How to study trends if you must -

modelling trends in time series using the dynamic linear model approach

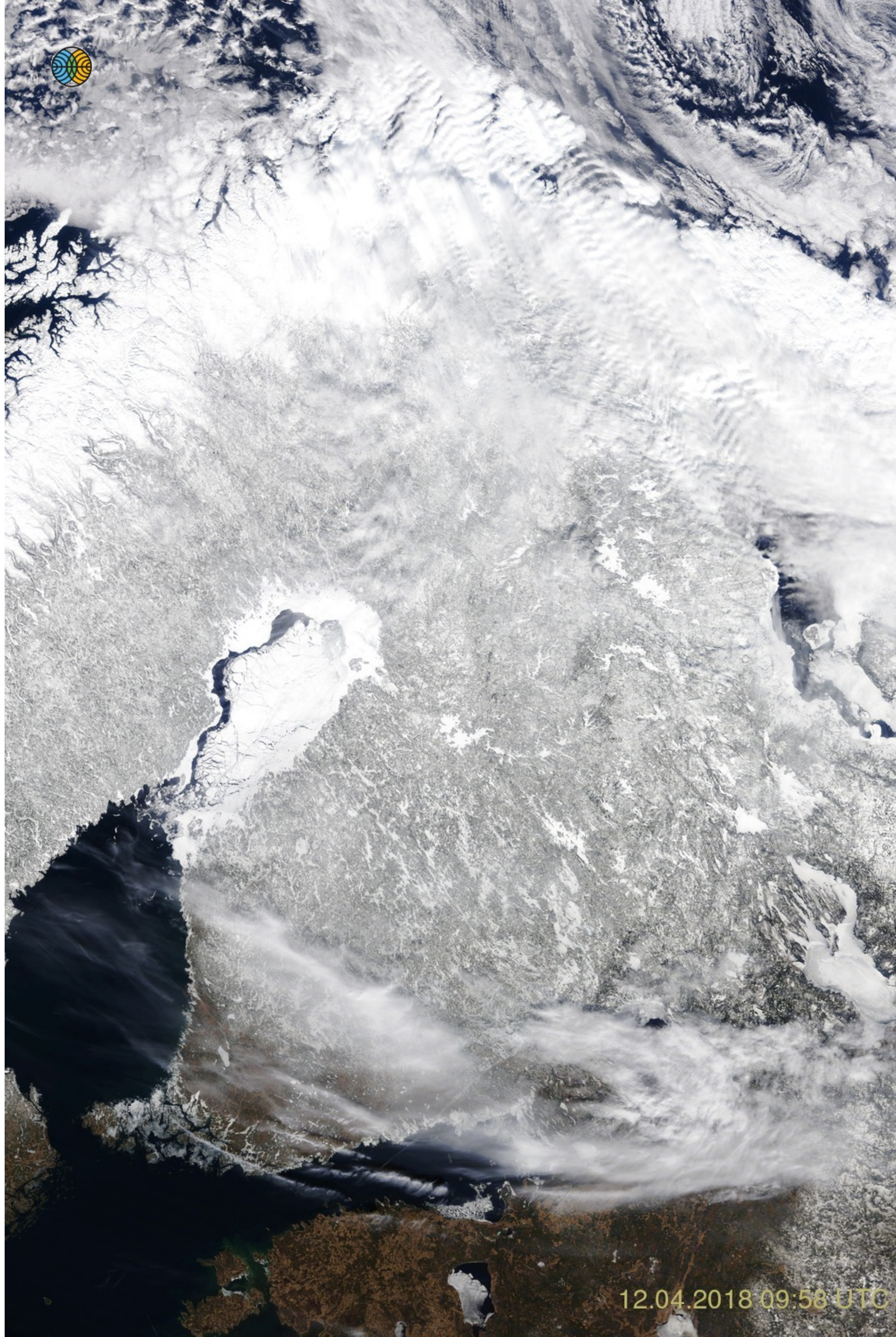
Marko Laine

- Finnish Meteorological Institute
- DTU seminar 2018-12-18

In this presentation

- **Inverse problems** related to different levels of satellite data.
- **Time series analysis** for environmental time series.
- **Dynamic linear model** (DLM) time series analysis by Kalman smoother and MCMC.
- **Data fusion** of satellite and in-situ data by DLM.
- [**Dimension reduction** techniques for data fusion]

On right: RGB True Color image of Finland 12. April 2018 by EOS-Terra satellite, MODIS instrument, http://fmiarc.fmi.fi/latestSat.php.



Academy of Finland Centre of Excellence in Inverse Methods and Imaging 2018-2025

- Continues the CoE of Inverse Problems research.
- Jointly with 6 Finnish Universities and FMI.

