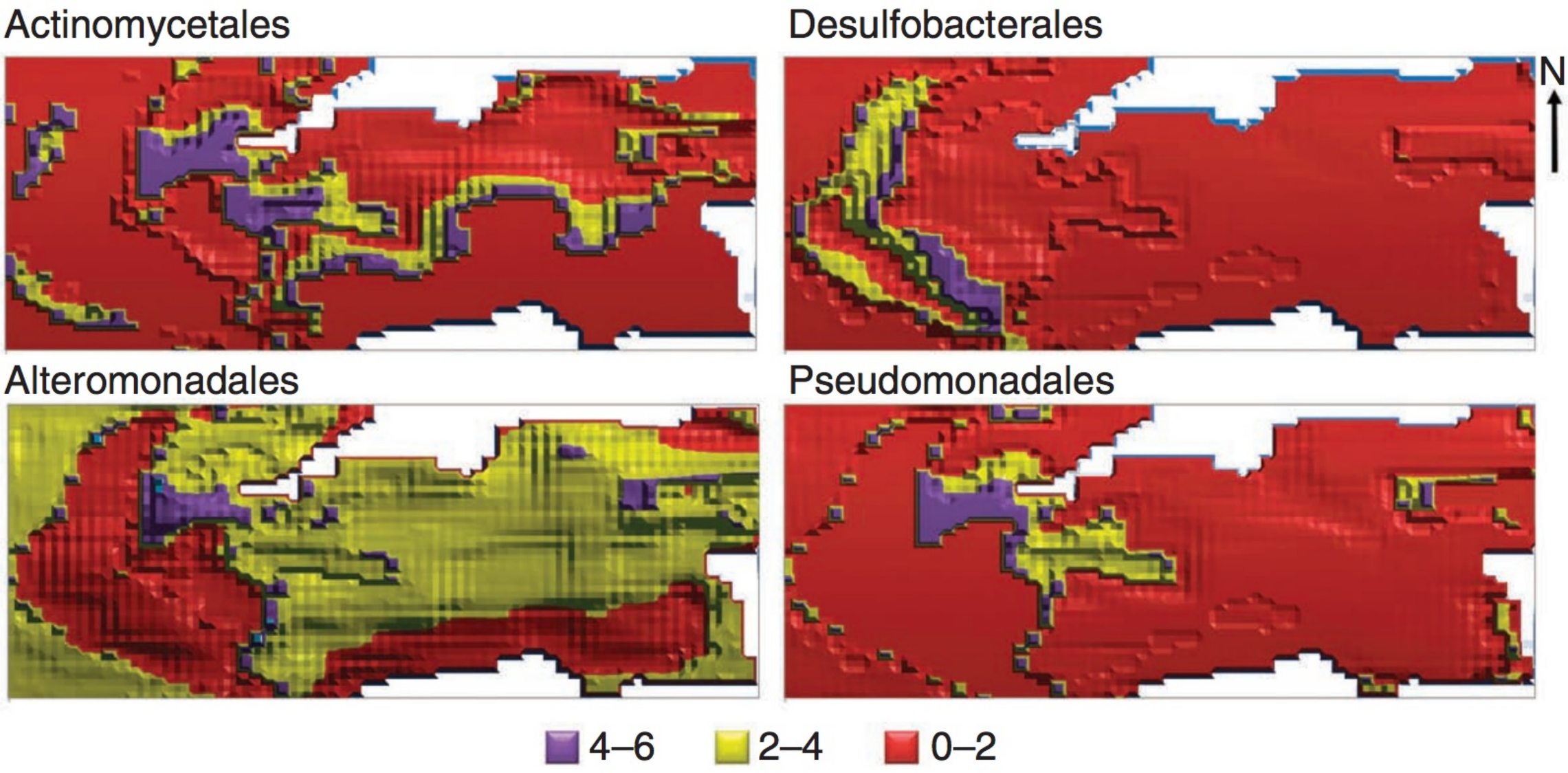
from knockout, observational, and synthetic data

		Artificial			*E. coli*				*S. cerevisiae*			
Method	Reference (abbrev.)	D5:100k	D5	Nar2011	D5:100k	D5	M3D	Nar2011	D5:100k	D5	M3D	Nar2011
1. ANOVA η^2	this paper	78.0^a	81.6^c		**67.1**^{a}	**74.6**^{c}	**79.8**^{d}		51.8^a	57.8^c	55.0^d	
2. Genie3	Huynh-Thu.. (2010)	**81.5**^{a}	**83.4**^{c}		61.7^a	69.0^c	67.3^d		51.8^a	54.5^c	51.3^d	
3. Team 395	unpublished	69.5^a			60.2^a				**53.9**^{a}			
4. Pearsons ρ^2	Butte.. (1999)	75.7^b	76.5^c		57.2^b	61.2^c	64.6^d		51.0^b	56.9^c	53.8^d	
5. MRNet	Meyer.. (2007)	71.5^b	73.0^c		58.1^b	66.2^c	64.5^d		50.9^b	52.2^c	52.3^d	
6. CLR	Faith.. (2007)	76.2^b	77.4^c	76.2^e	59.1^b	66.1^c	64.2^d	64.0^e	51.6^b	52.6^c	52.4^d	50.9^e
7. ARACNe	Margolin.. (2006)	76.3^b	77.5^c	76.7^e	57.2^b	64.2^c	63.5^d	64.4^e	50.4^b	51.3^c	49.9^d	49.1^e
8. qp graphs	Castelo.. (2009)			69.6^e				63.5^e				54.5^e
9. GeneNet	Opgen-Rhein.. (2007)			52.4^e				59.9^e				**55.2**^{e}

Compete Lasso, random forests, Bayesian networks, mutual information, ANOVA, etc.

Kuffner et al. 2012

Ecological niche modeling



An artificial neural network that included microbial interactions performed best.

