# Analysis time series using state space approach

Marko Laine

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

LUT 2015-12-07

# Outline

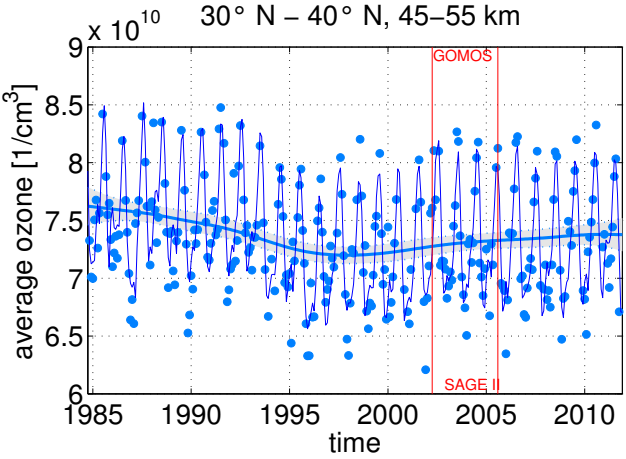
# Typical features in climatic series

- Non-stationary behaviour driven by external forcing.
- Low frequency oscillation.
- Long time behaviour hard to estimate from short series.
- Persistent long range correlations.
- Long memory, slower than exponential decay of the autocorrelation function.

Satellite remote sensing observations.

- Non uniform sampling in space and time.
- Data combined from different instruments.
- Missing values.
- Instrument bias, aging, ...

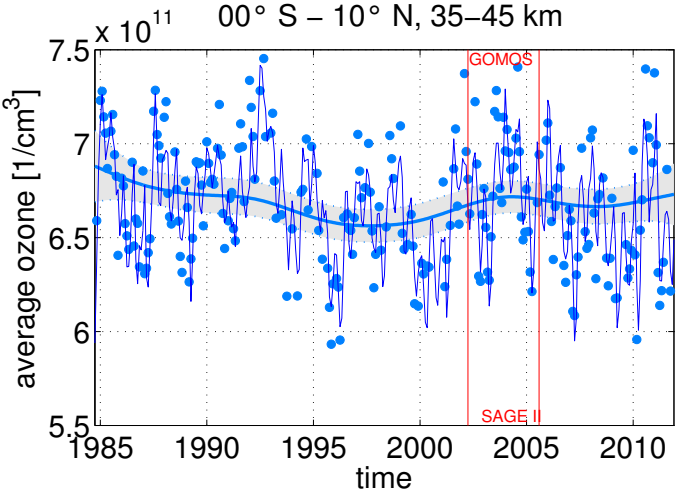
# Time series modelling by DLM




- DLM: dynamic linear model.
- Components: trend, seasonality, external forcing, random noise.
- Process description of time series.
- Dynamic regression: the regression coefficients can change in time.

# What is trend

- Trend is a change in the background level of the process.
- We are interested in (smooth) long term (decade) change.
- Need to filter out seasonality, external variation driven by known phenomena, random noise.



# Bayesian hierarchical model

- $y_t$: observations,
- $x_t$: hidden model states,
- $\theta$: model parameters,
- $t = 1, \ldots, n$: observation times.

- Observation model: $p(y_t|x_t, \theta)$
- Process model: $p(x_{t+1}|x_t, \theta)$
- Parameter model: $p(\theta)$

- Bayes formula:

$$p(x_{1:n}, \theta|y_{1:n}) \propto \prod_{t=1}^{n} p(y_t|x_t, \theta)p(x_t|x_{t-1}, \theta)p(\theta)$$

# Dynamic linear model

$$
\begin{aligned}
y_t &= F_t x_t + v_t, & v_t &\sim N_p(0, V_t) \\
x_t &= G_t x_{t-1} + w_t, & w_t &\sim N_q(0, W_t) \\
V_t &\sim p(\sigma_{\text{obs}}), \quad W_t \sim p(\sigma_{\text{mod}})
\end{aligned}
$$

- Observation operator $F_t$, model operator $G_t$.
- Observation uncertainty $v_t$, model uncertainty $w_t$. Assumed Gaussian with covariance matrices $V_t$ and $W_t$.
- Known, partly known, i.e. may contain parameters, unknown, i.e. must be estimated.