Inverse problems in atmospheric remote sensing

Atmosphere

Spectroscopy

Surface

Radiative transfer
absorption and scattering

Prior / regularization

Fitting algorithm

Uncertainty quantification

Noise, integration

Instrument function

Spectral stability

Instrument

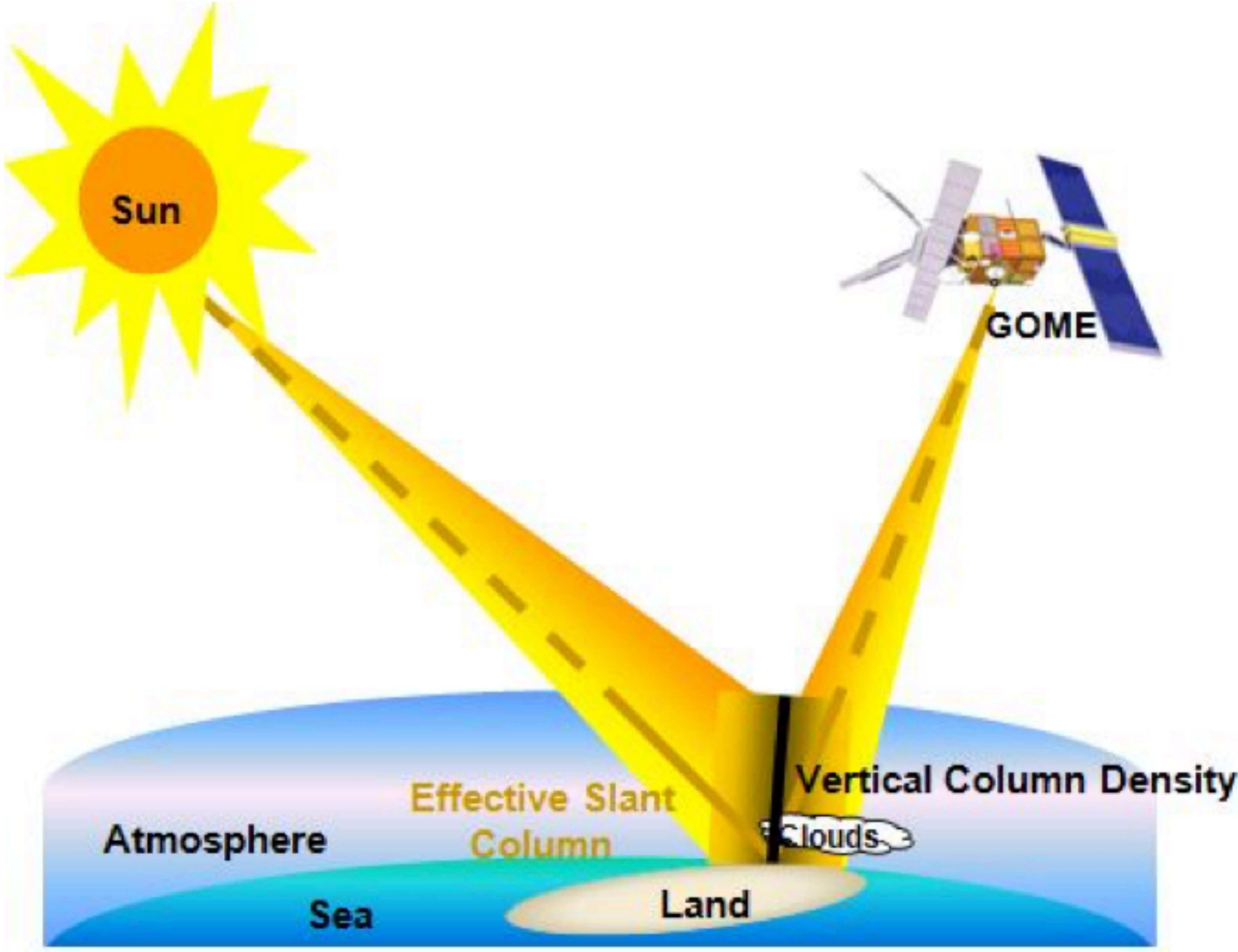
Satellite data (Level 1b)

Retrieval algorithm

Level 2 data: O_3, NO_2, HCHO, SO_2, CO, CH4, CO_2, aerosols, …

Level 3 data, Level 4 data

…



Scattered solar light observation

Credit: Johanna Tamminen

Satellite data processing levels

- **Level 1**: Reconstructed, unprocessed instrument data (e.g. radiances) at full resolution, and annotated with ancillary information. Input for the retrieval algorithm.
- **Level 2**: Retrieved (by inversion algorithm) geophysical variables at the same resolution and location as Level 1. (e.g. **vertical constituent profiles**).
- **Level 3**: Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.
- **Level 4**: Model output or results from analyses of lower-level data (e.g. **time series**, **data fusion**, assimilation).

Approaches to atmospheric inverse problems

Optimal estimation	hierarchical statistical model	conditional probabilities
forward model F CO_2 (x)→radiance (y) inverse problem	data model $Y = F(X\|\theta) + \varepsilon$	$p(Y\|X, \theta)$
prior x_α, S_α smoothness etc	process model	$p(X\|\theta)$
fixed tuning parameters	parameter model	$p(\theta)$
optimal? L2 loss → conditional mean 0-1 loss → MAP		$p(Y,X,\theta) = p(Y\|X,\theta)\, p(X\|\theta)\, p(\theta)$ $p(X\|Y,\theta) \propto p(Y\|X,\theta)p(X\|\theta)$

Components of the atmospheric inverse problem

What is X?

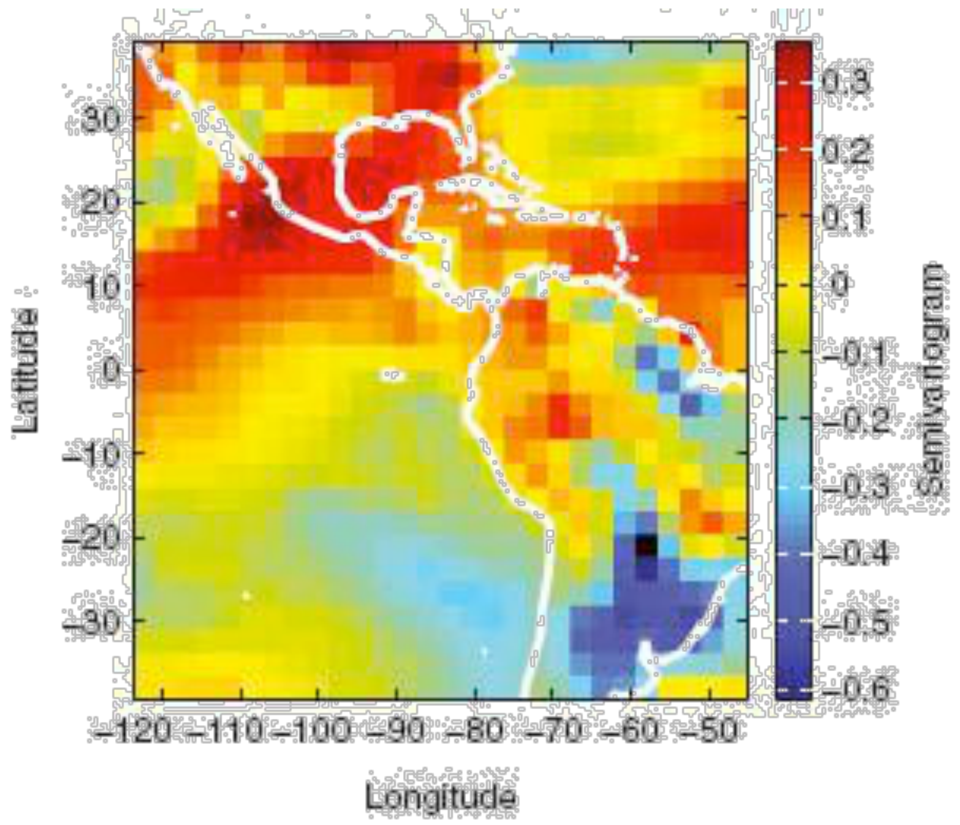
- 4D field (lat,lon,alt,time)
- might be interested in g(X), total column, surface flux, ..
- retrieved individual values $P(\hat{X}_i|Y_i,\theta)$ are not independent
- independent sounding by sounding retrieval not ok?