Larsen et al. 2012

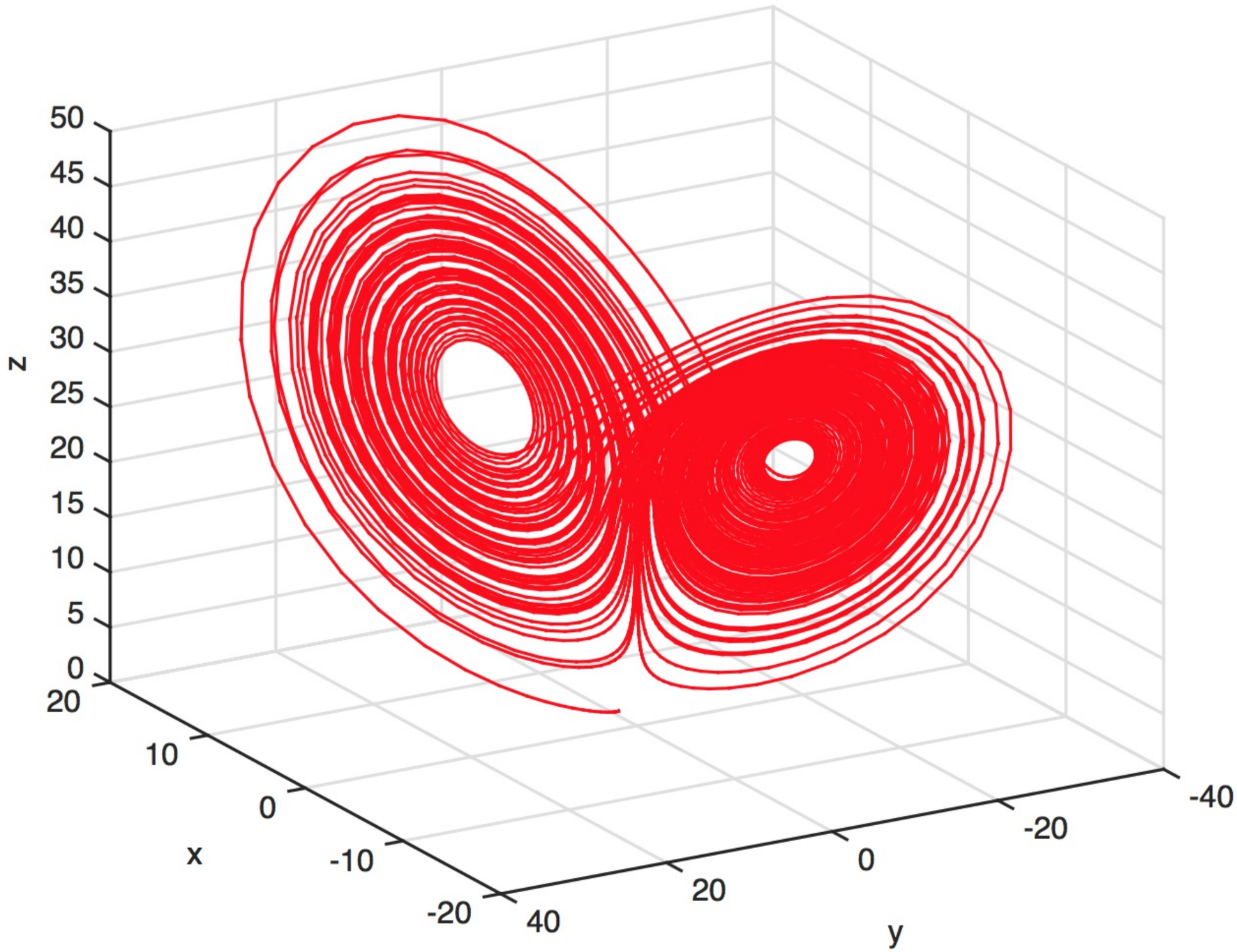
Nonlinear forecasting

Reconstruct attractor ("library") from time series

Use attractor to make short-term predictions

Sugihara and May 1990

Dynamics may be chaotic



Predictions with chaos: short shelf life

Trajectories in chaotic attractors diverge

$$|\delta \mathbf{Z}(t)| \approx e^{\lambda t} |\delta \mathbf{Z}_0|$$

λ is the Lyapunov exponent

(so with chaos, $\lambda > 0$)

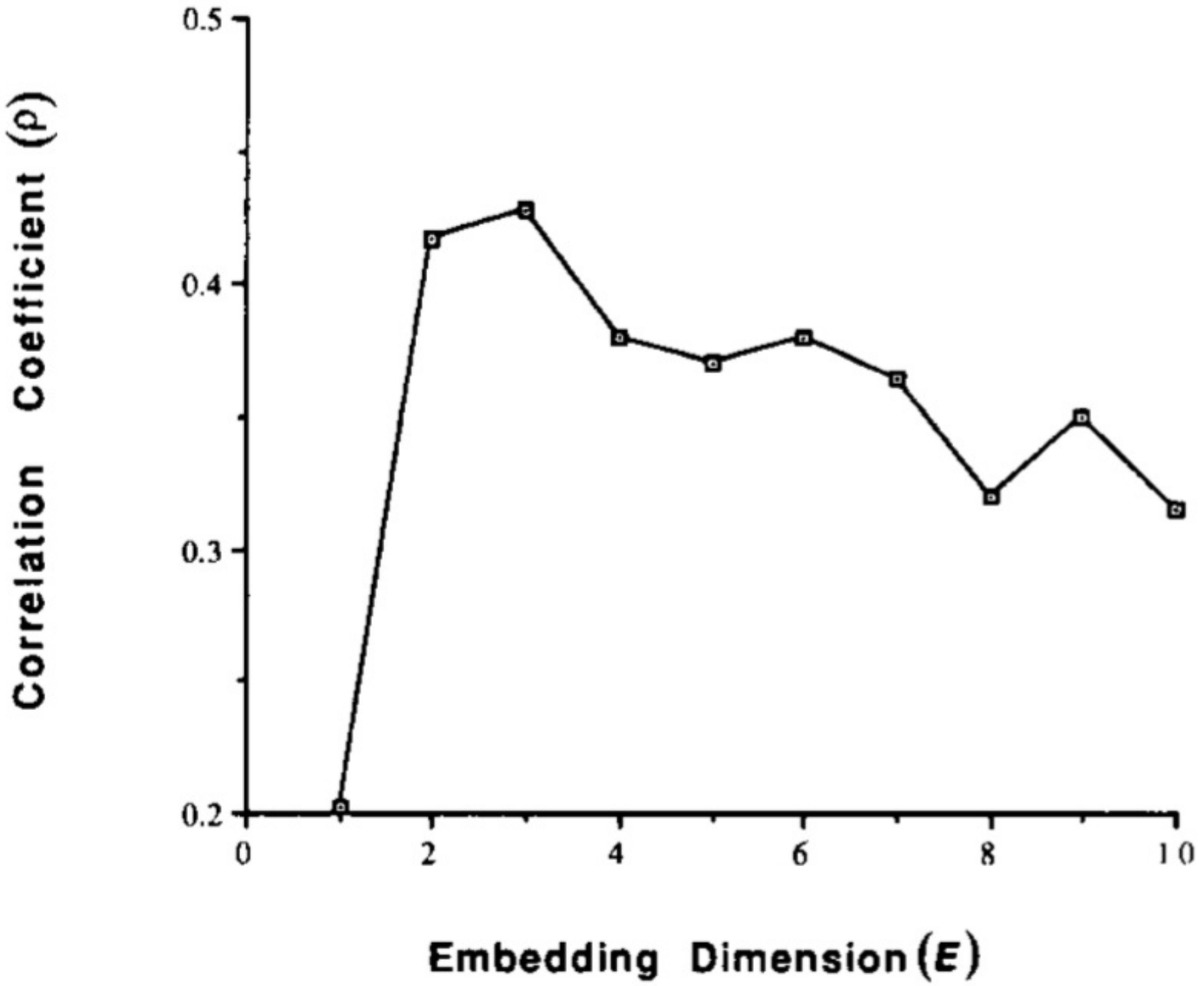
How to forecast

- Choose an embedding dimension E and lag τ
- Each point in E-dimensional space:
  $\{x_t, x_{t-\tau}, x_{t-2\tau}, \ldots, x_{t-(E-1)\tau}\}$
- Construct these points from the time series
- Define a point to predict ("predictee")
- See where predictee's $E+1$ nearest neighbors wind up t steps into the future
- Measure correlations ρ between predictee's observed future state and neighbors' weighted predictions