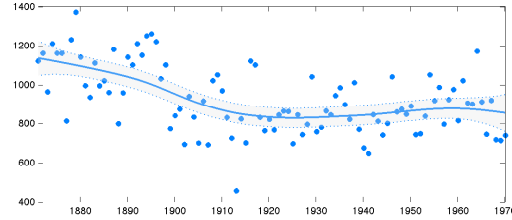
# Simple example

- Local level and trend model

$$
\begin{aligned}
y_t &= \mu_t + \epsilon_{\text{obs}}, & \epsilon_{\text{obs}} &\sim N(0, \sigma^2_{\text{obs}}), \text{ observations} \\
\mu_t &= \mu_{t-1} + \alpha_{t-1} + \epsilon_{\text{level}}, & \epsilon_{\text{level}} &\sim N(0, \sigma^2_{\text{level}}), \text{ local level} \\
\alpha_t &= \alpha_{t-1} + \epsilon_{\text{trend}}, & \epsilon_{\text{trend}} &\sim N(0, \sigma^2_{\text{trend}}), \text{ local trend}
\end{aligned}
$$

$$
G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad x_t = \begin{bmatrix} \mu_t & \alpha_t \end{bmatrix}^T.
$$

When $\sigma_{\text{level}} = 0$, this is cubic spline smoothing with smoothness parameter $\lambda = \sigma^2_{\text{trend}}/\sigma^2_{\text{obs}}$.



# Seasonal model

$$G_{\text{seas}} = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$F_{\text{seas}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

# Seasonal model, with harmonic functions

$$G_{\text{seas}(k)} = \begin{bmatrix} \cos(k2\pi/12) & \sin(k2\pi/12)) \\ -\sin(k2\pi/12) & \cos(k2\pi/12)) \end{bmatrix}, \; F =_{\text{seas}} \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

# Covariates, proxy variables

$$G_{\text{proxy}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$F_t = \begin{bmatrix} X_{1,t} & X_{2,t} & X_{3,t} \end{bmatrix}.$$

In stratospheric ozone analysis: proxies for solar activity, quasi biennial oscillation, ENSO.

# Then combine all

$$G = \begin{bmatrix} G_{\text{trend}} & 0 & 0 \\ 0 & G_{\text{seasonal}} & 0 \\ 0 & 0 & G_{\text{covariates}} \end{bmatrix},$$

$$F = \begin{bmatrix} F_{\text{trend}} & F_{\text{seasonal}} & F_{\text{covariates}} \end{bmatrix},$$

and estimate the hidden states by state space methods, i.e. estimate $p(x_{1:n}|y_{1:n})$, typically only the marginal distributions.

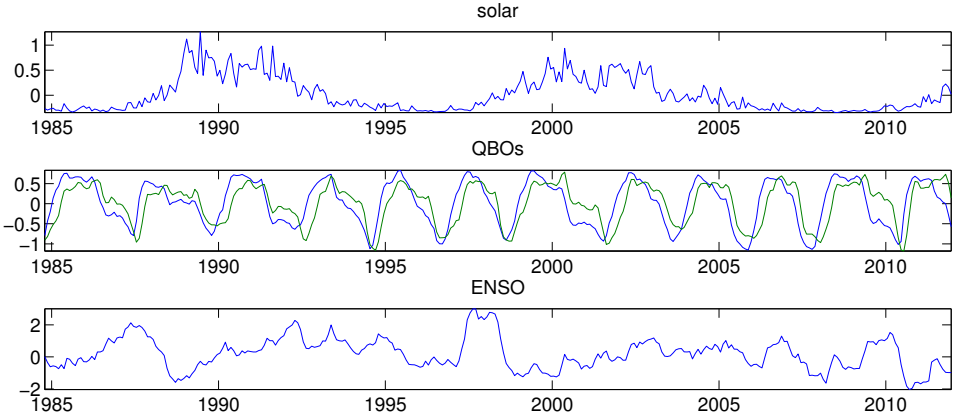
# Why DLM

- Provides hierarchical statistical model for uncertainties in data, process, and parameters.
- Uses verifiable statistical assumptions.
- No need to assume stationarity.
- No problems with missing values.
- Classical regression, spline smoothing, ARIMA analyses are special cases.
- Efficient calculations by Kalman formulas.
- Extendible to non-linear models, hierarchical parameters, non Gaussian errors.

# Model for ozone time series

$$y_t = \mu_t + \gamma_t + \beta_{1,t} X_{1,t} + \beta_{2,t} X_{2,t} + \beta_{3,t} X_{3,t} + \epsilon_t \quad (1)$$

- Local level and trend, $\sigma_{\text{level}} = 0$, $\sigma_{\text{trend}} > 0$.
- Seasonality with two harmonic functions, $\sigma_{\text{seas}} > 0$
- Proxy variables for solar effect and QBO and ENSO, fixed coefficients, $\sigma_{\text{proxy}} = 0$



