Case studies

Odd Kolbjørnsen & Geir Storvik

May 08 2017

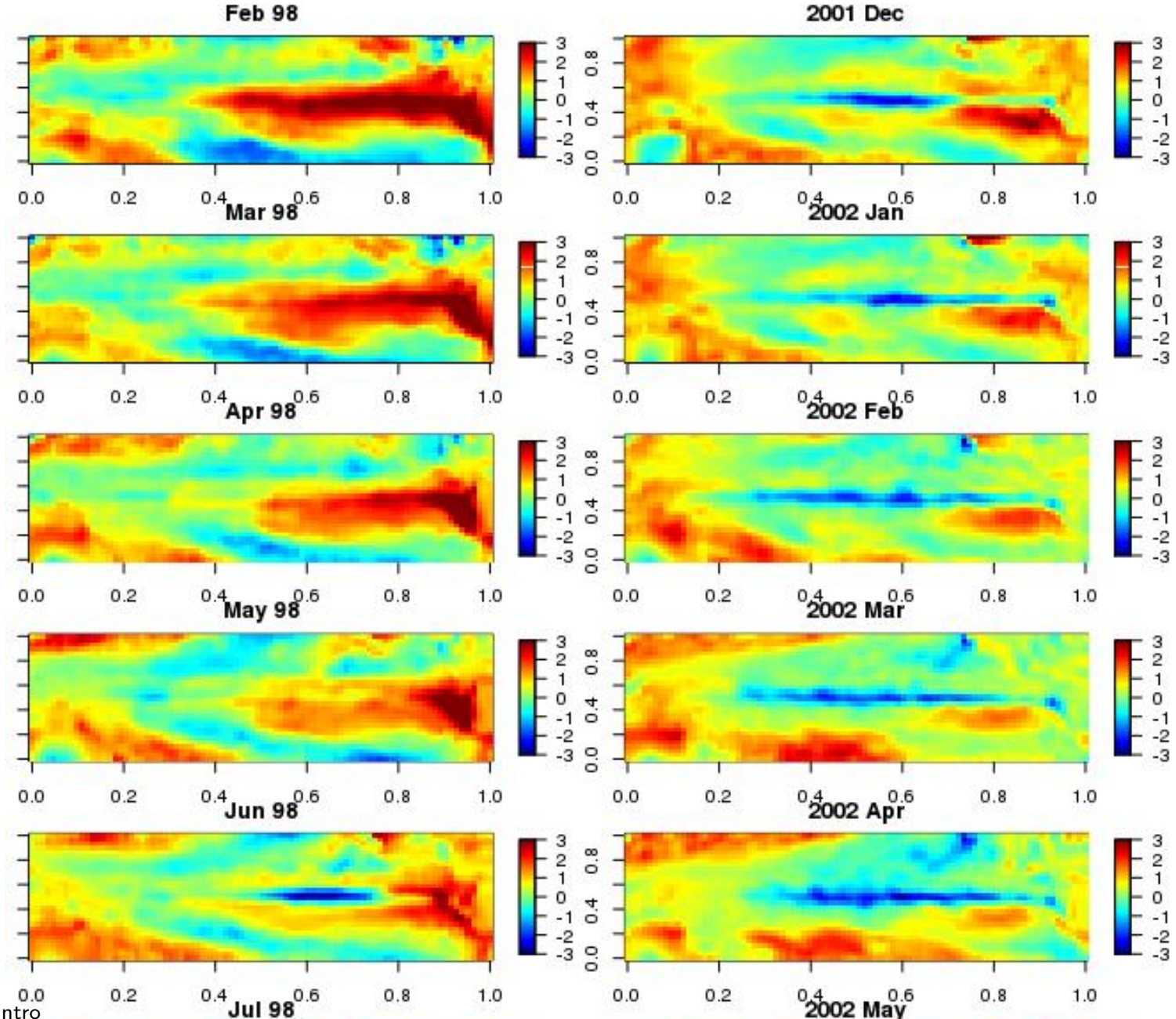
Long-lead forecasting of tropical pacific seas surface temperature

- Already considered the SST data several times
- On interannual time scales, the dominant features in the process are the so-called El Niño and La Niña phenomena
- Cressie and Wikle (2011):

  > *due to its complexity and uncertainties related to the fundamental physical mechanisms that drive these phenomena, SST is one of the few processes in oceanography in which "statistica" long-lead forecast models perform as well as or better than deterministic models*

- Now: Modelling through reduced-dimension
- Both linear and "nonlinear" dynamics
- Aim: 6 month ahead forecast

SST


Feb 98
2001 Dec
Mar 98
2002 Jan
Apr 98
2002 Feb
May 98
2002 Mar
Jun 98
2002 Apr
Jul 98
2002 May