Choosing E and τ

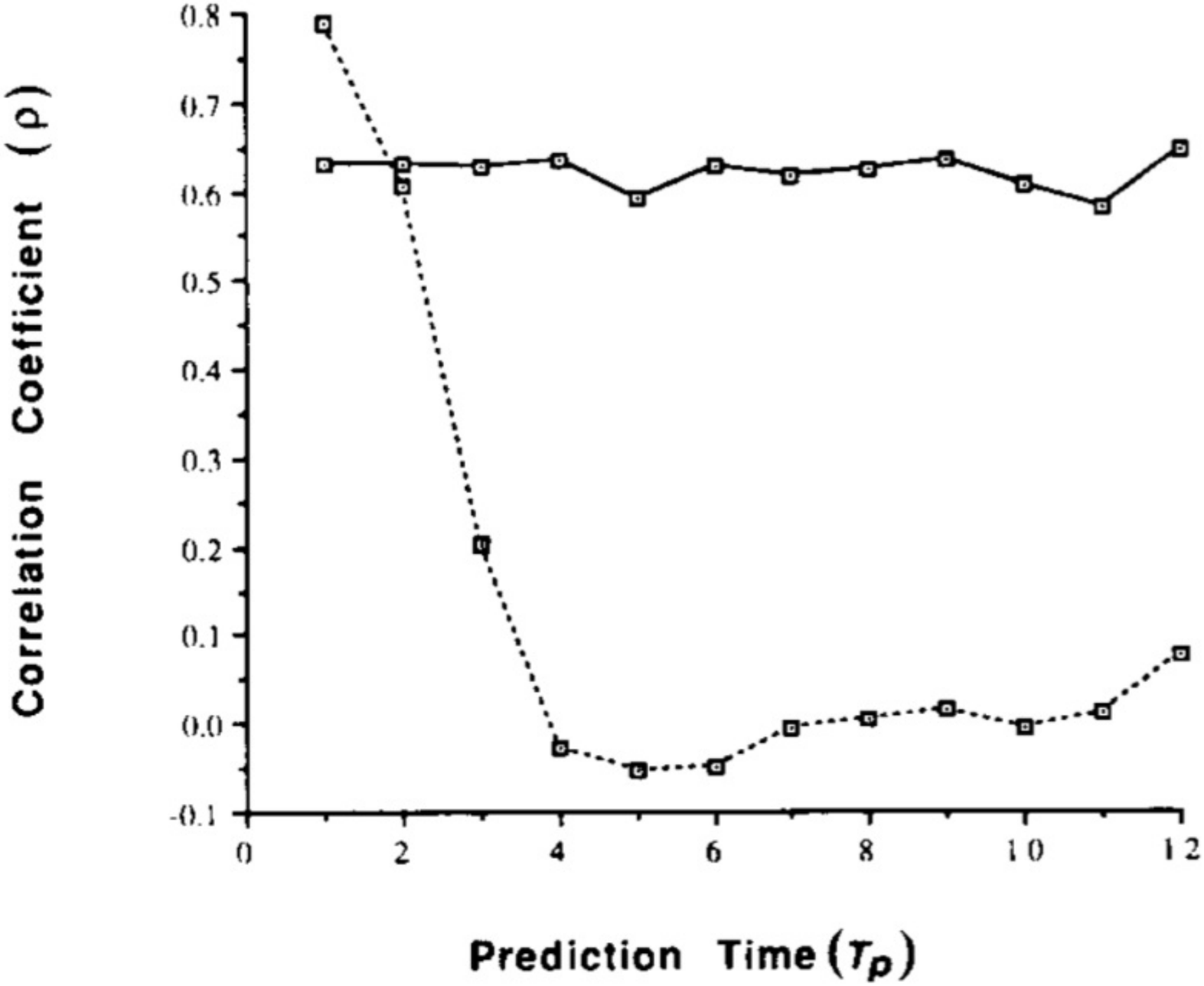
An unsolved problem

Use E (and τ) that yield best predictions



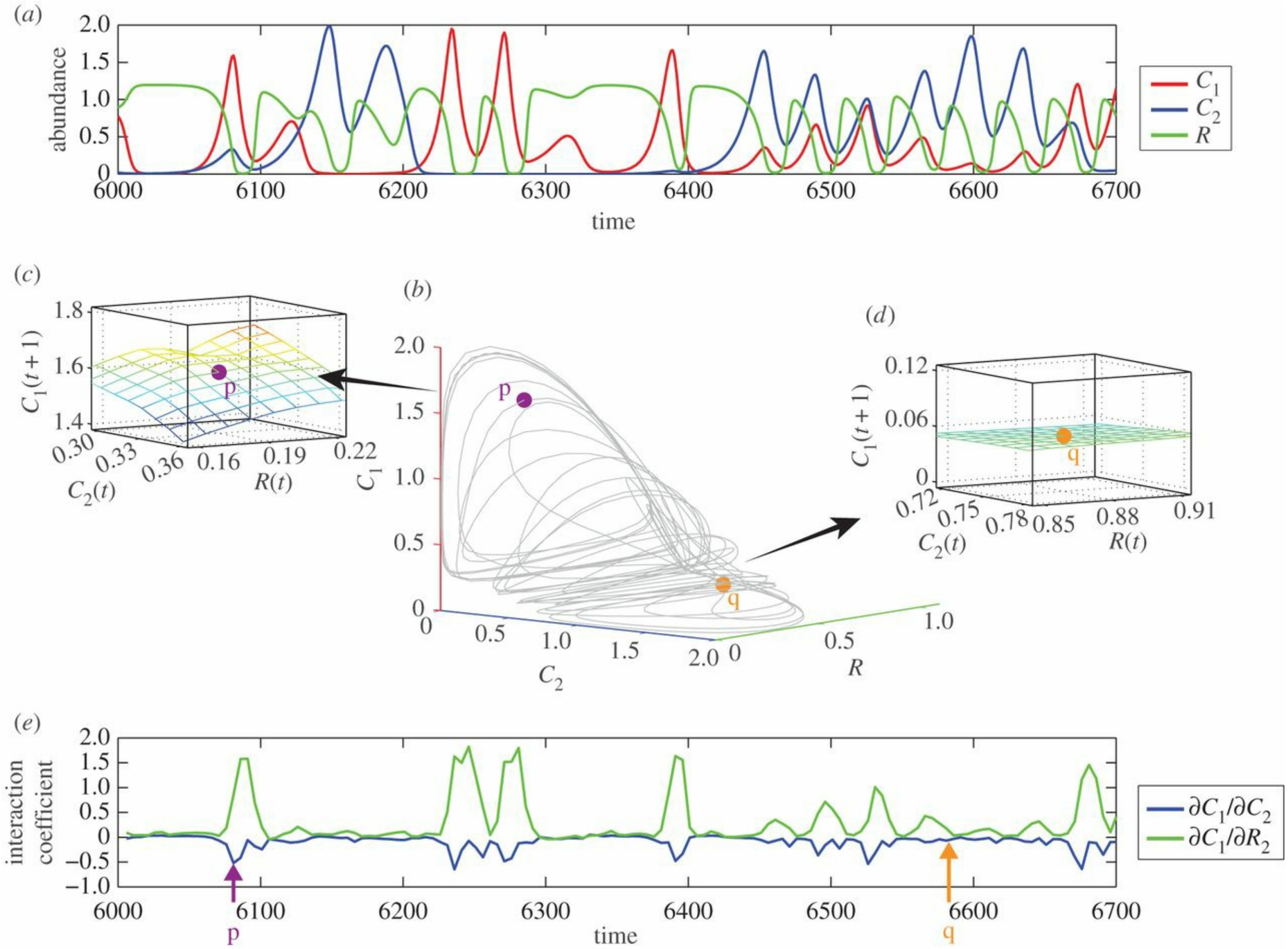
Sugihara and May 1990

Observational noise v. chaos



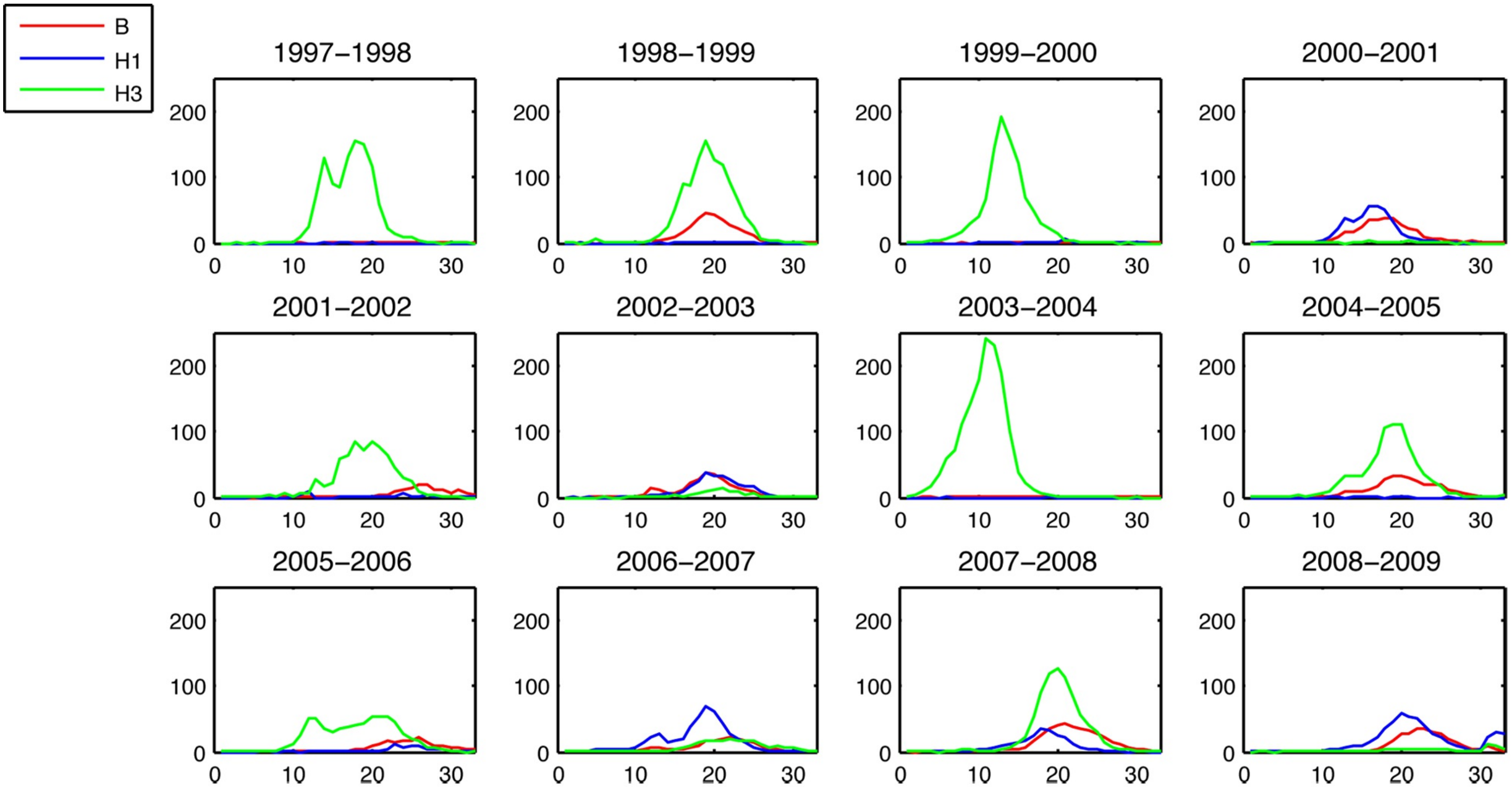
Sugihara and May 1990