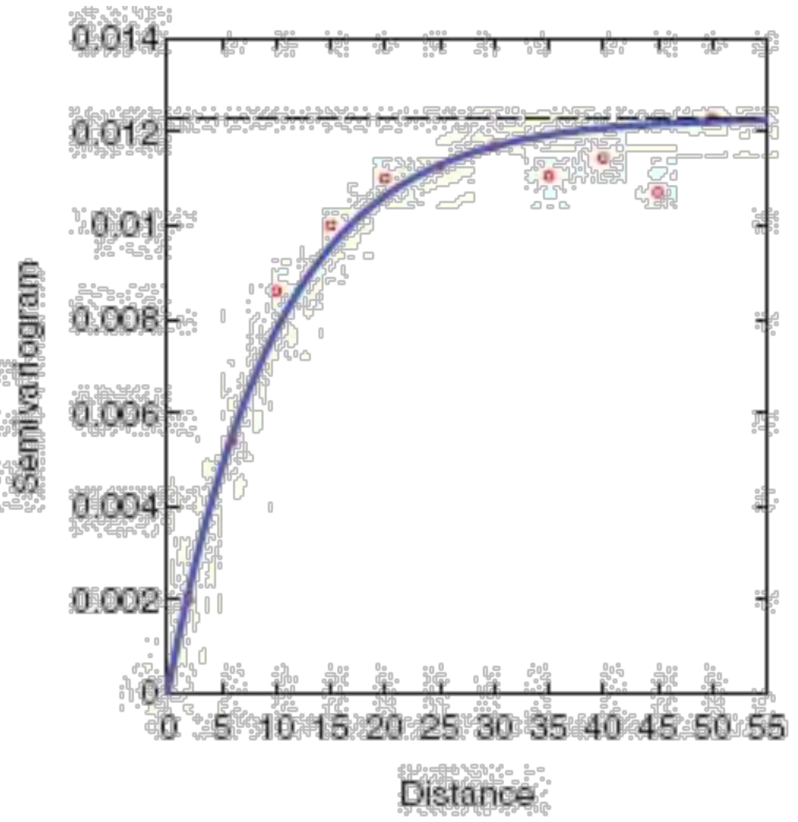
What is Y?

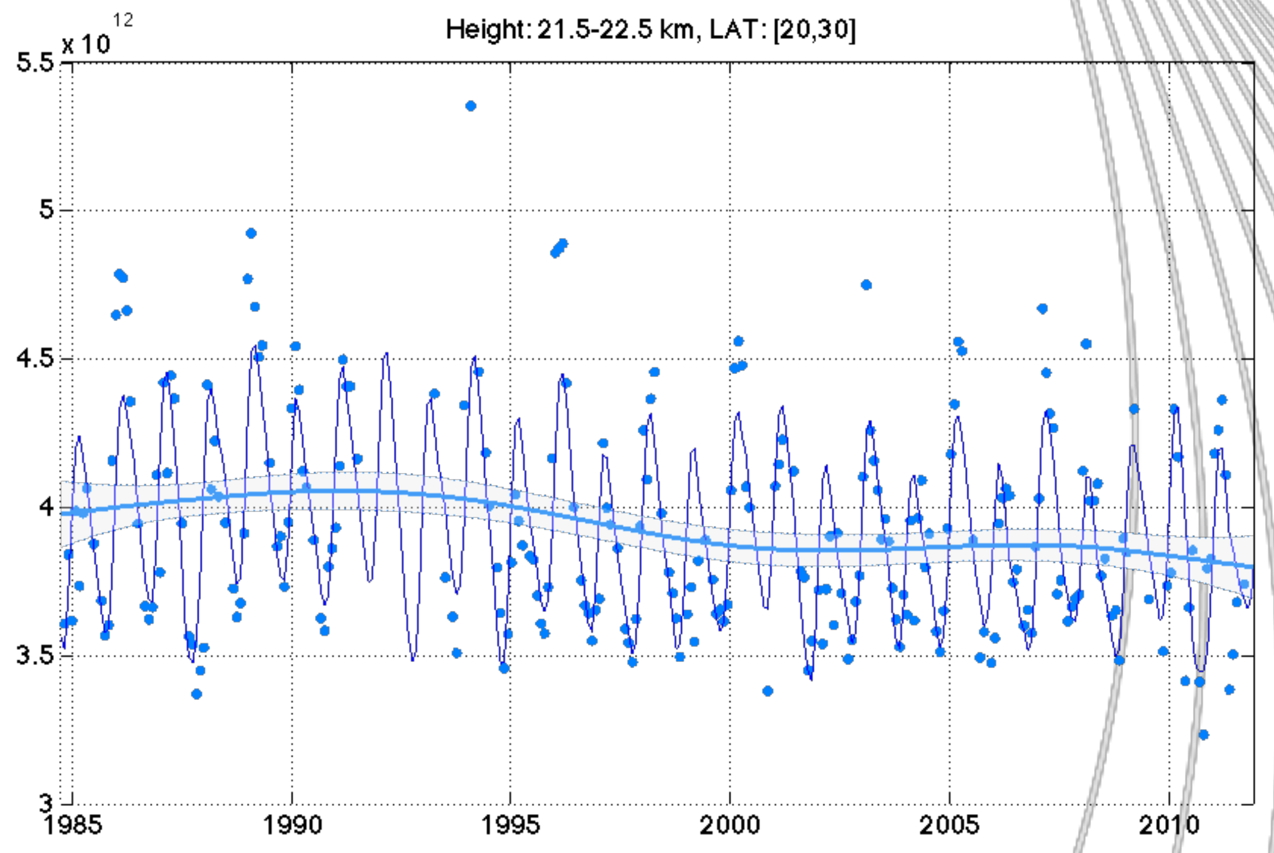
- radiances (Level 1)
- or retrieved individual CO2 columns $\tilde{Y}_i$ (Level 2)
- $P(\tilde{Y}_i|X,\theta)$ conditionally independent
- $P(X|\tilde{Y}, \theta) \propto P(\tilde{Y}|X, \theta)\, p(X|\theta)$ (Level 3-4)
- still need the process model $p(X|\theta)$