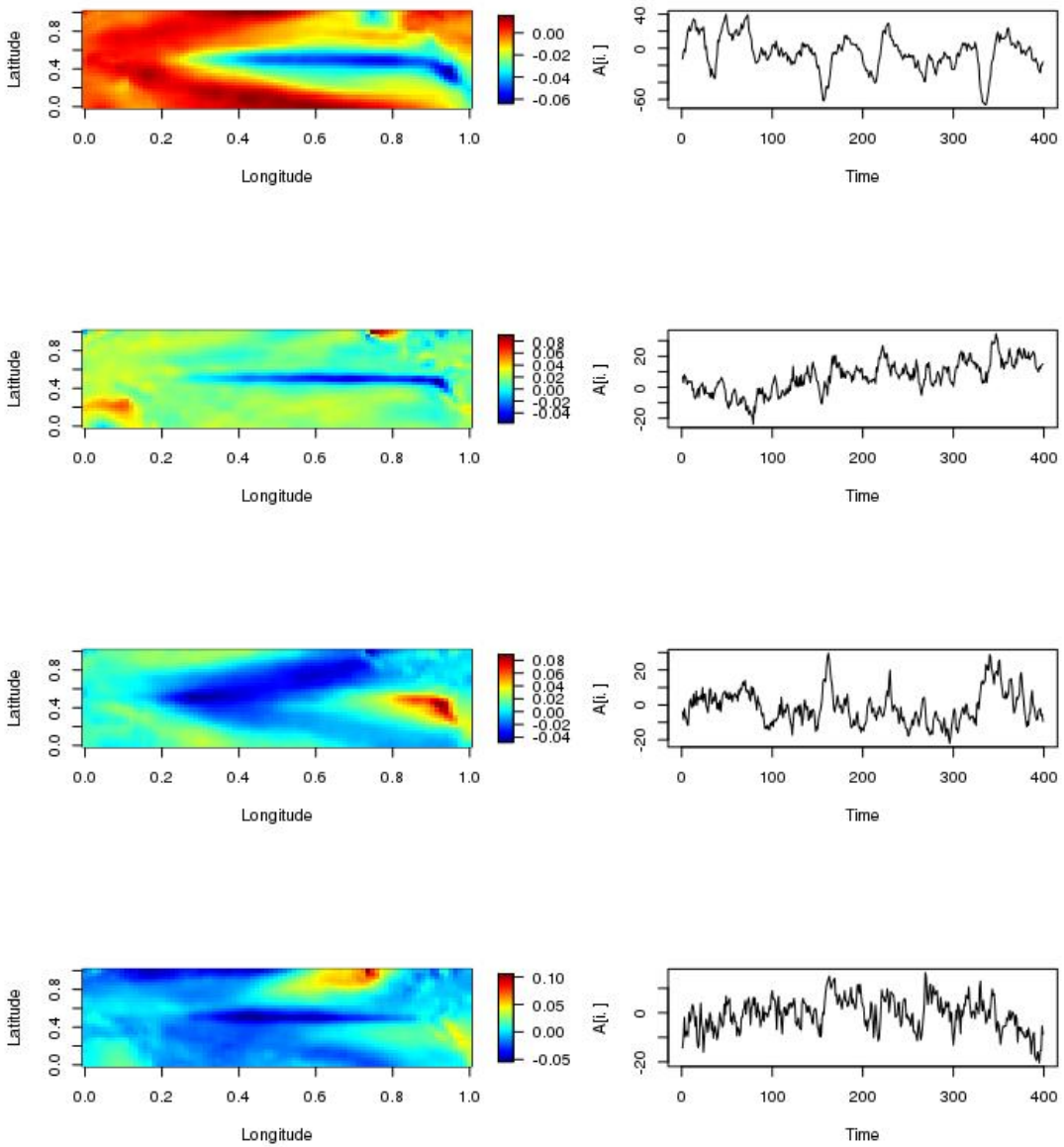
Reduced-dimension linear DSTM

Model

$$\begin{aligned} \mathbf{Z}_t &= \mathbf{Y}_t + \boldsymbol{\nu}_t, & \boldsymbol{\nu}_t &\overset{ind}{\sim} N(\mathbf{0}, \sigma_\nu^2 \mathbf{I}) \\ \mathbf{Y}_t &= \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \boldsymbol{\gamma}_t, & \boldsymbol{\gamma}_t &\overset{ind}{\sim} N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}) \\ \boldsymbol{\alpha}_t &= \mathbf{M}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, & \boldsymbol{\nu}_t &\overset{ind}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_\eta) \end{aligned}$$

Of interest: $\mathbf{Y}_{t+\tau}^p = \boldsymbol{\Phi}\boldsymbol{\alpha}_{t+\tau}$.
Note: $E[\boldsymbol{\Phi}\boldsymbol{\Phi}^T\mathbf{Z}_{t+\tau}|\boldsymbol{\alpha}] = \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\alpha}_{t+\tau} = \boldsymbol{\Phi}\boldsymbol{\alpha}_{t+\tau} = \mathbf{Y}_{t+\tau}^p$
Inference through MCMC