Forecasting communities



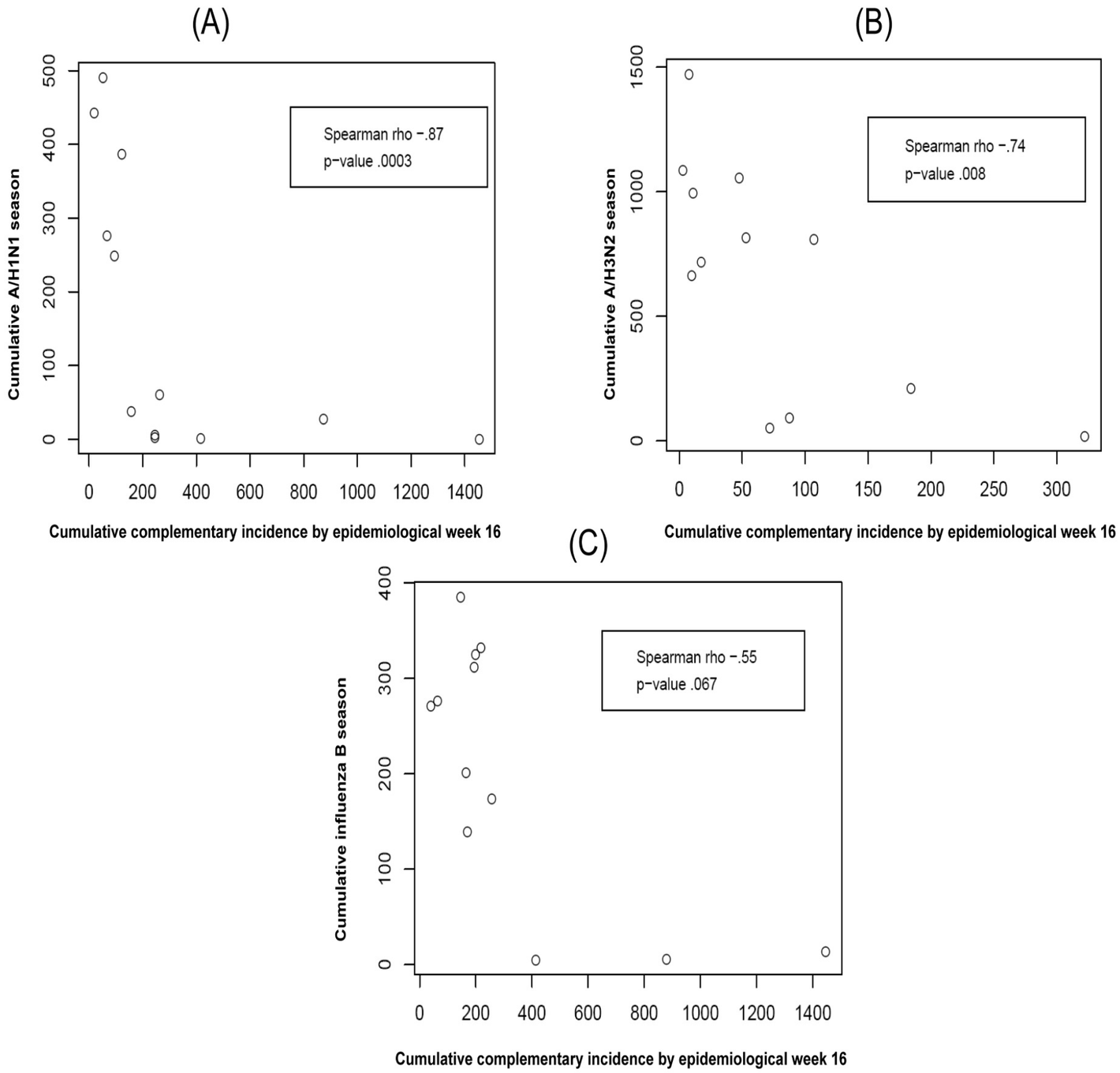
Deyle et al. 2015

Predicting flu



Goldstein et al. 2011

Epidemic sizes negatively correlated



Goldstein et al. 2011

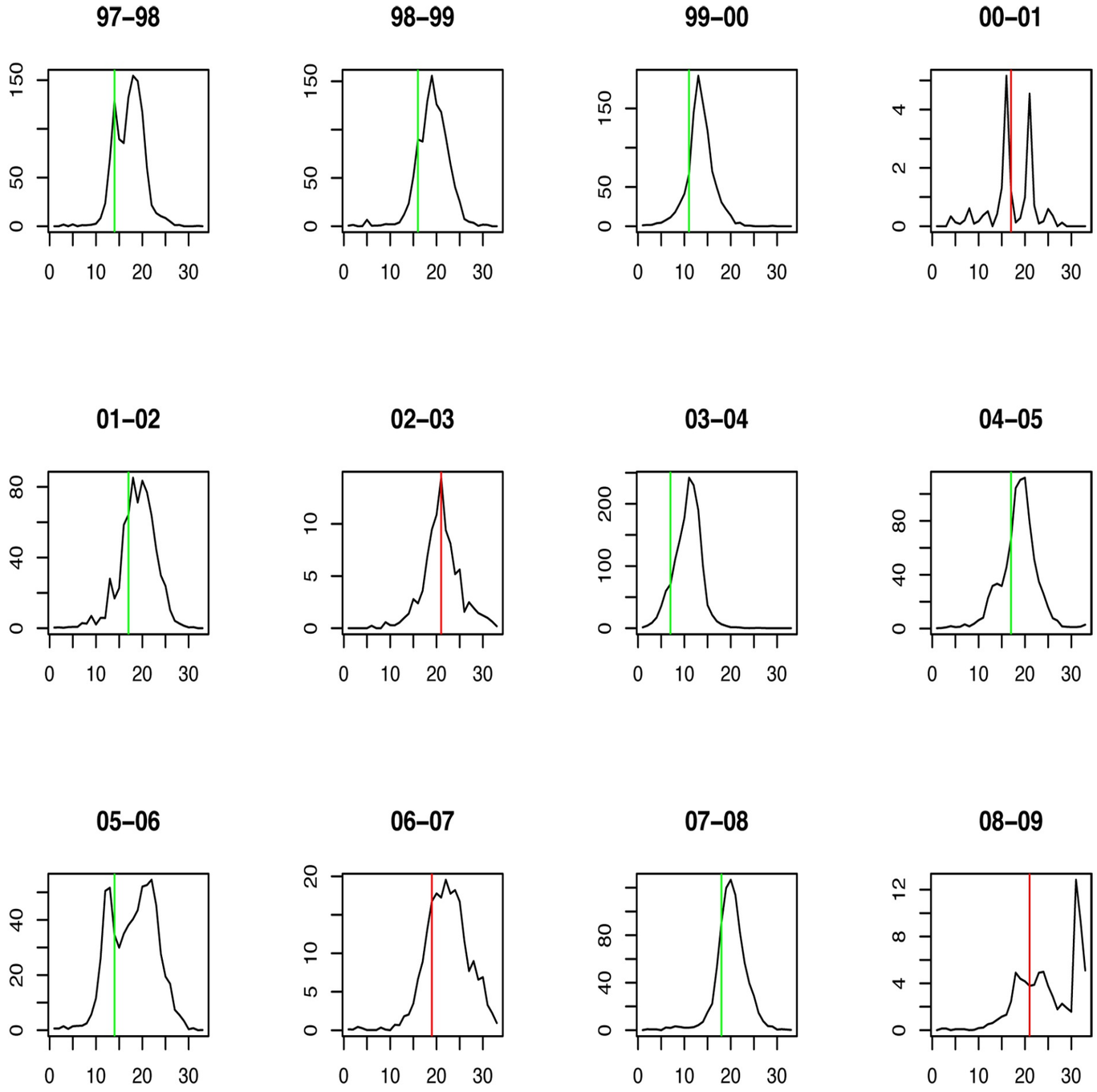
Predict based on cumulative incidence