# The matrices involved

$$x_t = \begin{bmatrix} \mu_t & \alpha_t & \psi_{t,1} & \psi^*_{t,1} & \psi_{t,2} & \psi^*_{t,2} & \beta_1 & \beta_2 & \beta_3 \end{bmatrix}^T$$

$$G = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos\left(\frac{\pi}{6}\right) & \sin\left(\frac{\pi}{6}\right) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\sin\left(\frac{\pi}{6}\right) & \cos\left(\frac{\pi}{6}\right) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos\left(\frac{\pi}{3}\right) & \sin\left(\frac{\pi}{3}\right) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\sin\left(\frac{\pi}{3}\right) & \cos\left(\frac{\pi}{3}\right) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$F_t = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & \text{solar}(t) & \text{qbo1}(t) & \text{qbo2}(t) \end{bmatrix}$$

$$W = \text{diag}\begin{bmatrix} 0 & \sigma^2_\alpha & \sigma^2_\psi & \sigma^2_\psi & \sigma^2_\psi & \sigma^2_\psi & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\theta = \begin{bmatrix} \sigma_\alpha & \sigma_\psi \end{bmatrix}^T$$

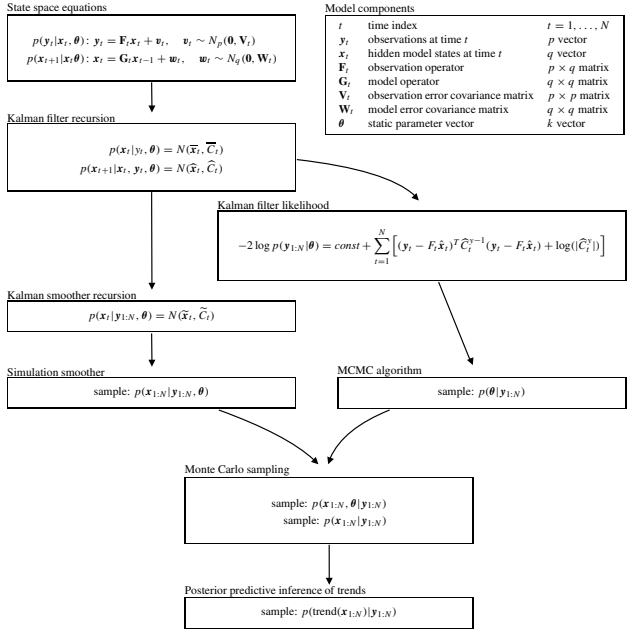
# Outline of the estimation

For dynamic linear models we have computational tools for all relevant statistical distributions.

- $p(x_t|y_{1:t}, \theta)$ by Kalman filter
- $p(x_{t+1}|x_t, y_{1:t}, \theta)$ by Kalman filter
- $p(x_t|y_{1:n}, \theta)$ by Kalman smoother
- $p(x_{1:n}|y_{1:n}, \theta)$ by simulation smoother
- $p(y_{1:n}|\theta)$ by Kalman filter likelihood
- $p(x_{1:n}, \theta|y_{1:n})$ by MCMC

Above, $\theta$ contains all the auxiliary model parameter, e.g., related to observation and model error covariances. They are either fixed, estimated by maximum likelihood, or marginalized over by MCMC.

# Flowchart


State space equations
$p(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\theta}):\ \boldsymbol{y}_t = \mathbf{F}_t\boldsymbol{x}_t + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim N_p(\mathbf{0},\mathbf{V}_t)$
$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t\boldsymbol{\theta}):\ \boldsymbol{x}_t = \mathbf{G}_t\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim N_q(\mathbf{0},\mathbf{W}_t)$
Model components
$t$ time index $t = 1,\ldots,N$
$\boldsymbol{y}_t$ observations at time $t$ $p$ vector
$\boldsymbol{x}_t$ hidden model states at time $t$ $q$ vector
$\mathbf{F}_t$ observation operator $p \times q$ matrix
$\mathbf{G}_t$ model operator $q \times q$ matrix
$\mathbf{V}_t$ observation error covariance matrix $p \times p$ matrix
$\mathbf{W}_t$ model error covariance matrix $q \times q$ matrix
$\boldsymbol{\theta}$ static parameter vector $k$ vector
Kalman filter recursion
$p(\boldsymbol{x}_t|y_t,\boldsymbol{\theta}) = N(\overline{\boldsymbol{x}}_t, \overline{\boldsymbol{C}}_t)$
$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t,\ \boldsymbol{y}_t,\boldsymbol{\theta}) = N(\widehat{\boldsymbol{x}}_t, \widehat{C}_t)$
Kalman filter likelihood
$-2\log p(\boldsymbol{y}_{1:N}|\boldsymbol{\theta}) = const + \sum_{t=1}^{N}\left[(\boldsymbol{y}_t - F_t\hat{\boldsymbol{x}}_t)^T\widehat{C}_t^{y-1}(\boldsymbol{y}_t - F_t\hat{\boldsymbol{x}}_t) + \log(|\widehat{C}_t^y|)\right]$
Kalman smoother recursion
$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:N},\boldsymbol{\theta}) = N(\widetilde{\boldsymbol{x}}_t, \widetilde{C}_t)$
Simulation smoother
sample: $p(\boldsymbol{x}_{1:N}|\boldsymbol{y}_{1:N},\boldsymbol{\theta})$
MCMC algorithm
sample: $p(\boldsymbol{\theta}|\boldsymbol{y}_{1:N})$
Monte Carlo sampling
sample: $p(\boldsymbol{x}_{1:N},\boldsymbol{\theta}|\boldsymbol{y}_{1:N})$
sample: $p(\boldsymbol{x}_{1:N}|\boldsymbol{y}_{1:N})$
Posterior predictive inference of trends
sample: $p(\text{trend}(\boldsymbol{x}_{1:N})|\boldsymbol{y}_{1:N})$