What is $p(X|\theta)$?

- spatio-temporal process model
- prior $X_i \sim N(X_{\alpha i}, S_{\alpha i})$
- GCM, CTM, statistical models GP, GRMF
- spatial statistics tools needed

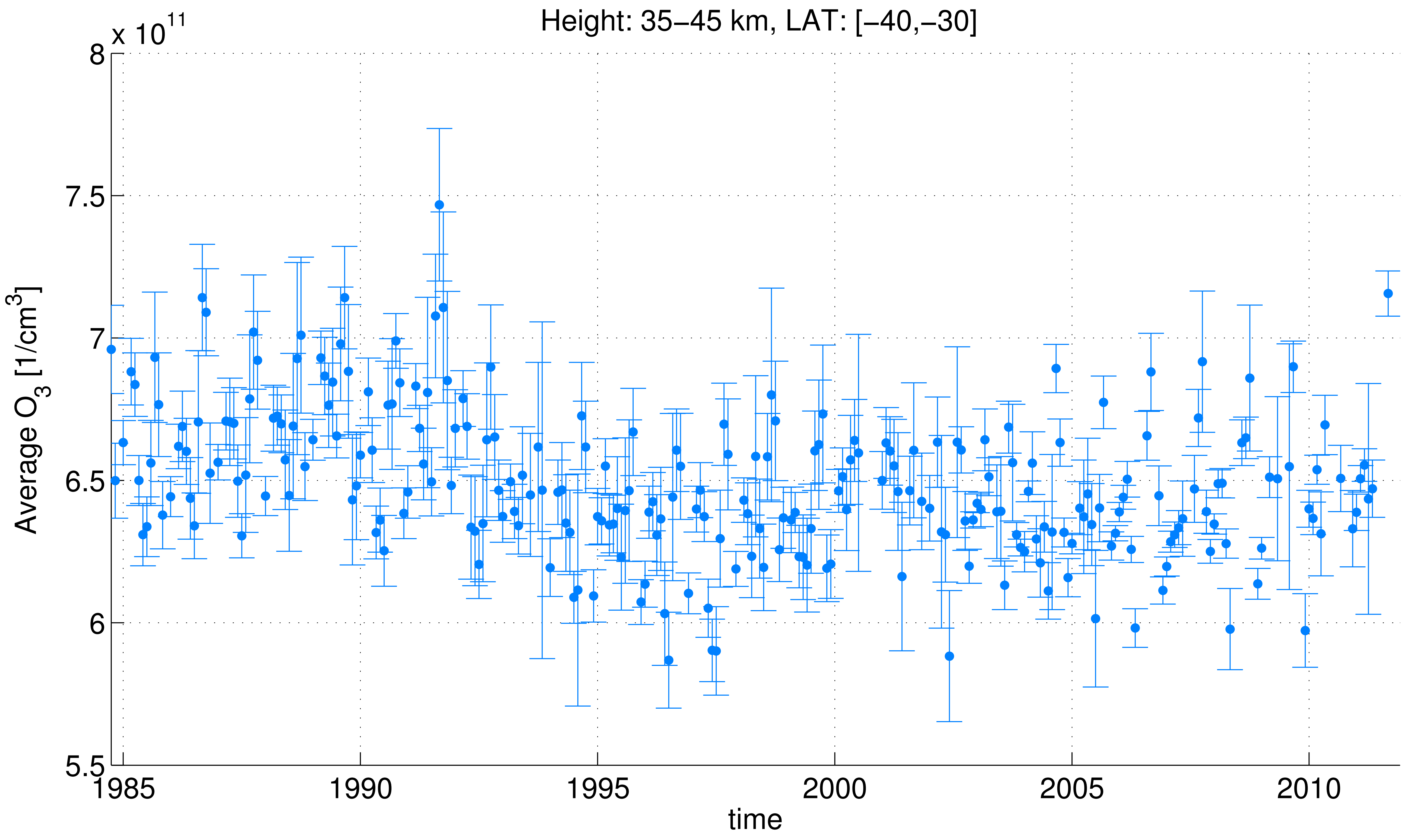






Time series analysis — example 1

- Has stratospheric ozone recovered from human caused depletion by CFC compounds?