$$X = \frac{I(s) + I(s-1)}{\max(h, I(s) + \ldots + I(s-4))}$$

$$Y = \beta_0 + \beta_X \cdot X + \beta_T \cdot T$$

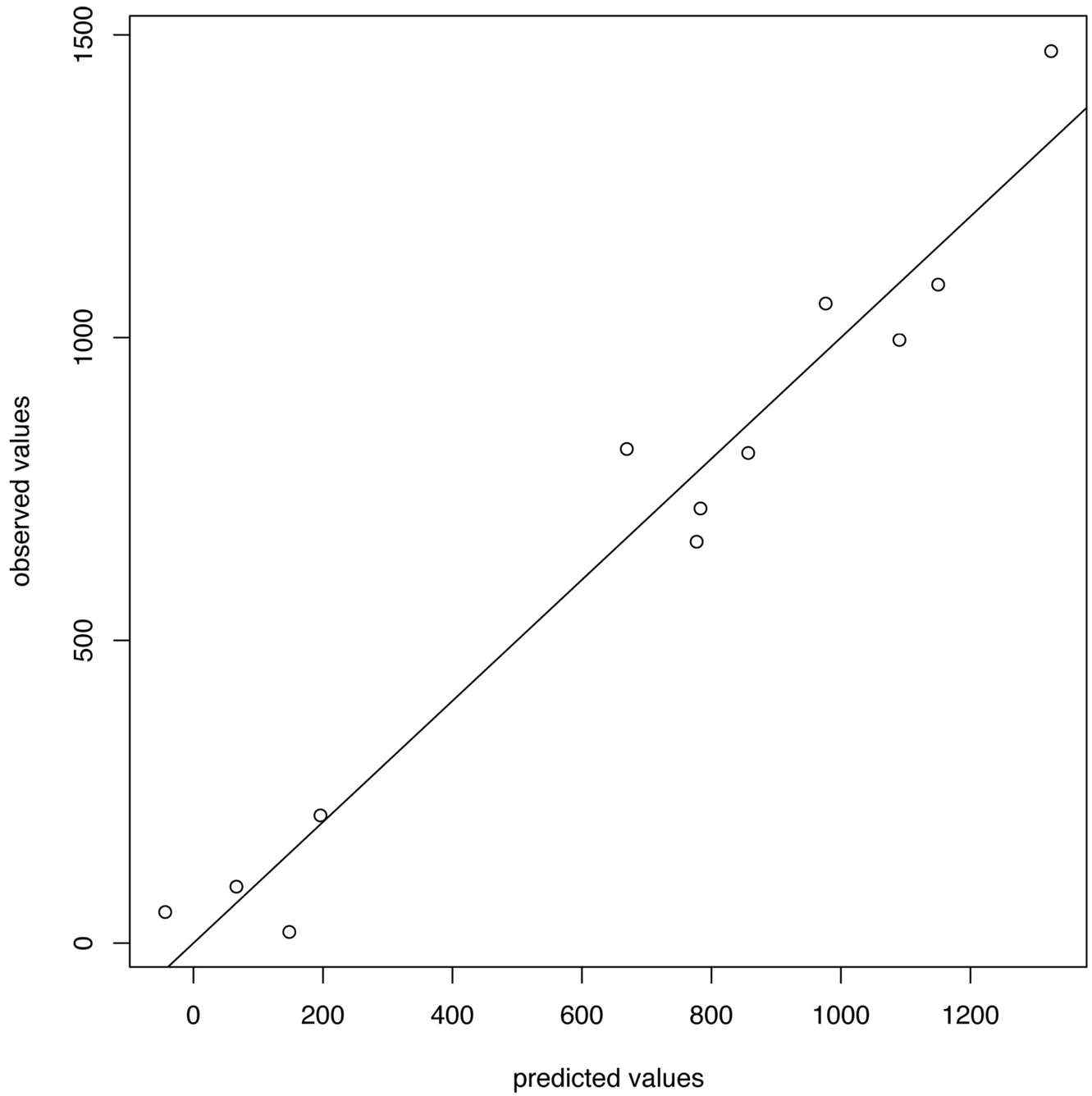
Goldstein et al. 2011

Stopping times for H3N2



Goldstein et al. 2011

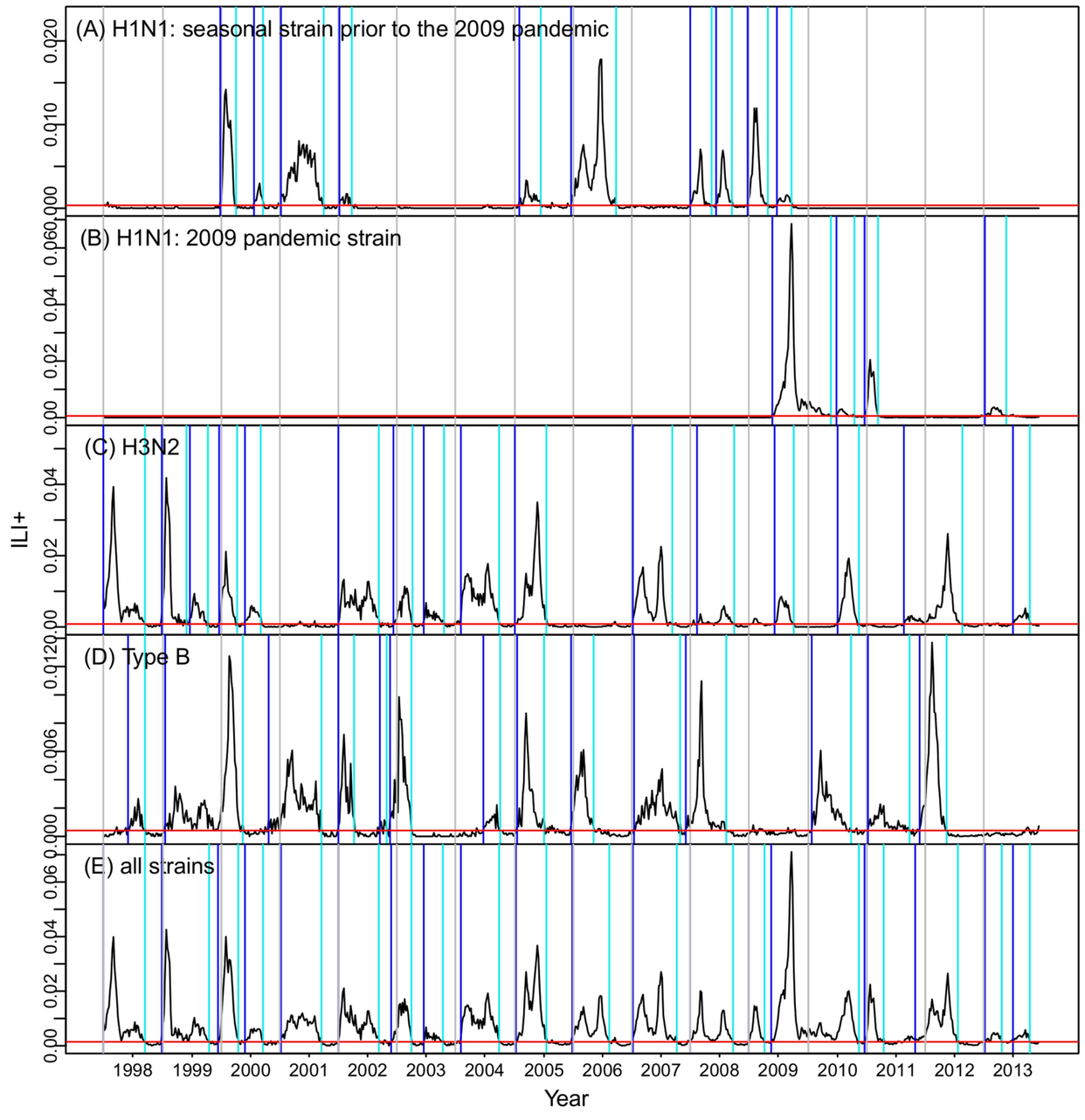
Predictions v. observations (H3N2)