SST - EOF



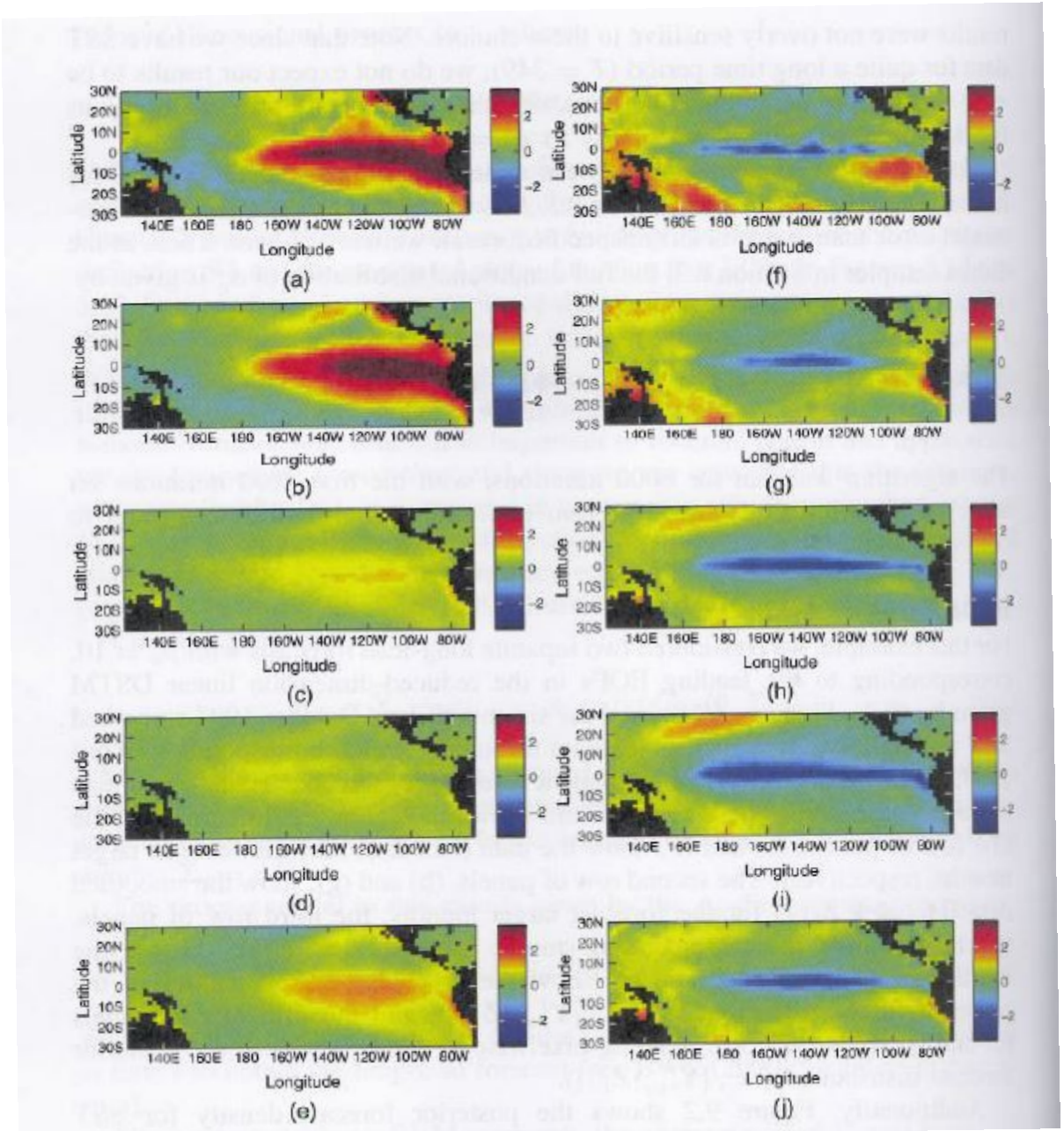
Bayesian priors

$$\begin{aligned}
\mathbf{Z}_t &= \mathbf{Y}_t + \boldsymbol{\nu}_t, & \boldsymbol{\nu}_t &\overset{ind}{\sim} N(\mathbf{0}, \sigma_\nu^2 \mathbf{I}) \\
\mathbf{Y}_t &= \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \boldsymbol{\gamma}_t, & \boldsymbol{\gamma}_t &\overset{ind}{\sim} N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}) \\
\boldsymbol{\alpha}_t &= \mathbf{M}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, & \boldsymbol{\nu}_t &\overset{ind}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_\eta)
\end{aligned}$$

$$\begin{aligned}
\text{vec}(\mathbf{M}) &\sim N(\text{vec}(0.9\mathbf{I}), 100\mathbf{I}) \\
\boldsymbol{\Sigma}_\eta^{-1} &\sim \text{Wishart}\left(\frac{1}{p-1}(100\mathbf{I})^{-1}, p-1\right) \\
\sigma_\nu^2 &\sim \text{Inverse Gamma}(0.1, 100) \\
\sigma_\gamma^2 &\sim \text{Inverse Gamma}(0.1, 100)
\end{aligned}$$

- Fix $\alpha_t = \boldsymbol{\Phi}^T \mathbf{Z}_t, t = 1, ..., 6$
  With $T = 349$ it should not be very sensitive to this initial assumption
- Inference through MCMC using $p_\alpha = 10$.

Linear model - True and predicted processes



Linear model - Posterior forecast of montly Niño region

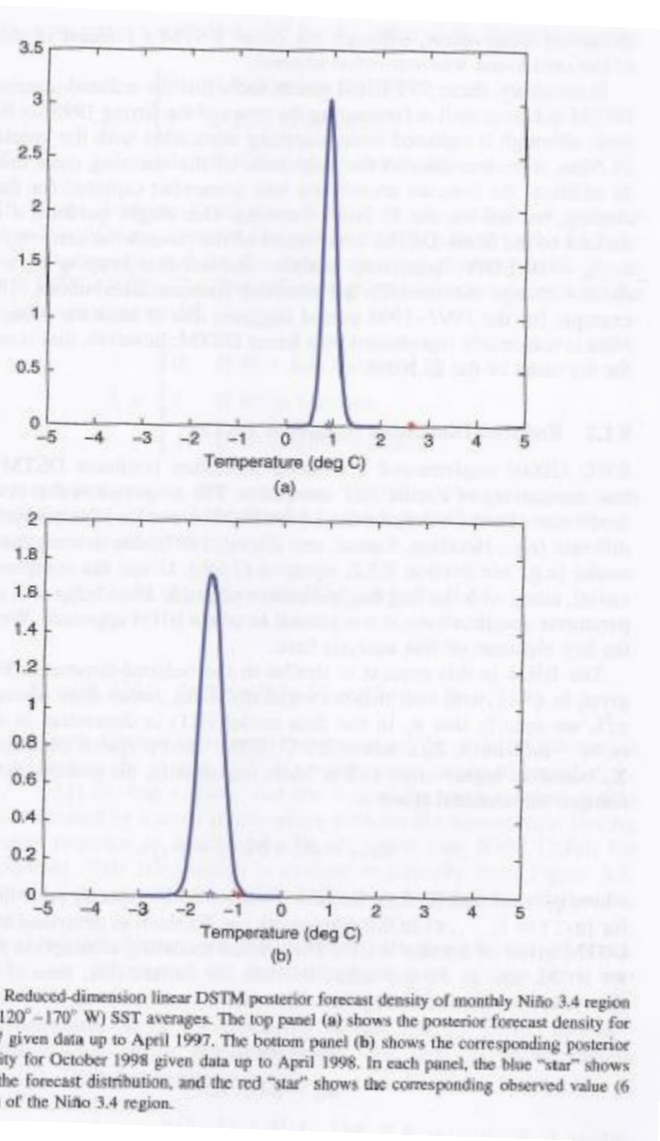


Figure 9.2 Reduced-dimension linear DSTM posterior forecast density of monthly Niño 3.4 region (5° S–5° N, 120°–170° W) SST averages. The top panel **(a)** shows the posterior forecast density for October 1997 given data up to April 1997. The bottom panel **(b)** shows the corresponding posterior forecast density for October 1998 given data up to April 1998. In each panel, the blue "star" shows the mean of the forecast distribution, and the red "star" shows the corresponding observed value (6 months later) of the Niño 3.4 region.

Nonlinear model

Want different dynamics for the El Niño and La Niña phenomena
Assume now $\mathbf{M}_t$ changes with time:

$$\boldsymbol{\alpha}_t = \boldsymbol{\mu}_t + \mathbf{M}_t \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \overset{ind}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$$

Assume $\boldsymbol{\mu}_t$ and $\mathbf{M}_t$ depend on which *climate regime* we are in:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}(l_t), \quad \mathbf{M}_t = \mathbf{M}(l_t)$$

Also extend model to

$$\mathbf{Z}_t = \mathbf{Y}_t + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \overset{ind}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_\nu)$$

Climate regimes

For current/previous time points:

$$I_t = \begin{cases} 0 & \text{if } SOI_t < \text{lower threshold} \\ 1 & \text{if } SOI_t \text{ in between} \\ 2 & \text{if } SOI_t > \text{upper threshold} \end{cases}$$

where SOI_t is the Southern Oscilation index.
For future time points, we *model* I_t as a stochastic process:

$$I_t = \begin{cases} 0 & \text{if } W_t < \text{lower threshold} \\ 1 & \text{if } W_t \text{ in between} \\ 2 & \text{if } W_t > \text{upper threshold} \end{cases}$$

where we assume

$$W_t|\boldsymbol{\beta}_w, \sigma_w^2 \sim \text{Gau}(\mathbf{X}_t^T\boldsymbol{\beta}, \sigma_w)$$
$$\mathbf{X}_t \equiv (1, U_t, U_t \sin(2\pi b_t/12), U_t \cos(2\pi b_t/12), U_t^2)^T$$

where U_t is an east-west component of wind at 10 meters above surface and b_t is an index of month.