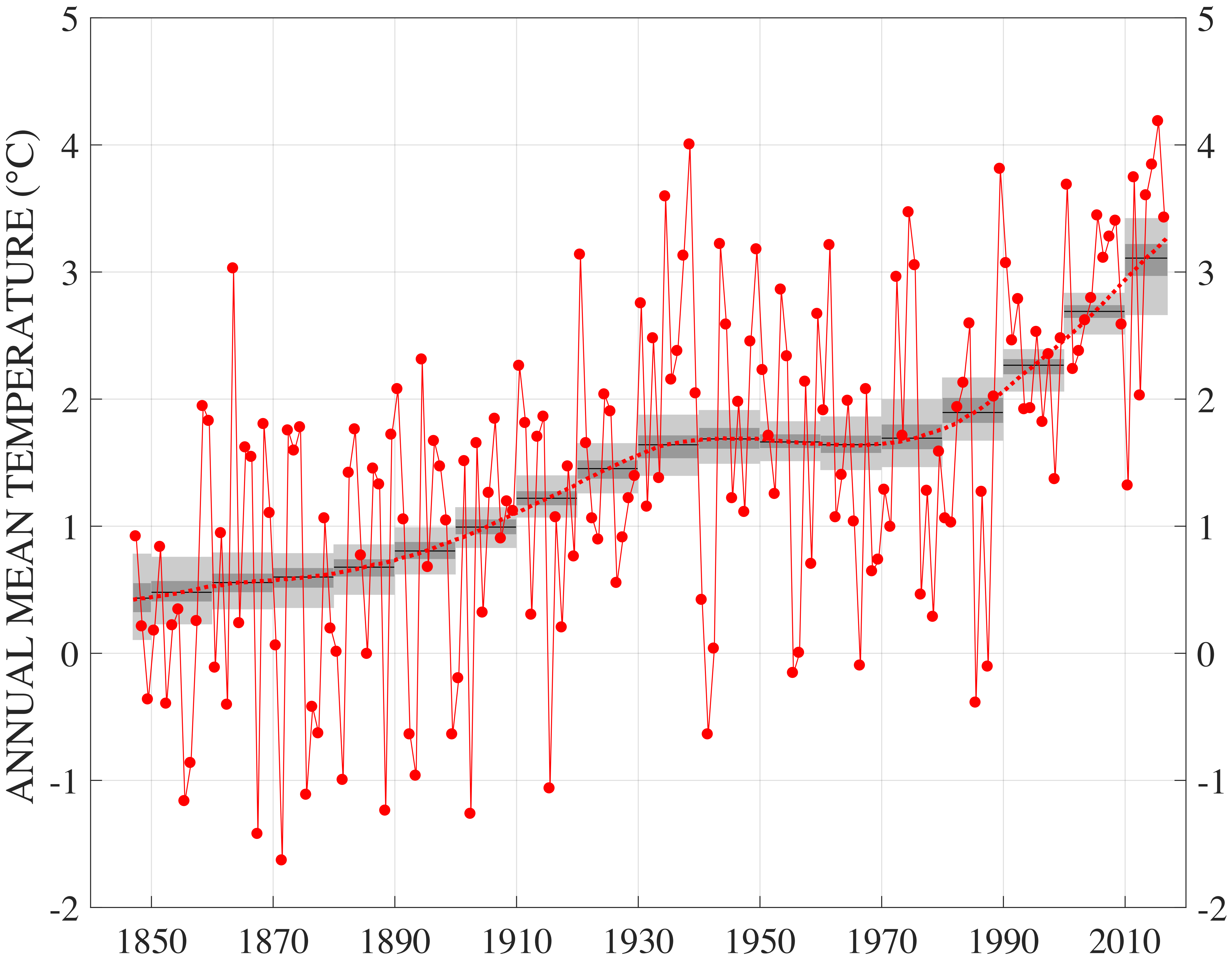


- Answer: recovery started from year 1997.

Time series analysis — example 2

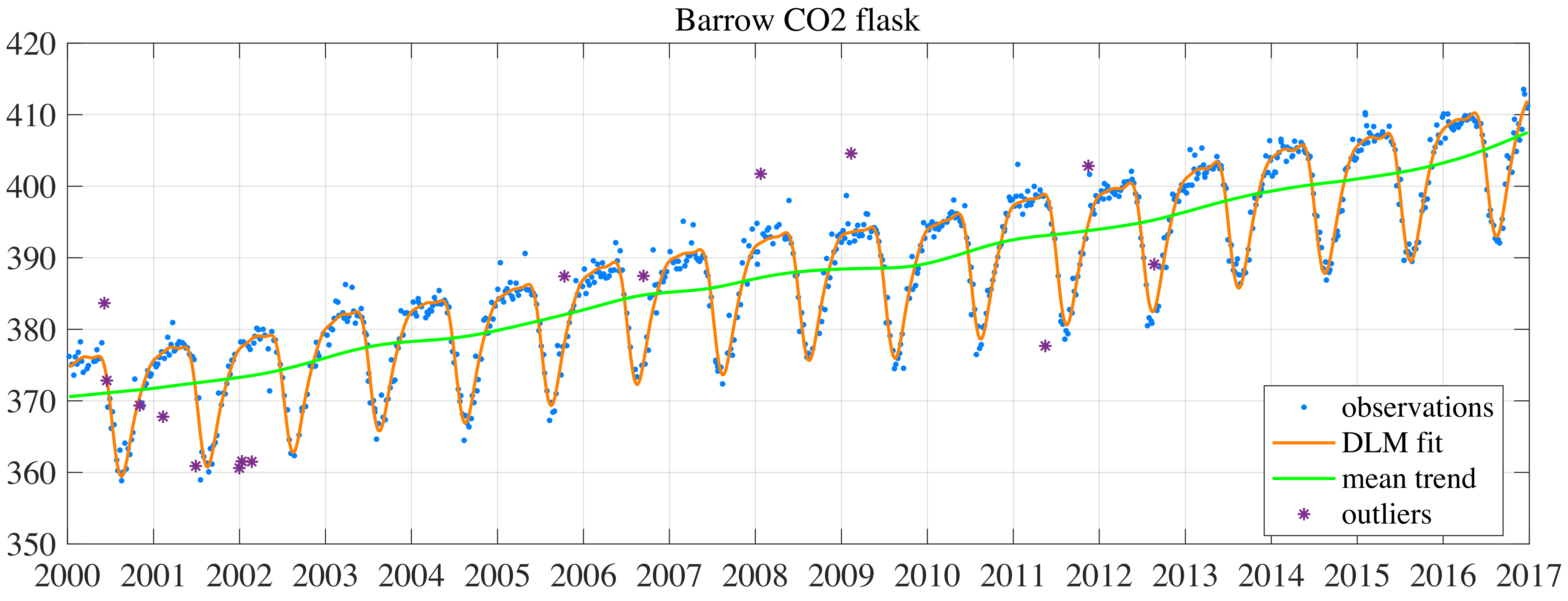
- Can the increase in the temperatures in Finland be attributed to natural variability?