Goldstein et al. 2011

Predicting flu in Hong King

Aim: Predict peak timing and magnitude



Yang et al. 2015

Mechanistic model and particle filter

$$\frac{dS(t)}{dt} = -\frac{R_0}{D} \cdot \frac{I(t)S(t)}{N} - \alpha$$

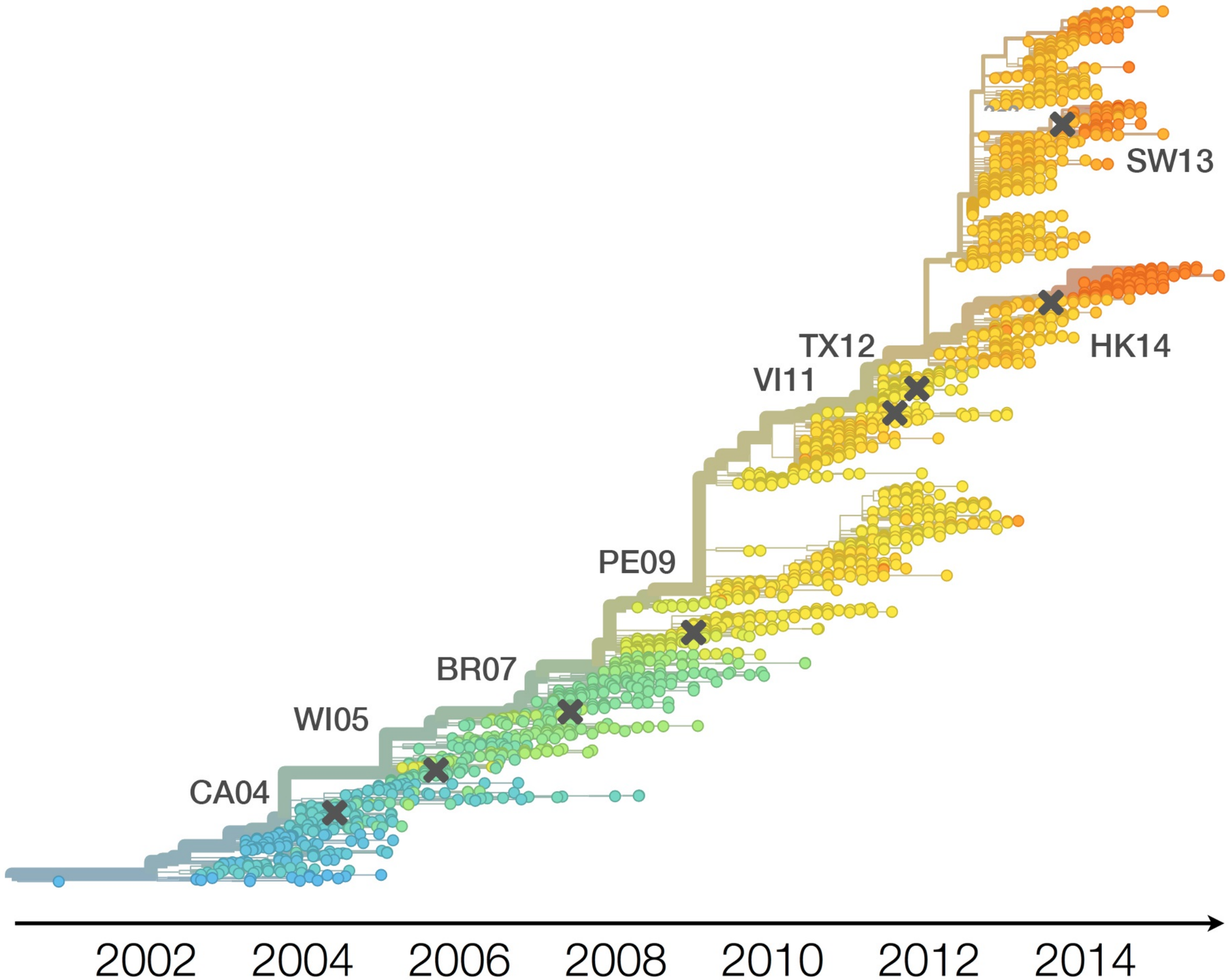
$$\frac{dI(t)}{dt} = \frac{R_0}{D} \cdot \frac{I(t)S(t)}{N} - \frac{I(t)}{D} + \alpha$$