# Kalman filter

Assume the initial state distribution $N(x_1, C_1)$ is known. One step prediction prior mean and covariance, $p(x_t|x_{t-1}, y_{1:t-1}, \theta) = N(\widehat{x}_t, \widehat{C}_t)$ and covariance matrix for predicted observations $C_{y,t}$:

$$\begin{aligned}
\widehat{x}_t &= G_t x_{t-1} && \text{prior mean for state } x_t, \\
\widehat{C}_t &= G_t C_{t-1} G_t^T + W_t && \text{prior covariance for state } x_t, \\
C_{y,t} &= F_t \widehat{C}_t F_t^T + V_t && \text{Covariance for prediction } y_t
\end{aligned}$$

and posterior mean and covariance, $p(x_t|y_{1:t}, \theta) = N(x_t, C_t)$

$$\begin{aligned}
K_t &= \widehat{C}_t F_t^T C_{y,t}^{-1} && \text{Kalman Gain}, \\
v_t &= y_t - F_t \widehat{x}_t && \text{prediction residuals}, \\
x_t &= \widehat{x}_t + K_t v_t && \text{tilan } x_t \text{ posterior mean}, \\
C_t &= \widehat{C}_t - K_t F_t \widehat{C}_t && \text{tilan } x_t \text{ posterior covariance},
\end{aligned}$$

for each time $t = 2, 3, \ldots, n$.

# Kalman filter, Matlab code

```
for i=1:n
  v(:,i) = y(i,:)' - F*x(:,i);
  Cp(:,:,i) = F*C(:,:,i)*F' + diag(V(i,:).^2);
  K(:,:,i) = G*C(:,:,i)*F'/Cp(:,:,i);
  if i<n
    L = G-K(:,:,i)*F;
    x(:,i+1) = G*x(:,i) + K(:,:,i)*v(:,i);
    C(:,:,i+1) = G*C(:,:,i)*L' + W;
  end
end
```

# Kalman smoother

Smoothed states $p(x_t|y_{1:n}, \theta) = N(\widetilde{x}_t, \widetilde{C}_t)$ are obtained by backward recursion:

$$
\begin{aligned}
L_t &= G_t - G_t K_t F_t \\
r_t &= F_t^T C_{y,t}^{-1} v_t + L_t^T r_{t+1} \\
N_t &= F_t^T C_{y,t}^{-1} F_t + L_t^T N_{t+1} L_t \\
\widetilde{x}_t &= \widehat{x}_t + \widehat{C}_t r_t && \text{smoothed state mean} \\
\widetilde{C}_t &= \widehat{C}_t - \widehat{C}_t N_t \widehat{C}_t && \text{smoothed state covariance,}
\end{aligned}
$$

for each time $t = n, n-1, \ldots, 1$, with initialization $r_{n+1}$ and $N_{n+1}$ as zero.

# Kalman smoother, Matlab code

```
for i=n:-1:1
  L = G-K(:,:,i)*F;
  r = F'/Cp(:,:,i)*v(:,i) + L'*r;
  N = F'/Cp(:,:,i)*F + L'*N*L;
  x(:,i) = x(:,i) + C(:,:,i)*r;
  C(:,:,i) = C(:,:,i) - C(:,:,i)*N*C(:,:,i);
end
```

# Likelihood function from the Kalman filter

Marginal likelihood $p(y_{1:n}|\theta)$, with the state $x_{1:n}$ "integrated out" is obtained from one step predictions:

$$p(y_{1:n}|\theta) = p(y_1|\theta) \prod_{t=2}^{n} p(y_t|y_{1:t-1}, \theta).$$

Individual predictions are obtained from Kalman filter recursion for a fixed parameter $\theta$. For a linear model with Gaussian errors, we have:

$$p(y_{1:n}|\theta) \propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{n} \left[ (y_t - F_t \widehat{x}_t)^T \, C_{y,t}^{-1} \, (y_t - F_t \widehat{x}_t) + \log(|C_{y,t}|) \right] \right\}.$$