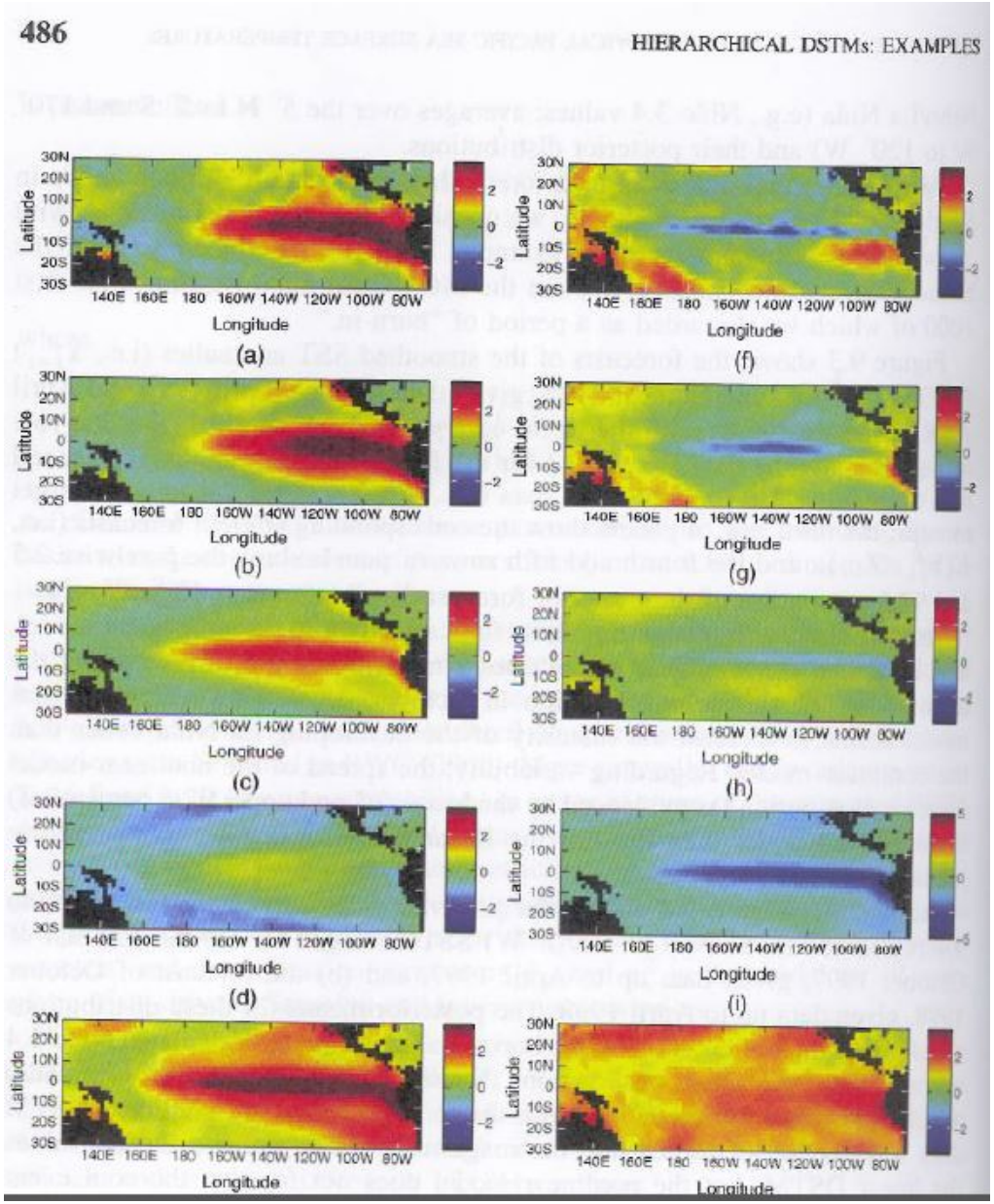
Further details

$$\boldsymbol{\alpha}_t = \begin{pmatrix} \boldsymbol{\alpha}^{(1)} \\ \boldsymbol{\alpha}^{(2)} \end{pmatrix}, \quad \boldsymbol{\mu}_t = \begin{pmatrix} \boldsymbol{\mu}^{(1)}(l_t) \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{M}_t = \begin{pmatrix} \mathbf{M}^{(1,1)}(l_t) & \mathbf{M}^{(1,2)} \\ \mathbf{M}^{(2,1)} & \mathbf{M}^{(2,2)} \end{pmatrix}$$

- Fitting performed by MCMC
- We can get posterior probabilities for climate regimes. This can be used for prediction:

$$E[\boldsymbol{\alpha}_{t+1}|\mathbf{Z}_{1:t}] = \sum_{j=0}^{2} \Pr(l_{t+1} = j|\mathbf{Z}_{1:t}) E[\boldsymbol{\mu}(j) + \mathbf{M}(j)\boldsymbol{\alpha}_t|\mathbf{Z}_{1:t}]$$