37% accuracy with 1-3 week lead, ~50% at 0 week lead

Yang et al. 2015

FluSight

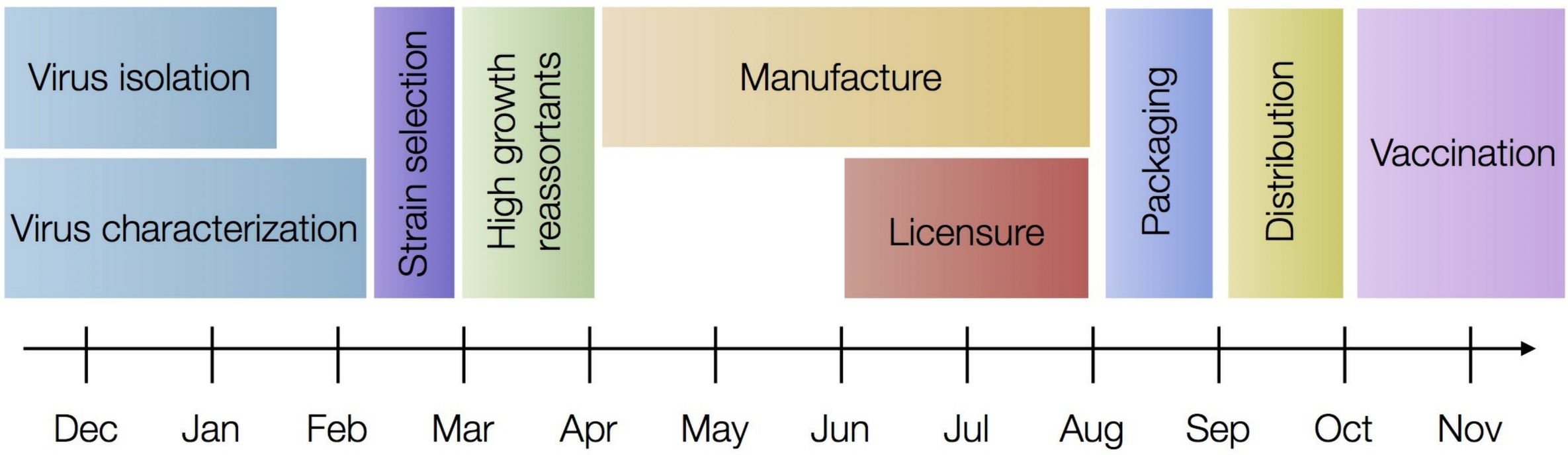
Forecasting flu evolution



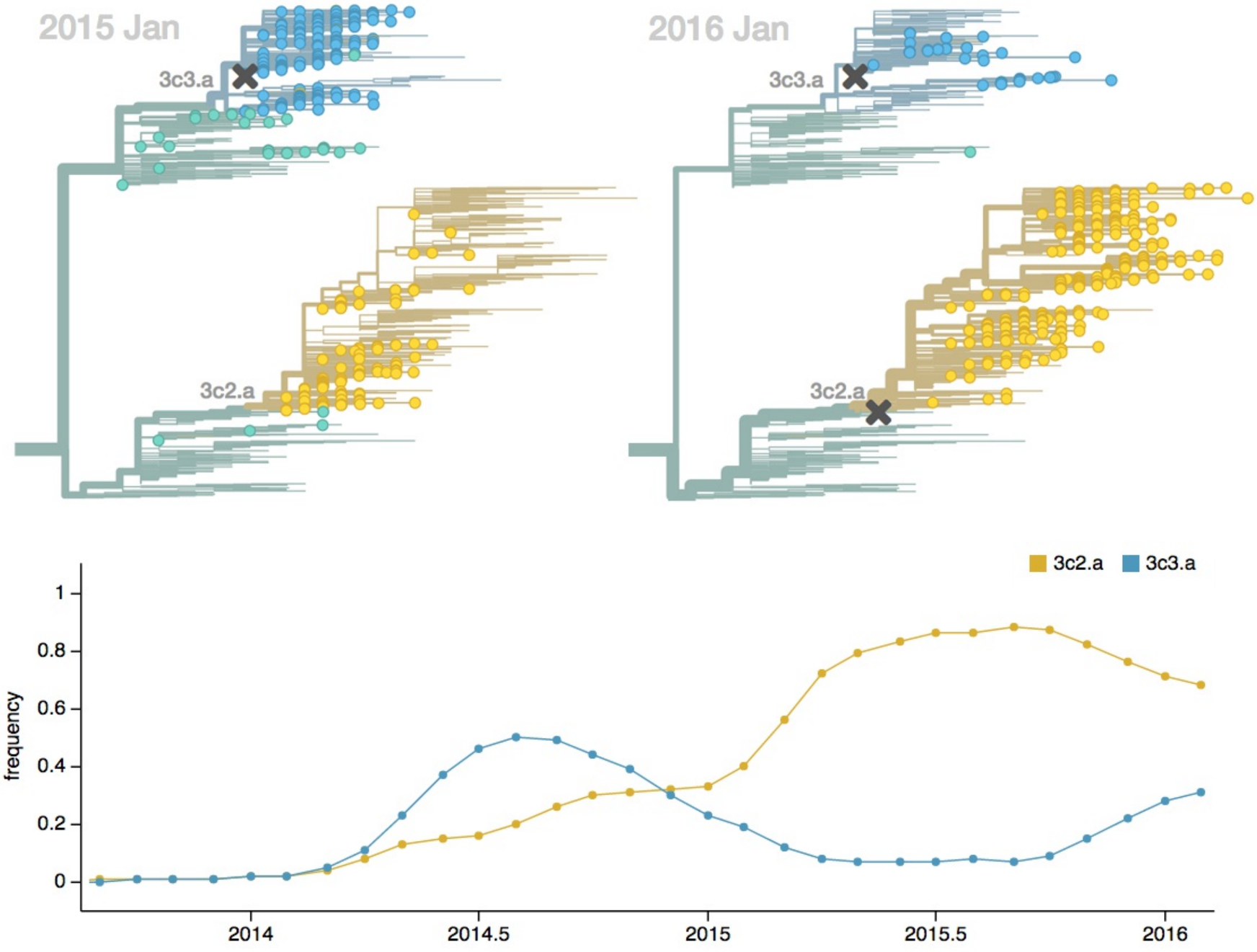
Vaccine strain selection timeline

Collection by WHO National Influenza Centres

Characterization by WHO Collaborating Centres



Seek to explain change in clade frequencies over 1 year

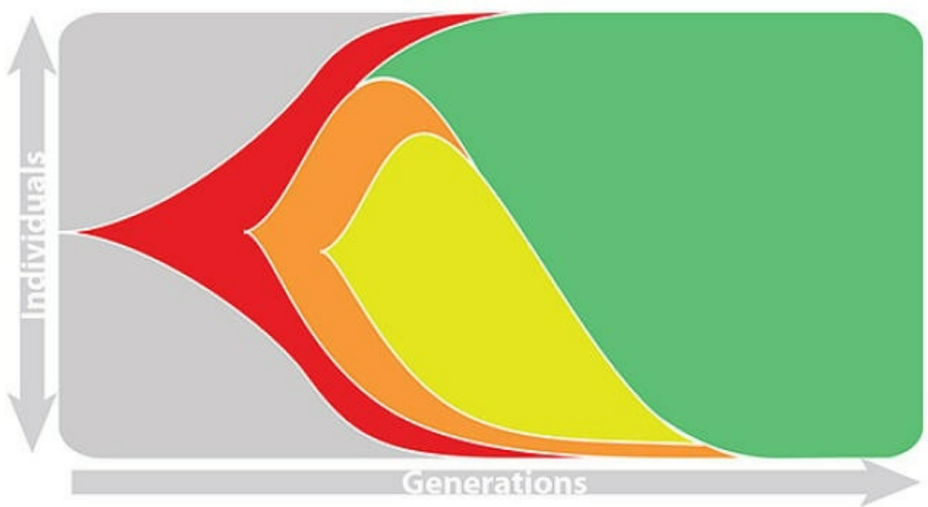


Fitness models can project clade frequencies

Clade frequencies $\boldsymbol{X}$ derive from the fitnesses $\boldsymbol{f}$ and frequencies $\boldsymbol{x}$ of constituent viruses, such that

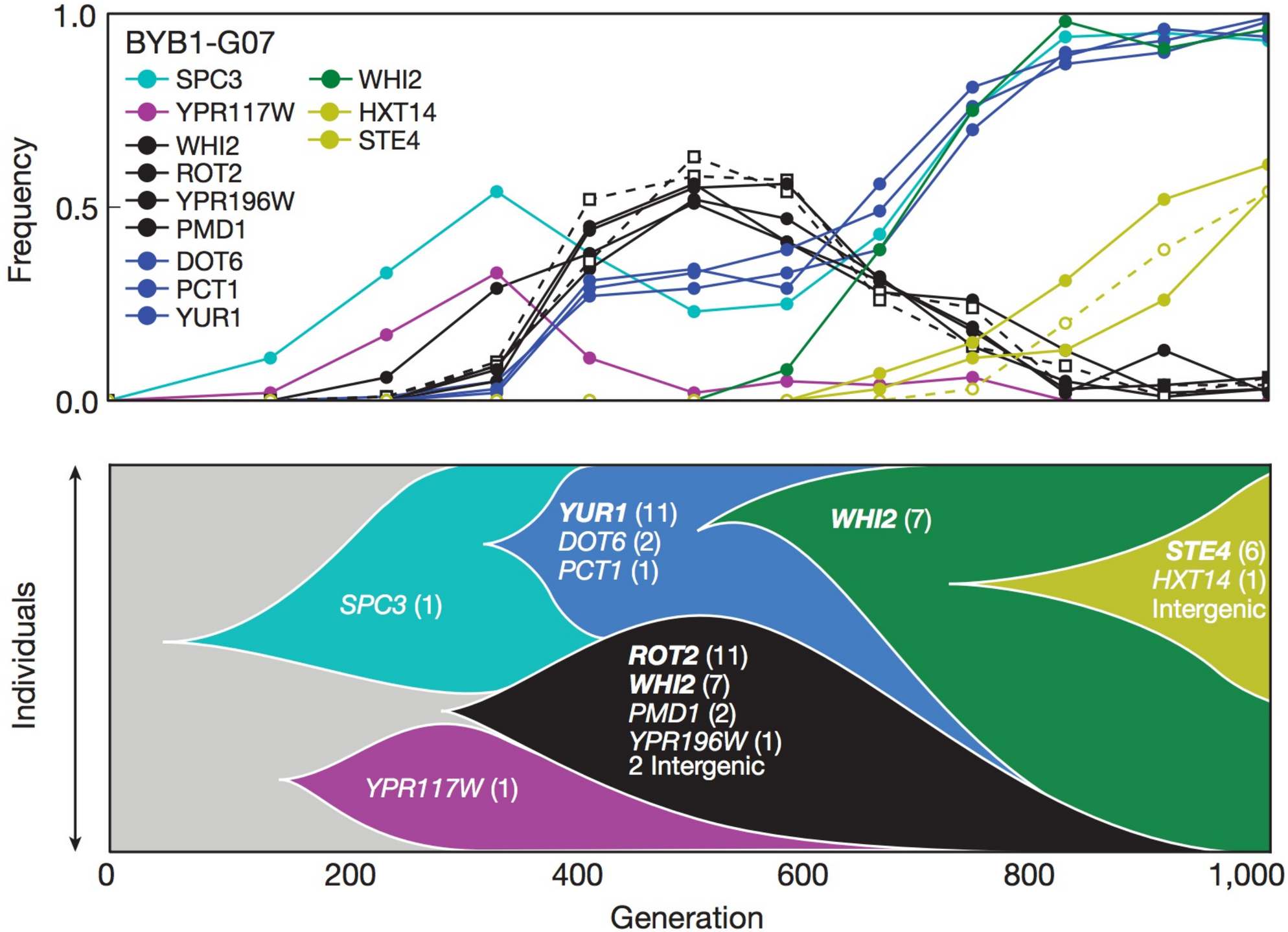
$$\hat{X}_v(t + \Delta t) = \sum_{i:v} x_i(t) \exp(f_i \, \Delta t)$$

This captures clonal interference between competing lineages



1uksza and Lässig 2014

Clonal interference



Predictive fitness models

A simple predictive model estimates the fitness f of virus i as

$$\hat{f}_i = \beta^{\mathbf{ep}} f_i^{\mathbf{ep}} + \beta^{\mathbf{ne}} f_i^{\mathbf{ne}}$$

where $f_i^{\mathbf{ep}}$ measures cross-immunity via substitutions at epitope sites and $f_i^{\mathbf{ne}}$ measures mutational load via substitutions at non-epitope sites