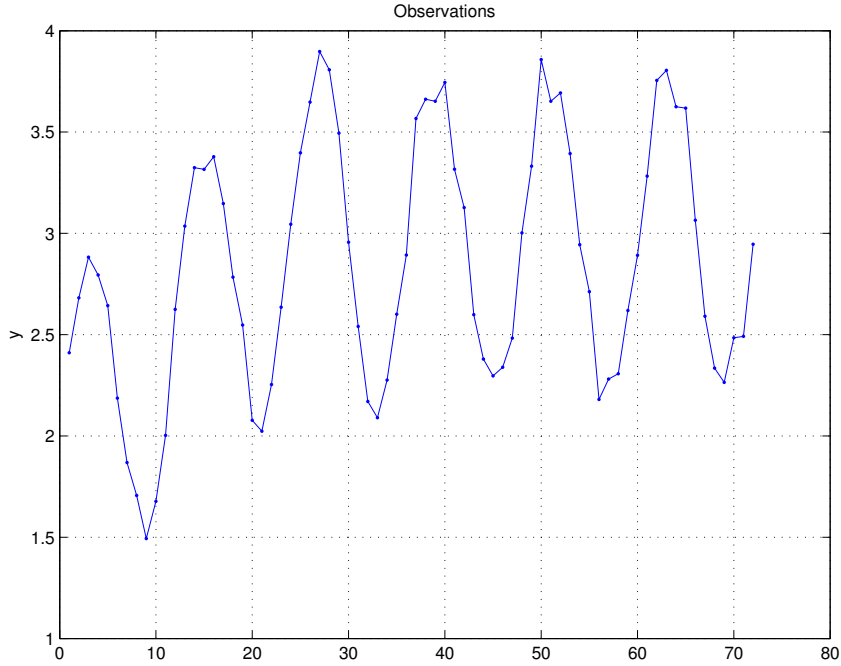
# Sample from the estimated state

Kalman formulas give marginal distributions $p(x_t|y_{1:n})$ but there is efficient way to simulate values from the whole posterior distribution of the states $p(x_{1:n}|y_{1:n})$:

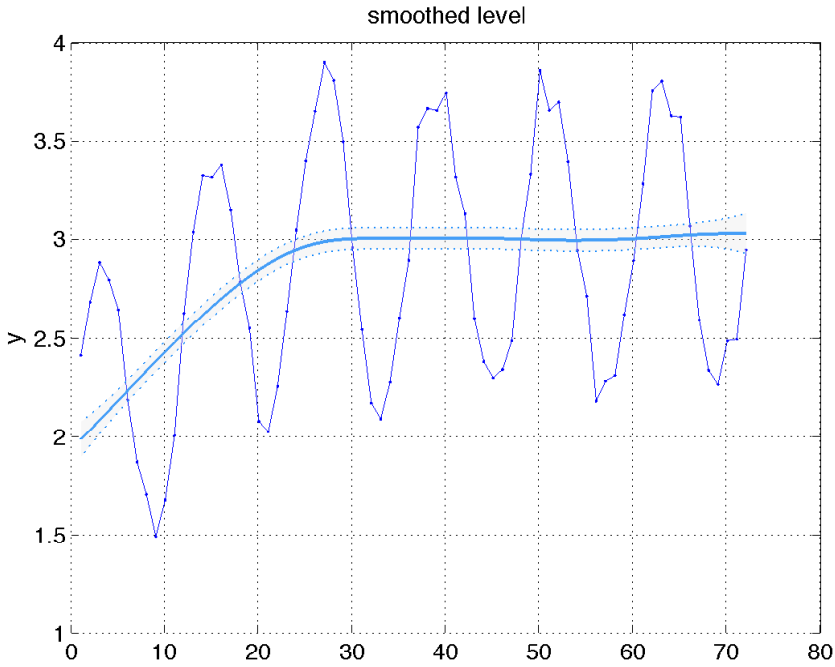
1. Sample from the system equations, $\widetilde{x}_{1:n}$, $\widetilde{y}_{1:n}$.
2. Smooth $\widetilde{y}_{1:n}$ to get $\widetilde{\widetilde{x}}_{1:n}$.
3. Add the residuals to the original smoothed state, $x^*_{1:n} = \widetilde{x}_{1:n} - \widetilde{\widetilde{x}}_{1:n} + x_{1:n}$.

This is needed to get uncertainty estimates of trend related statistics.