Nonlinear model - Posterior forecast of montly Niño region



Nonlinear model - Posterior forecast of montly Niño region

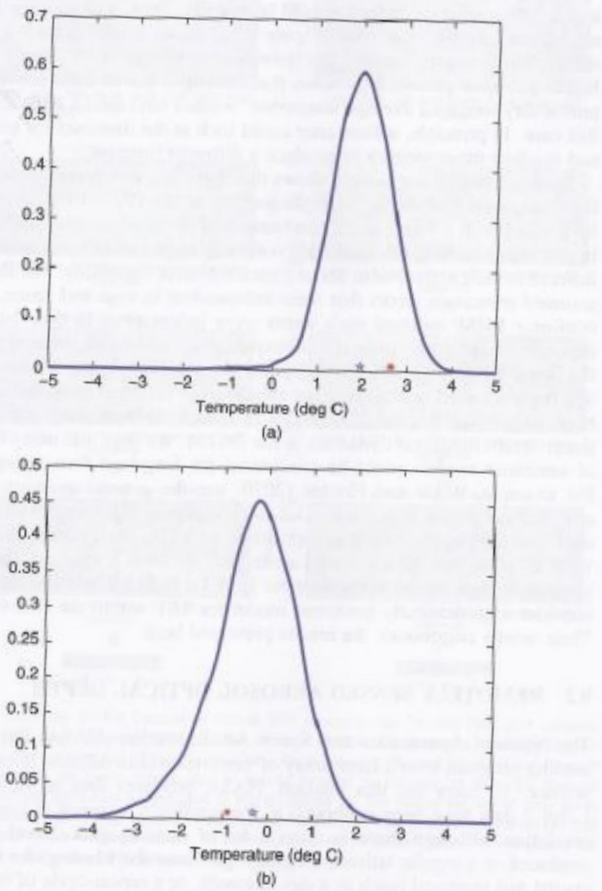
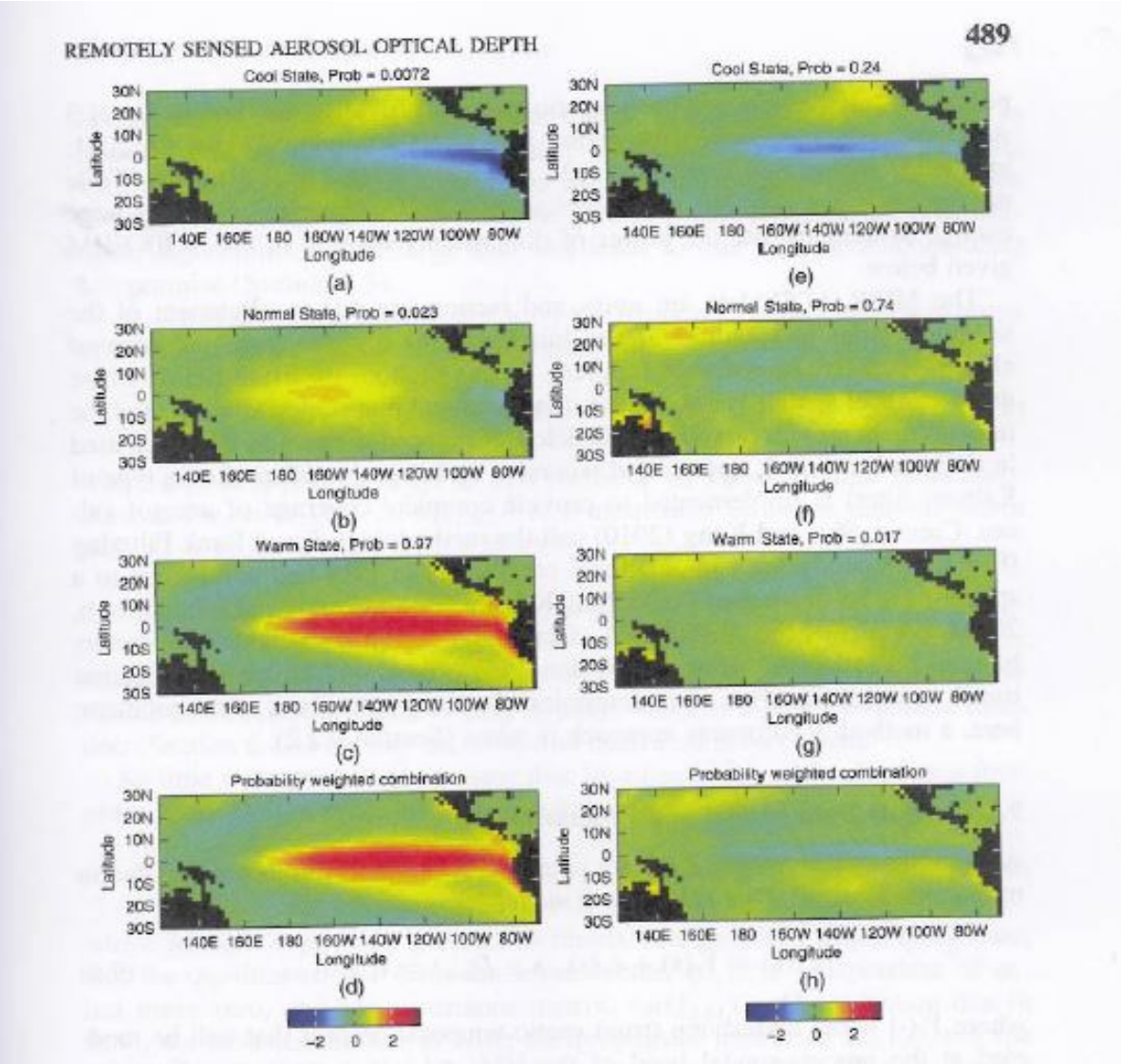


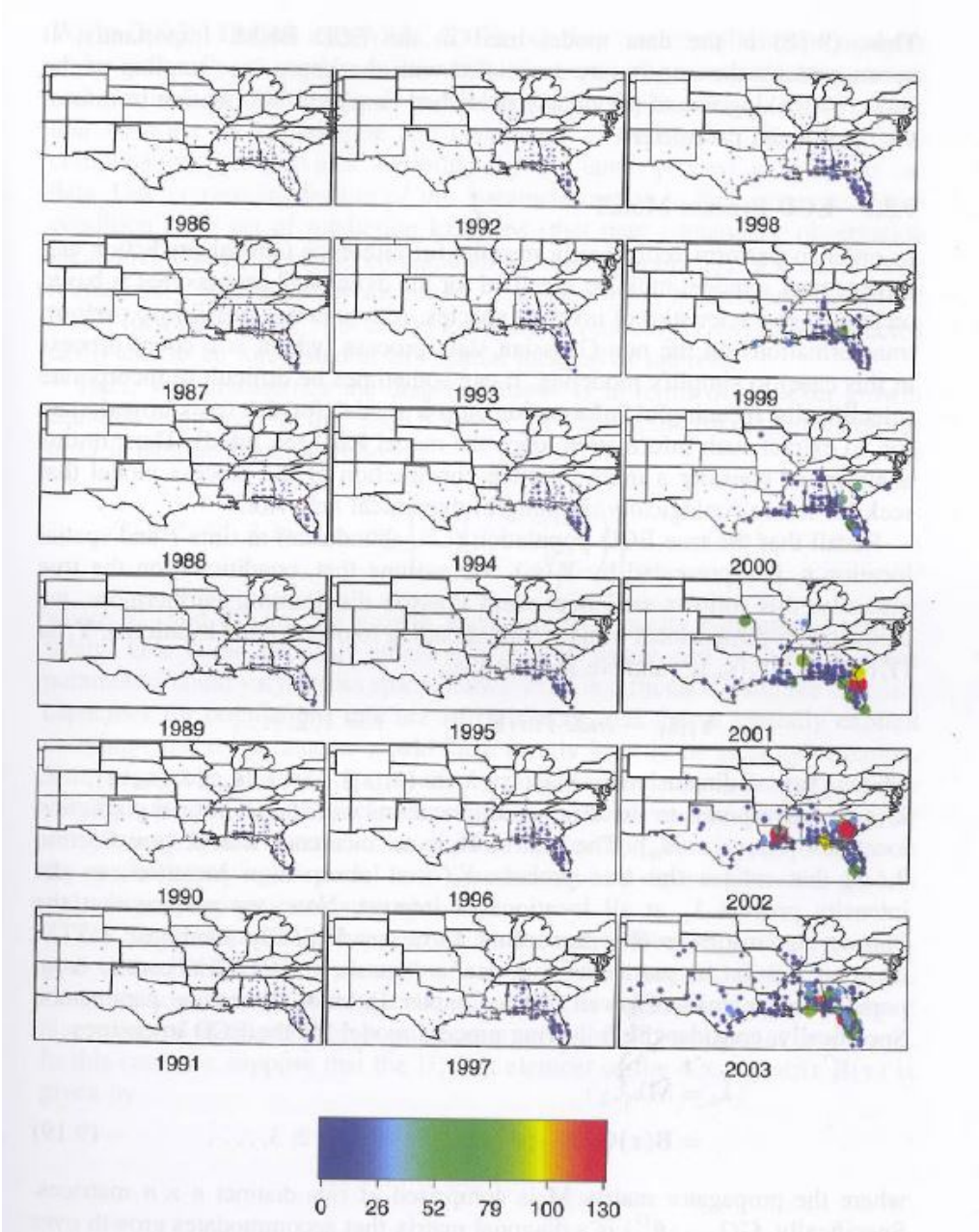
Figure 9.4 Reduced-dimension nonlinear DSTM posterior forecast density of monthly Niño 3.4 region (5° S–5° N, 120°–170° W) SST averages. The top panel **(a)** shows the posterior forecast distribution for October 1997, given data up to April 1997. The bottom panel **(b)** shows the corresponding forecast for October 1998, given data up to April 1998. In each panel, the blue "star" shows the mean of the forecast distribution and the red "star" shows the corresponding observed value (6 months later) of the Niño 3.4 region.