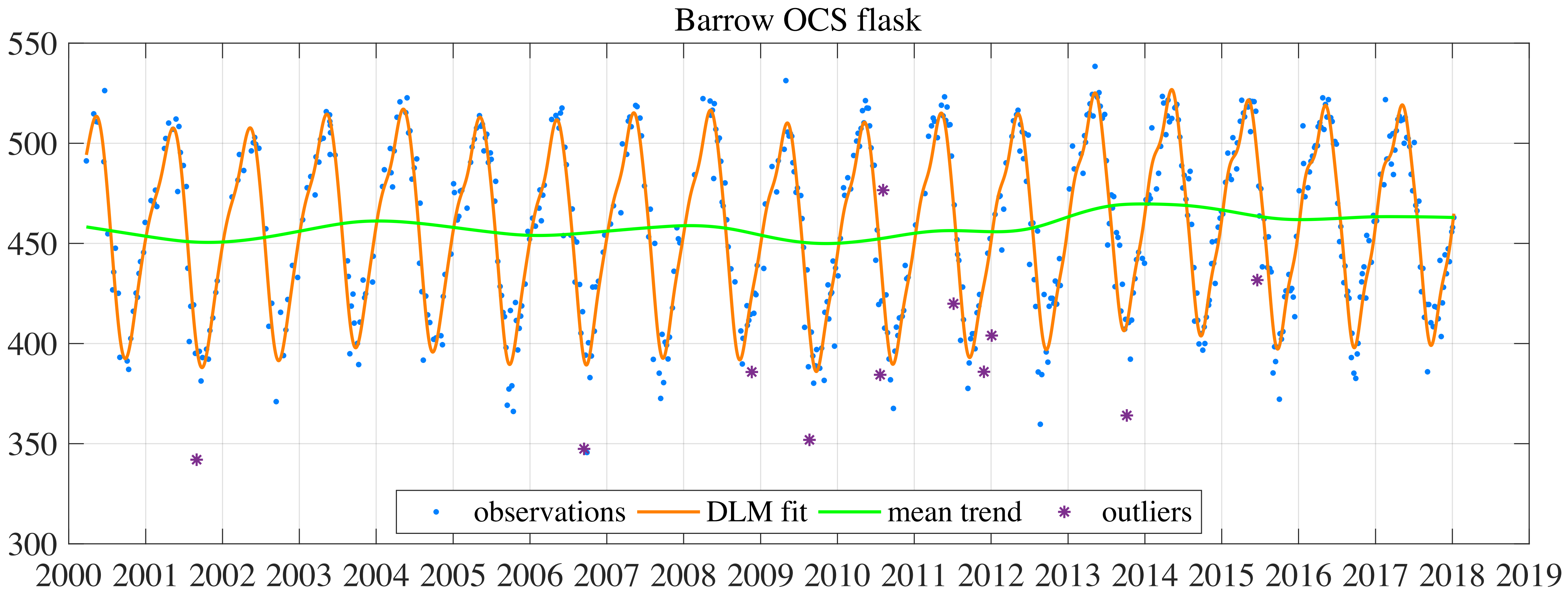
- Answer: no.