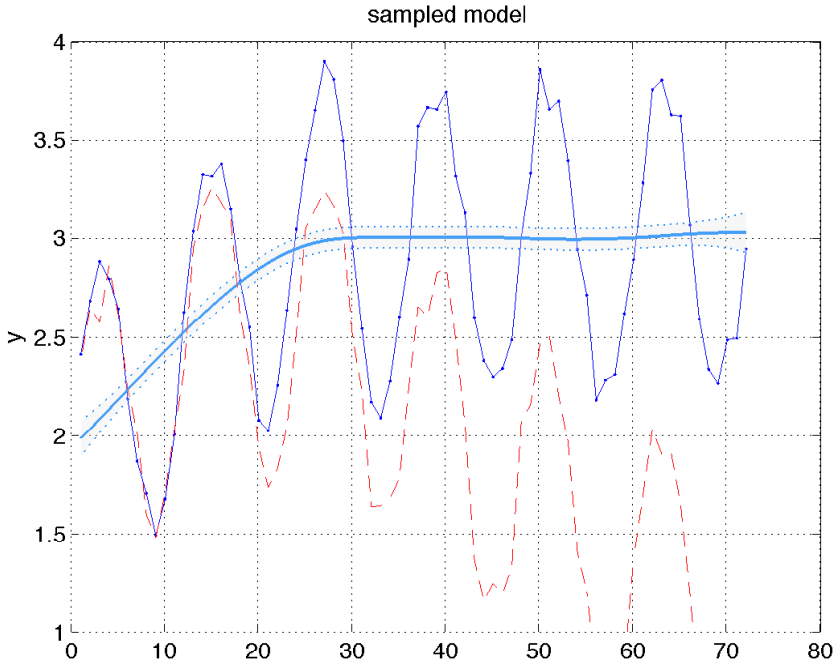
# Example



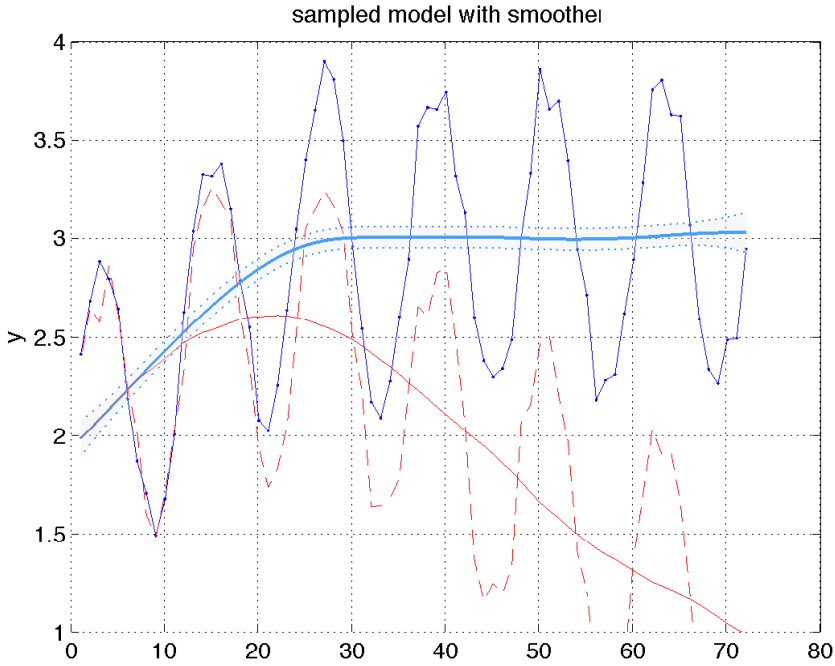
# Example



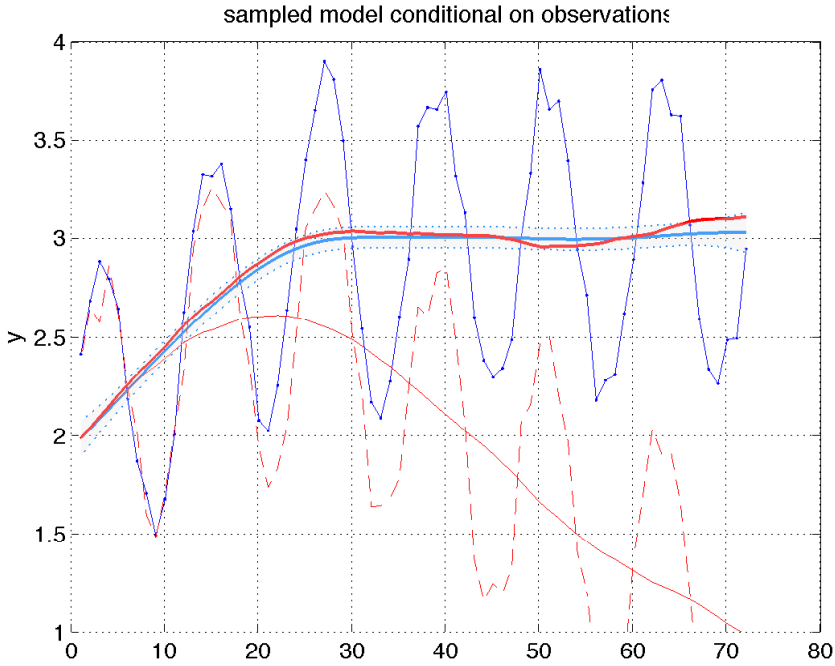
# Example



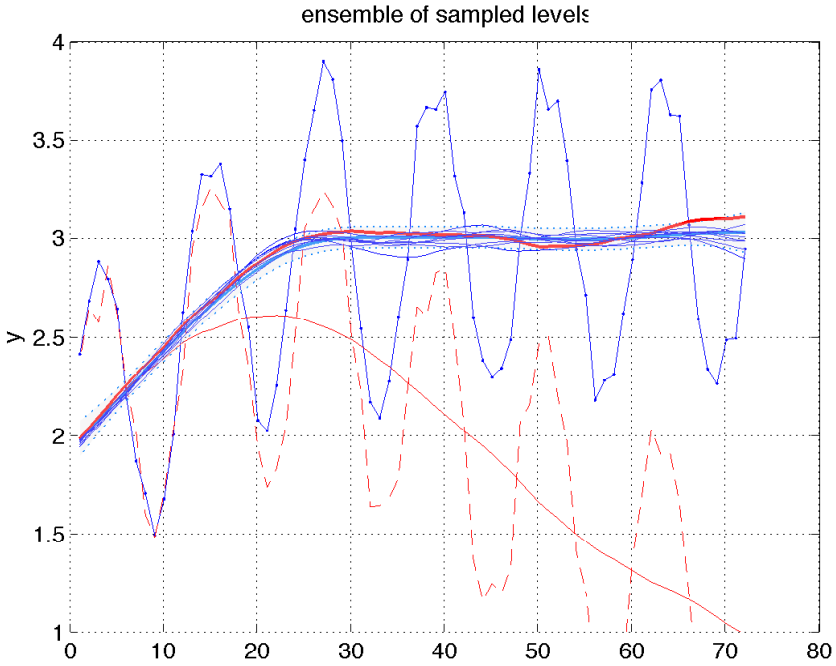
# Example


sampled model with smoothe
y
4
3.5
3
2.5
2
1.5
1
0
10
20
30
40
50
60
70
80

# Example


sampled model conditional on observations
y
4
3.5
3
2.5
2
1.5
1
0
10
20
30
40
50
60
70
80

# Example

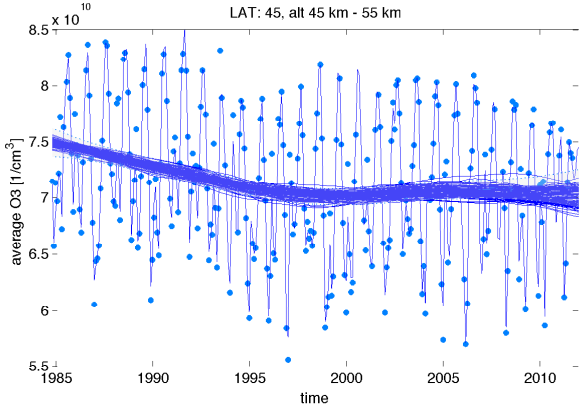

ensemble of sampled levels
y
4
3.5
3
2.5
2
1.5
1
0
10
20
30
40
50
60
70
80

# Trend analysis by simulation

- Kalman formulas give marginal distributions $p(x_t|y_{1:n}, \theta)$.
- We can simulate DLM states from $p(x_{1:n}|y_{1:n}, \theta)$.
- Need MCMC to simulate from

$$p(x_{1:n}|y_{1:n}) = \int p(x_{1:n}|y_{1:n}, \theta)\, d\theta.$$

- This is needed to get uncertainty estimates of trend related statistics.