1uksza and Lässig 2014

Model selection

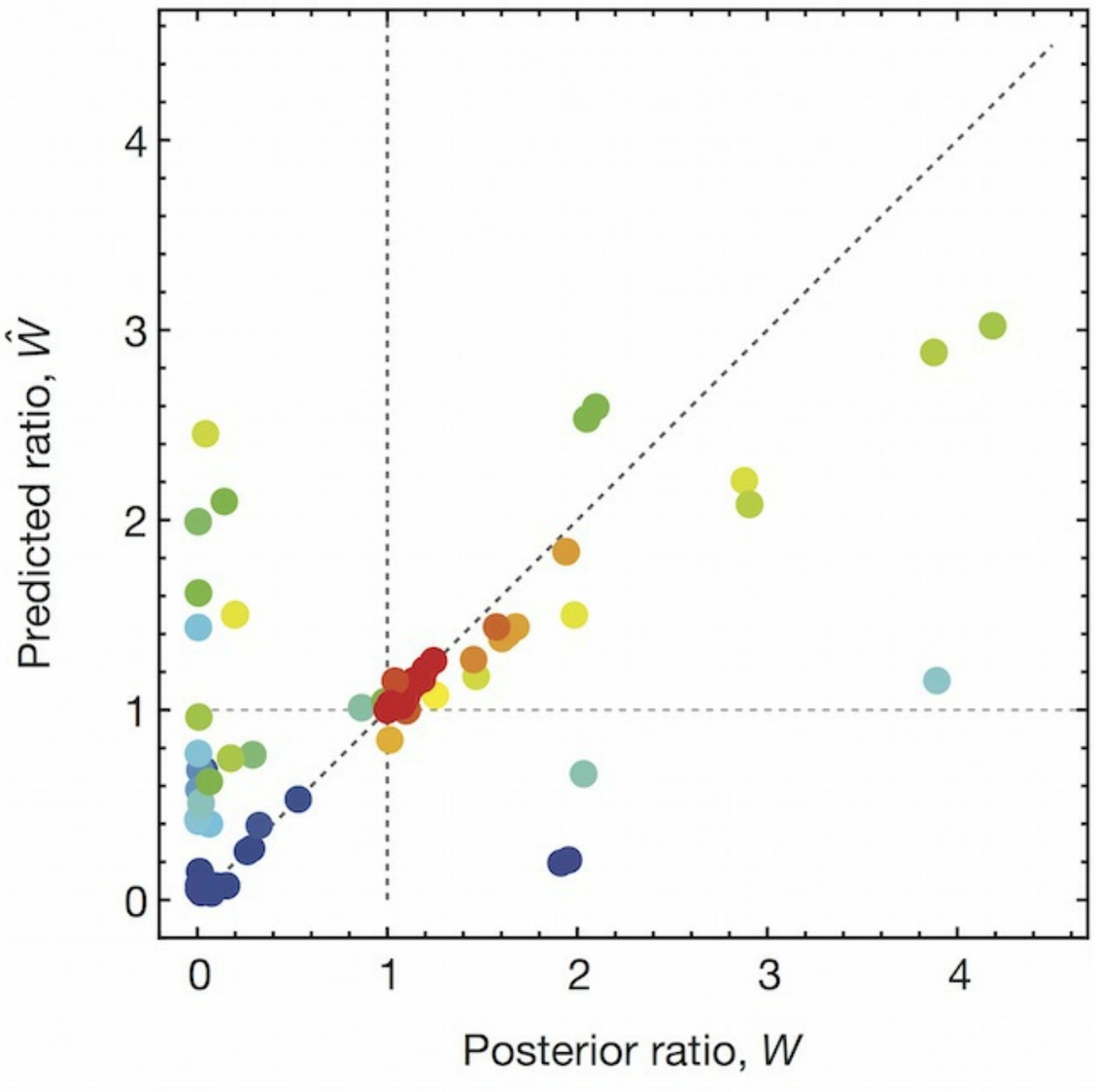
$$f_i = f_i^{\mathrm{ep}} + f_i^{\mathrm{ne}} + f_i^{\mathrm{nl}}$$

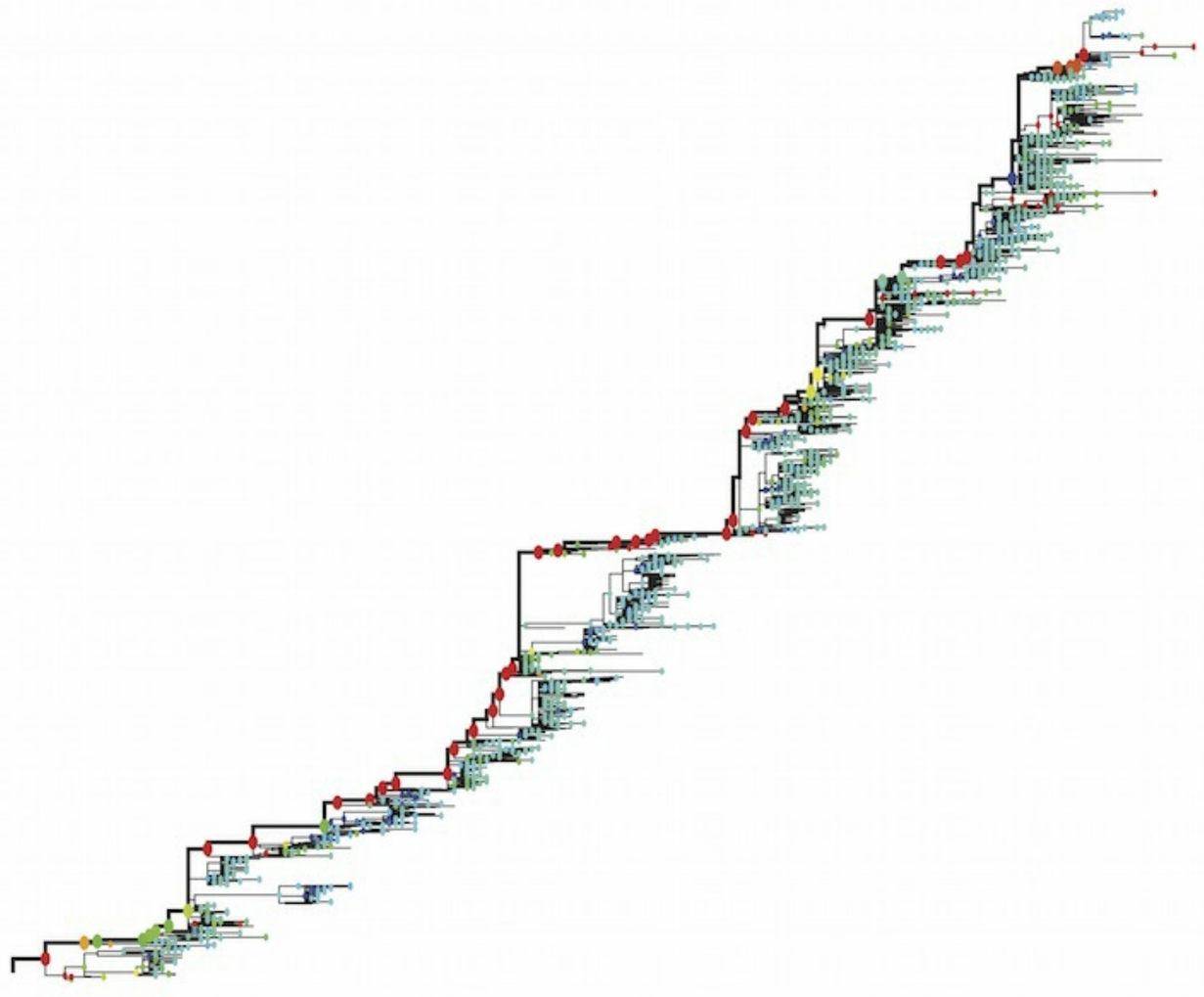
$$f_v^{\mathrm{nl}}(t) = \log\left[\frac{1}{X_v(t)} \sum_{j:t,v} x_j \exp\left[\lambda D_0\left(\mathbf{a}_j, \mathbf{a}^*(t)\right)\right]\right] \equiv D_0^{\mathrm{ave}}\left(v(i), \mathbf{a}^*(t)\right)$$

model	σ^*_{ep}	D^*_0	σ^*_{ne}	λ^*	σ^*_{gl}	$\Delta\mathcal{H}_{\mathrm{tot}}$	$\bar{\Phi}$	V
full haplotype-based	1.15 ± 0.29	14.	-0.5	0.31 ± 0.26	0	541.	26.	2.1
linear	0.52 ± 0.07	0	-0.5	0	0	326.	15.	0.9
epitope variants								
– excl. receptor binding sites [7]	1.06 ± 0.38	14.	-0.5	0.30 ± 0.25	0	451.	23.	1.9
– codon subset [3]	0.37 ± 0.36	14.	-0.5	0.39 ± 0.32	0	187.	14.	1.5
– extended codon set [11]	1.14 ± 0.29	14.	-0.5	0.32 ± 0.26	0	523.	25.	2.1
– incl. glycosylation [42, 48–51]	1.01 ± 0.32	14.	-0.5	0.29 ± 0.24	1.	549.	29.	2.2
epitope-only								
– all codons [7]	1.39 ± 0.50	14.	0	0	0	297.	16.	1.2
– excl. receptor binding sites [7]	1.20 ± 0.53	14.	0	0	0	201.	12.	0.8
– codon subset [3]	0.60 ± 0.44	14.	0	0	0	33.	2.	0.1
– extended codon set [11]	1.36 ± 0.49	14.	0	0	0	270.	15.	1.1
– only glycosylation [42, 48–51]	0	0	0	0	1.	136.	5.	0.2
– random codons	0.34 ± 0.36	14.	0	0	0	-69.	1.	.04

1uksza and Lässig 2014

Performance





1uksza and Lässig 2014

What other models would you test?

What limits prediction with other pathogens?