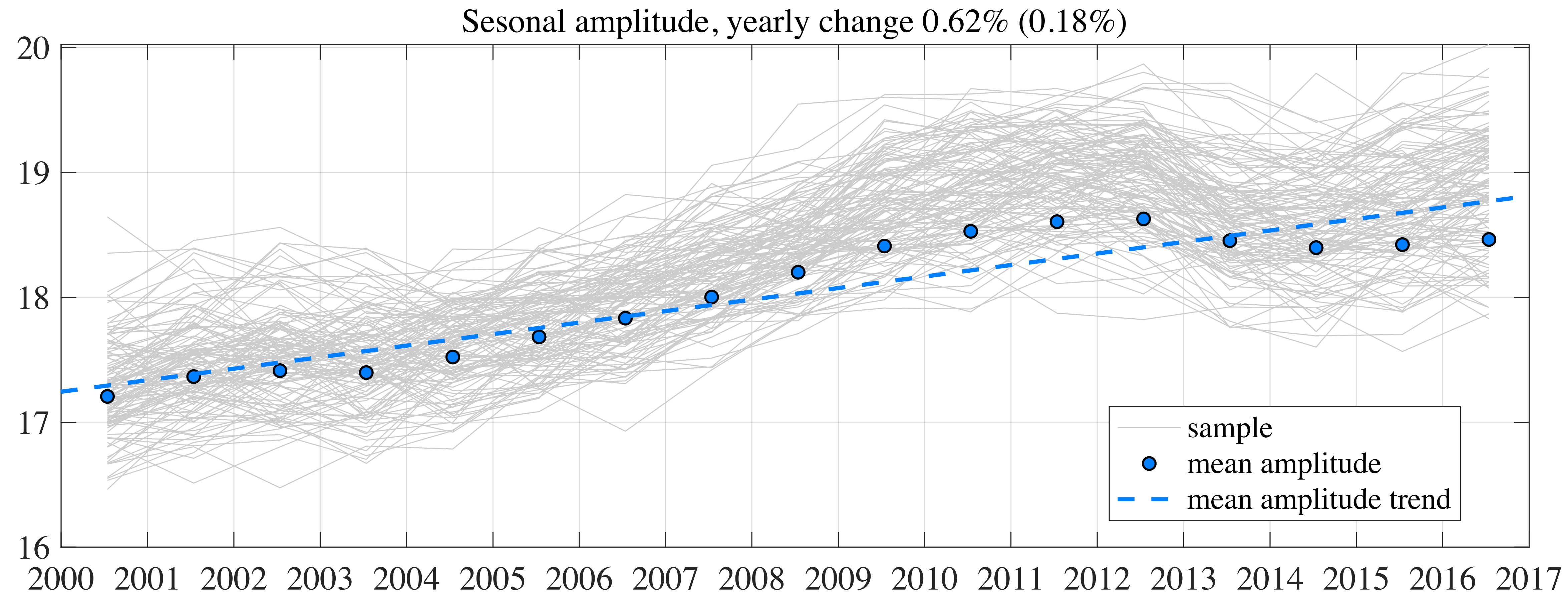


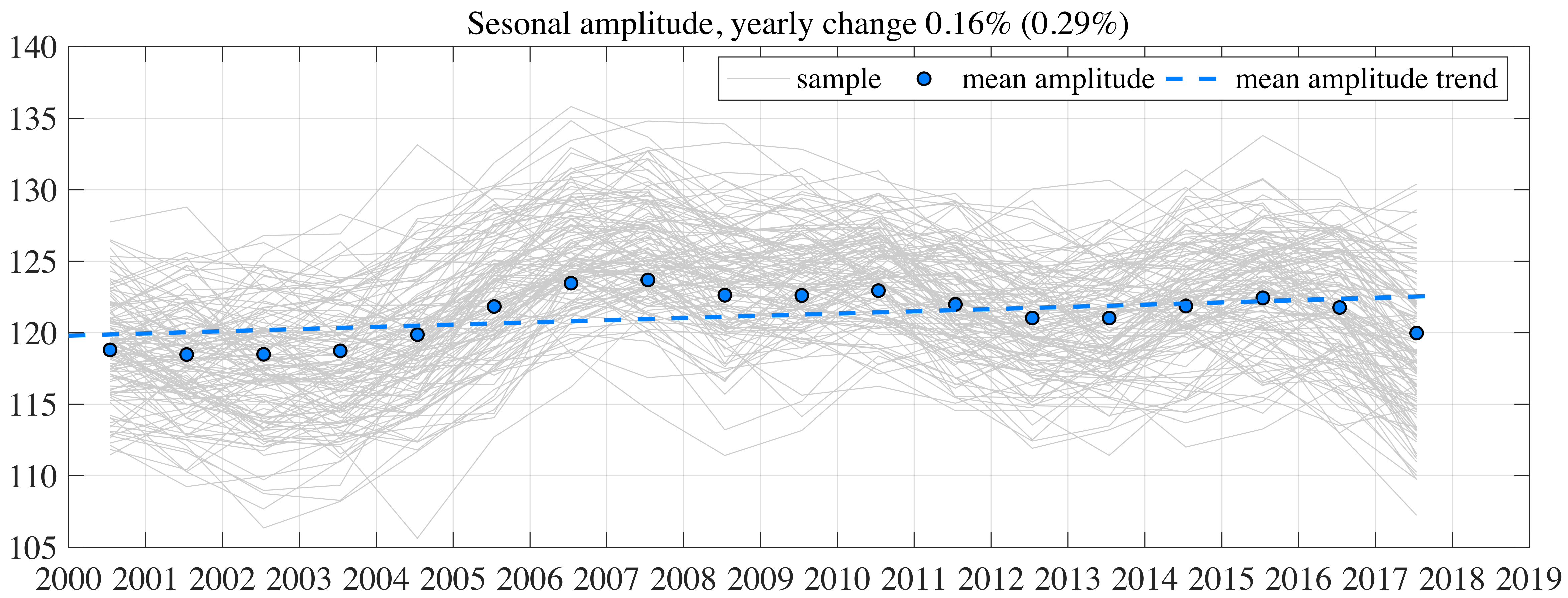
Time series analysis — example 3

- Is there growth in seasonal amplitude of atmospheric CO_2 and OCS?