Regime-dependent predictions



Modelling and forecasting the Eurasian Collared Dove invasion (ECD)

- Ecologist interested in growth and dispersal of biotic organisms.
- Dynamics particularly important
- ECD: Counts 1986 - 2002 at peak of the avian breeding season
- Survey routes of approximately 39.4 km
- 50 stops, observer counts and identifies by sight and sound for a 3-minutes period


1986
1992
1998
1987
1993
1999
1988
1994
2000
1989
1995
2001
1990
1996
2002
1991
1997
2003
0
26
52
79
100
130

Modelling observations

Assume $Y_t(\mathbf{s}_i)$ is the true (unknown) number of ECD counts.
Assume

$$Z_t(\mathbf{s}_i)|\mathbf{Y}_t(\mathbf{s}_i), \theta \overset{ind}{\sim} \text{Binomial}(Y_t(\mathbf{s}_i), \theta)$$

Modelling observations

Assume $Y_t(\mathbf{s}_i)$ is the true (unknown) number of ECD counts.
Assume

$$Z_t(\mathbf{s}_i)|\mathbf{Y}_t(\mathbf{s}_i), \theta \overset{ind}{\sim} \text{Binomial}(Y_t(\mathbf{s}_i), \theta)$$

- Both $Y_t(\mathbf{s}_i)$ and θ are not identifiable without additional information
- Information on θ possible through *replicable* samples. Not available

Modelling observations

Assume $Y_t(\mathbf{s}_i)$ is the true (unknown) number of ECD counts.
Assume

$$Z_t(\mathbf{s}_i)|\mathbf{Y}_t(\mathbf{s}_i), \theta \overset{ind}{\sim} \text{Binomial}(Y_t(\mathbf{s}_i), \theta)$$

- Both $Y_t(\mathbf{s}_i)$ and θ are not identifiable without additional information
- Information on θ possible through *replicable* samples. Not available
- Replicate counts were taken on a different, but related species. Can be used to construct an *informative* prior for θ: $\theta \sim \text{Beta}(a_\theta, b_\theta)$

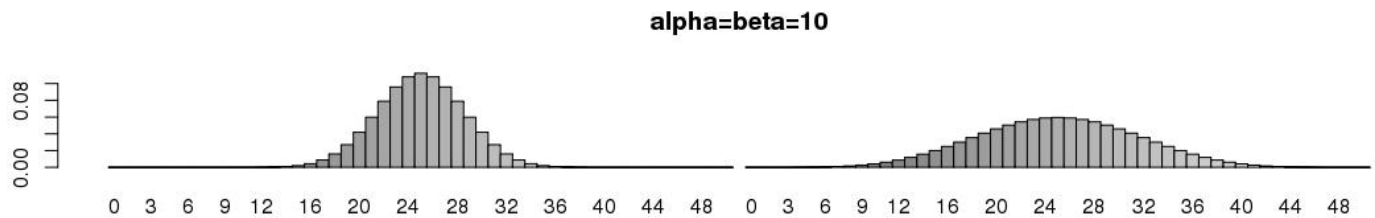
Modelling observations

Assume $Y_t(\mathbf{s}_i)$ is the true (unknown) number of ECD counts.
Assume

$$Z_t(\mathbf{s}_i)|\mathbf{Y}_t(\mathbf{s}_i), \theta \overset{ind}{\sim} \text{Binomial}(Y_t(\mathbf{s}_i), \theta)$$

- Both $Y_t(\mathbf{s}_i)$ and θ are not identifiable without additional information
- Information on θ possible through *replicable* samples. Not available
- Replicate counts were taken on a different, but related species. Can be used to construct an *informative* prior for θ: $\theta \sim \text{Beta}(a_\theta, b_\theta)$
- Gives (assuming different θ's for each observation)

$$[Z_t(\mathbf{s}_i)|\mathbf{Y}_t(\mathbf{s}_i)] = \int_0^1 [Z_t(\mathbf{s}_i)|\mathbf{Y}_t(\mathbf{s}_i), \theta][\theta]d\theta = \text{Beta-Bin}(Y_t(\mathbf{s}), a_\theta, b_\theta)$$



Process model

Numbers at observation locations

$$\mathbf{Y}_t = (Y_t(\mathbf{s}_1), ..., Y_t(\mathbf{s}_m))^T$$

Intensity at observation and prediction locations

$$\boldsymbol{\lambda}_t = (\lambda_t(\mathbf{s}_1), ..., \lambda_t(\mathbf{s}_m), \lambda_t(\mathbf{s}_{m+1}), ..., \lambda_t(\mathbf{s}_n))^T$$