# Estimating parameters

- How to select the model matrix $G$?
  - By considering all the relevant processes.
  - Diagnosing the residuals.
- How to choose initial state distribution $N(x_0, C_0)$?
  - By assuming diffuse priors ....
- How to estimate error covariances $W_t$ and $V_t$?
  - By Kalman filter likelihood:

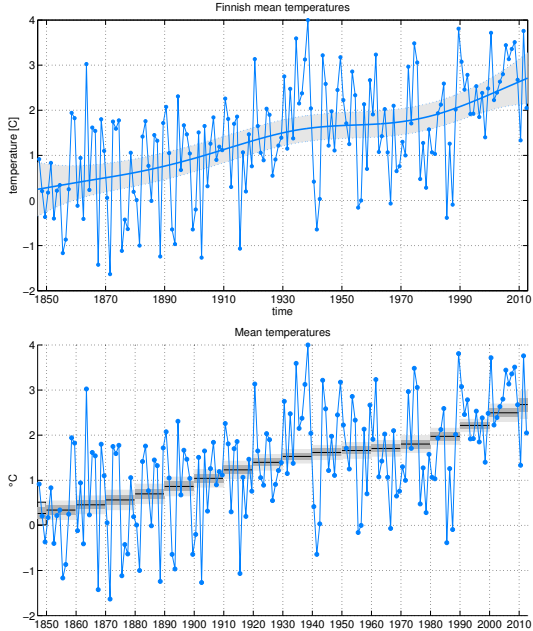
$$-2\log(p(y_{1:n}|x_{1:n},\theta)) = \text{constant} + \sum_{t=1}^{n}\left[(y_t - F_t\widehat{x}_t)^T C_t^{y-1}(y_t - F_t\widehat{x}_t) + \log(|C_t^*|)\right]$$

- Maximum aposteriori by optimization, or Markov chain Monte Carlo (MCMC) to sample from the posterior distribution.

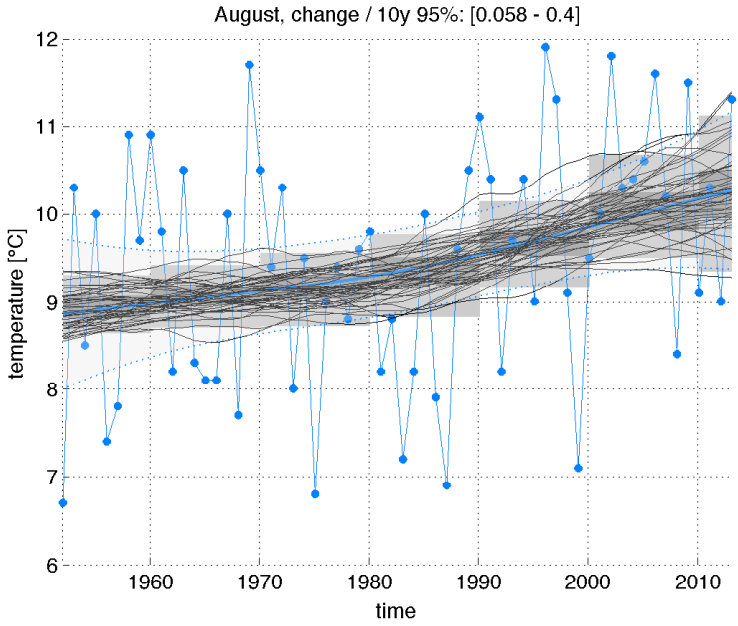
# Matlab toolbox for DLM

- For DLM calculations at `http://helios.fmi.fi/~lainema/dlm`.
- You will need the MCMC toolbox, also, `http://helios.fmi.fi/~lainema/mcmc`.
- DLM tutorial at `http://helios.fmi.fi/~lainema/dlm/dlmtut.html`.

# Finnish temperatures



# Kilpisjärvi summer temperatures Matlab demo



# Conclusion

- DLM framework offers flexible, intuitive and statistically sound way to analyse trends in time series.
- Matlab toolbox for DLM calculations for time series at `http://helios.fmi.fi/~lainema/dlm`.
- Kyrölä, E., Laine, M., Sofieva, V., Tamminen, J., Päivärinta, S.-M., Tukiainen, S., Zawodny, J., and Thomason, L. (2013): Combined SAGE II-GOMOS ozone profile data set for 1984–2011 and trend analysis of the vertical distribution of ozone *Atmos. Chem. Phys.*, 13, 10645–10658.
- Laine, M., Latva-Pukkila, N., and Kyrölä, E. (2013): Analyzing time varying trends in stratospheric ozone time series using state space approach, *Atmos. Chem. Phys. Discuss.*, 13, 20503–20530.
- Mikkonen, S., Laine, M., Mäkelä, H., Gregow, H., Tuomenvirta, H., Lahtinen, M., Laaksonen, A. (2015): Trends in the average temperature in Finland, 1847–2013, *Stochastic Environmental Research and Risk Assessment*, 29(6), 1521–1529.
- Roininen, L., Laine, M., Ulich, T. (2015): Time-varying ionosonde trend: Case study of Sodankylä hmF2 data 1957–2014, *Journal of Geophysical Research: Space Physics*, 120.