Assume

$$Y_t(\mathbf{s}_i)|\boldsymbol{\lambda}_t \overset{ind}{\sim} \text{Poisson}(\lambda_t(\mathbf{s}_i)), i = 1, ..., n$$

$$\mathbf{Y}_t|\boldsymbol{\lambda}_t \overset{ind}{\sim} \text{Poisson}(\mathbf{H}\boldsymbol{\lambda}_t)$$

$\mathbf{H}$ is an incidence matrix

Model for intensity

$$\begin{aligned}\boldsymbol{\lambda}_t &= \mathbf{M}\boldsymbol{\lambda}_{t-1} \\ &= \mathbf{B}(\boldsymbol{\tau})\mathbf{G}(\boldsymbol{\lambda}_{t-1};\boldsymbol{\theta}^G)\boldsymbol{\lambda}_{t-1}\end{aligned}$$

$\mathbf{G}(\boldsymbol{\lambda}_{t-1};\boldsymbol{\theta}^G)$:Diagonal matrix for growth over time

$$G_{ii} = \exp\left\{\theta_1^G\left(1 - \frac{\lambda_t(\mathbf{s}_i)}{\theta_2^G}\right)\right\}$$

$\mathbf{B}(\boldsymbol{\tau})$:Dispersal

$$B_{ij} \propto \exp\left\{-\frac{-d_{ij}^2}{\tau(\mathbf{s}_i)}\right\}, \quad \sum_j B_{ij} = 1$$

- **G**: Ricker growth model
- Only random parts: $\boldsymbol{\lambda}_1$ and priors on parameters $(\boldsymbol{\theta}^G, \boldsymbol{\tau})$
- Corresponds to a discretized diffusion equation with a population that can pass through "open boundaries"

Model for intensity

$$\begin{aligned}\boldsymbol{\lambda}_t &= \mathbf{M}\boldsymbol{\lambda}_{t-1} \\ &= \mathbf{B}(\boldsymbol{\tau})\mathbf{G}(\boldsymbol{\lambda}_{t-1};\boldsymbol{\theta}^G)\boldsymbol{\lambda}_{t-1}\end{aligned}$$

$\mathbf{G}(\boldsymbol{\lambda}_{t-1};\boldsymbol{\theta}^G)$:Diagonal matrix for growth over time

$$G_{ii} = \exp\left\{\theta_1^G\left(1 - \frac{\lambda_t(\mathbf{s}_i)}{\theta_2^G}\right)\right\}$$

$\mathbf{B}(\boldsymbol{\tau})$:Dispersal

$$B_{ij} \propto \exp\left\{-\frac{-d_{ij}^2}{\tau(\mathbf{s}_i)}\right\}, \quad \sum_j B_{ij} = 1$$

Model for intensity

$$\begin{aligned}\boldsymbol{\lambda}_t &= \mathbf{M}\boldsymbol{\lambda}_{t-1} \\ &= \mathbf{B}(\boldsymbol{\tau})\mathbf{G}(\boldsymbol{\lambda}_{t-1};\boldsymbol{\theta}^G)\boldsymbol{\lambda}_{t-1}\end{aligned}$$

$\mathbf{G}(\boldsymbol{\lambda}_{t-1};\boldsymbol{\theta}^G)$:Diagonal matrix for growth over time

$$G_{ii} = \exp\left\{\theta_1^G\left(1 - \frac{\lambda_t(\mathbf{s}_i)}{\theta_2^G}\right)\right\}$$

$\mathbf{B}(\boldsymbol{\tau})$:Dispersal

$$B_{ij} \propto \exp\left\{-\frac{-d_{ij}^2}{\tau(\mathbf{s}_i)}\right\}, \quad \sum_j B_{ij} = 1$$

- **G**: Ricker growth model

Model for intensity

$$\begin{aligned}\boldsymbol{\lambda}_t &= \mathbf{M}\boldsymbol{\lambda}_{t-1} \\ &= \mathbf{B}(\boldsymbol{\tau})\mathbf{G}(\boldsymbol{\lambda}_{t-1};\boldsymbol{\theta}^G)\boldsymbol{\lambda}_{t-1}\end{aligned}$$

$\mathbf{G}(\boldsymbol{\lambda}_{t-1};\boldsymbol{\theta}^G)$:Diagonal matrix for growth over time

$$G_{ii} = \exp\left\{\theta_1^G\left(1 - \frac{\lambda_t(\mathbf{s}_i)}{\theta_2^G}\right)\right\}$$

$\mathbf{B}(\boldsymbol{\tau})$:Dispersal

$$B_{ij} \propto \exp\left\{-\frac{-d_{ij}^2}{\tau(\mathbf{s}_i)}\right\}, \quad \sum_j B_{ij} = 1$$

- **G**: Ricker growth model
- Only random parts: $\boldsymbol{\lambda}_1$ and priors on parameters $(\boldsymbol{\theta}^G, \boldsymbol{\tau})$
- Corresponds to a discretized diffusion equation with a population that can pass through "open boundaries"

Spatial modelling

- $\boldsymbol{\lambda}_1$ is a Gaussian spatial process with exponential covariance function at log scale
- Similar for $\boldsymbol{\tau}$
- Allow for predictions at points $\mathbf{s}_{m+1}, ..., \mathbf{s}_n$.
- Dynamical model allows or forcasting

Dove: Spatial predictions and forecast

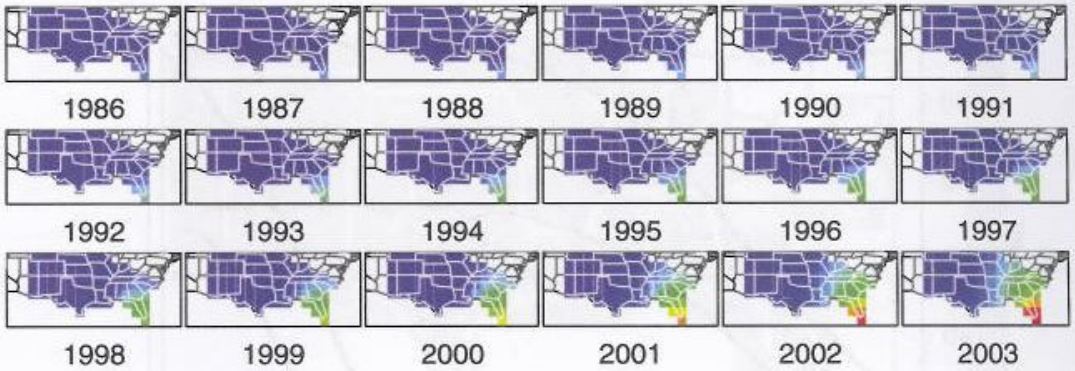


Figure 9.11 Posterior means of ECD abundance in the United States for the period 1986 (corresponding to the sampling period considered in this analysis). Areas of the map s "white" were not included in the prediction grid.

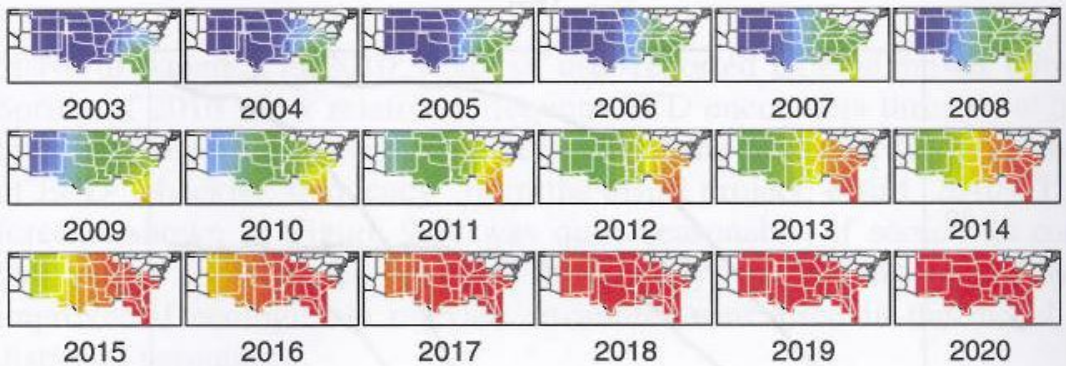


Figure 9.12 Posterior mean of the ECD-abundance forecast in the United States for th 2003 to 2020 (where the years 2004–2020 correspond to out-of-sample forecasts). Areas of shown as "white" were not included in the prediction grid.

Dove: Specific locations

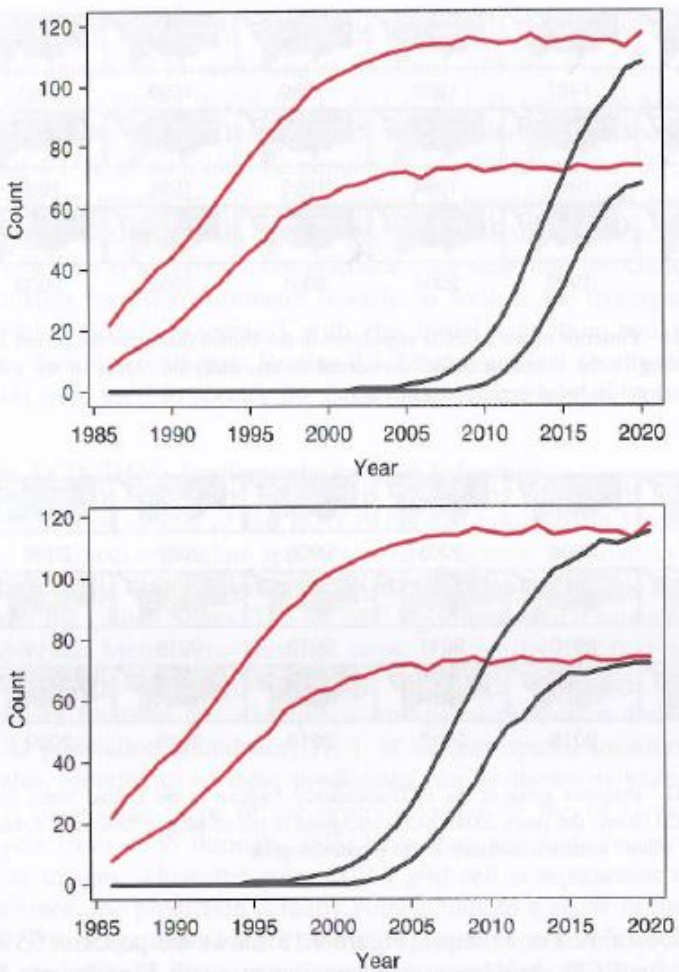


Figure 9.13 Comparisons of posterior credible intervals for the ECD population from 1986 to 2020. **Top panel:** Credible intervals for a location in south Florida (red) and for eastern Kansas (black). **Bottom panel:** Credible intervals for the south Florida location (red) and for northern Utah (black).

Dove: Dispersal τ

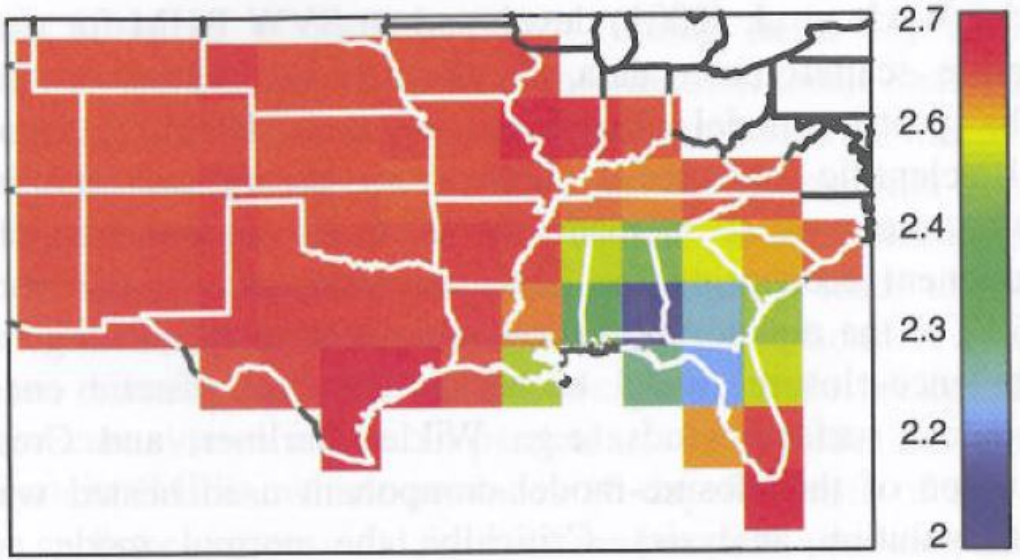


Figure 9.14 Posterior mean of the spatially varying dispersal coefficients (τ) in the ECD process model (9.19).

Remarks

- Implementation through MCMC
- Can also obtain uncertainties
  - Look similar to posterior mean (variance equal to expectation in Poisson distribution)
- Posterior distributions of growth and dispersal parameters where more narrow than their priors, indicating learning from data
- Spatially varying τ's was supported by data