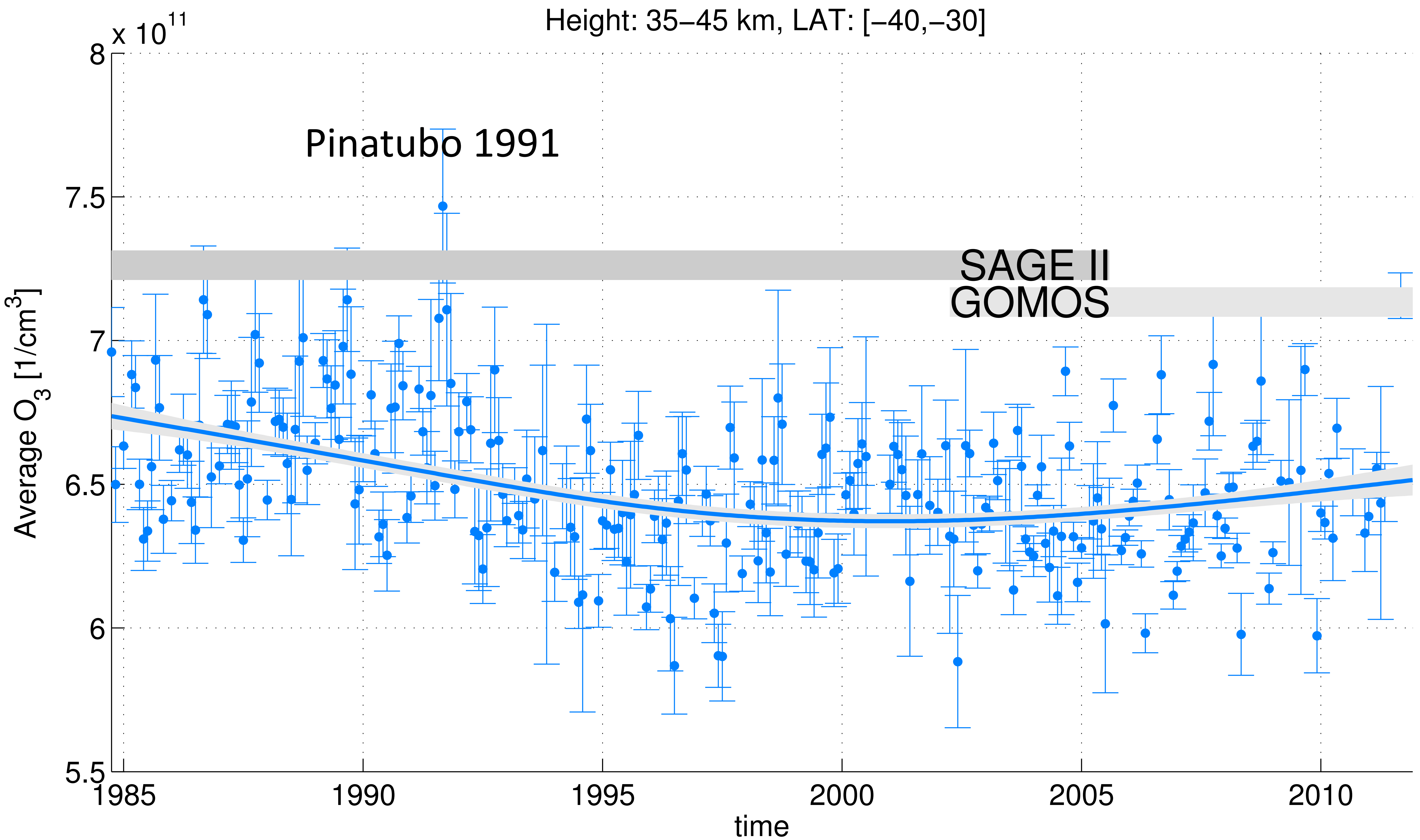




Challenges in climatic time series

Seasonality
Change of the background level
Data gaps, different instruments, volcanic events

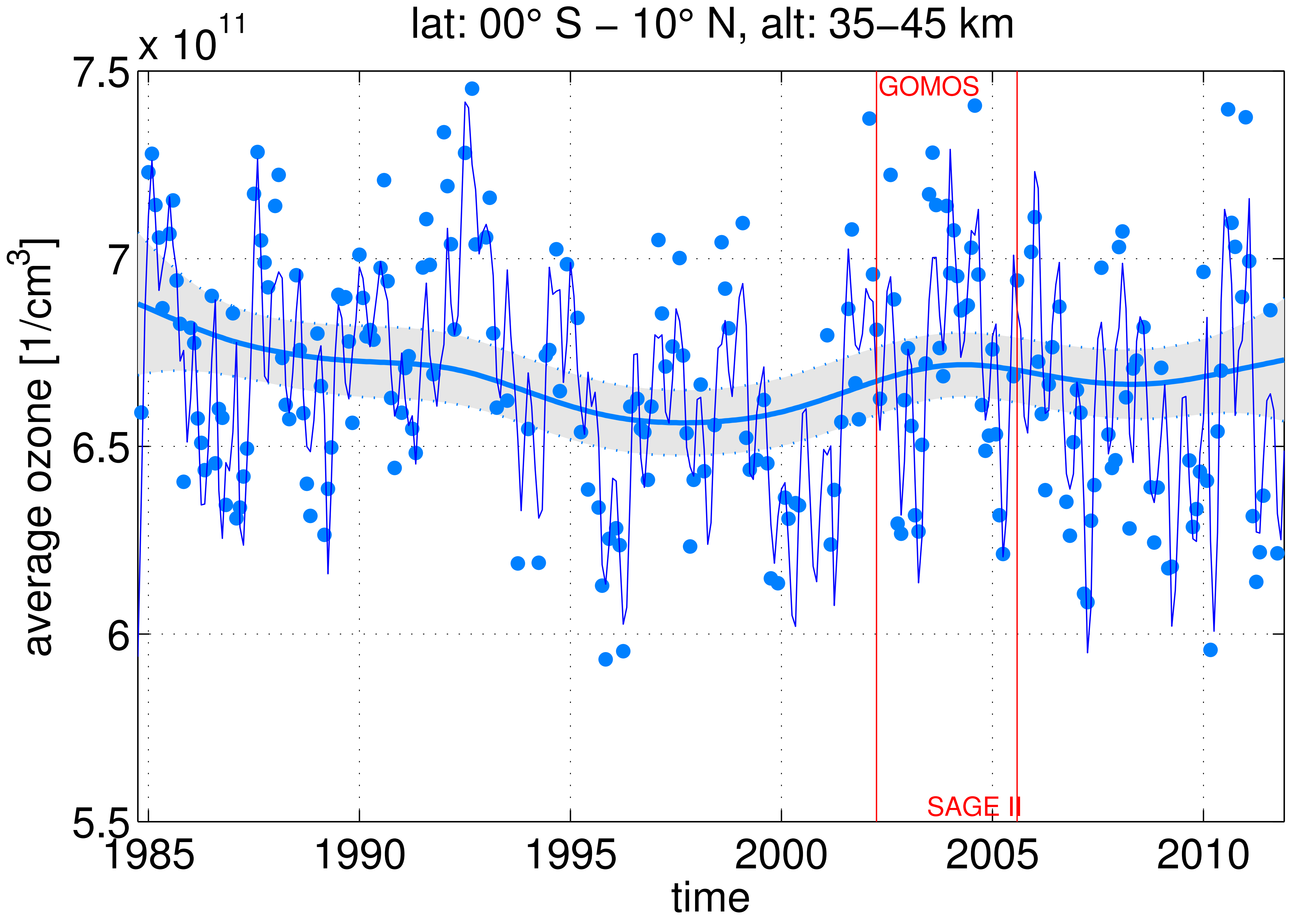
Components of ozone time series …



What is trend?

- Trend is a change in the background mean level of the process.
- For example: we are interested in smooth long term (decadal) change attributed to ozone recovery.
- Need to model seasonality, external forcing driven by known phenomena, long range correlations, …

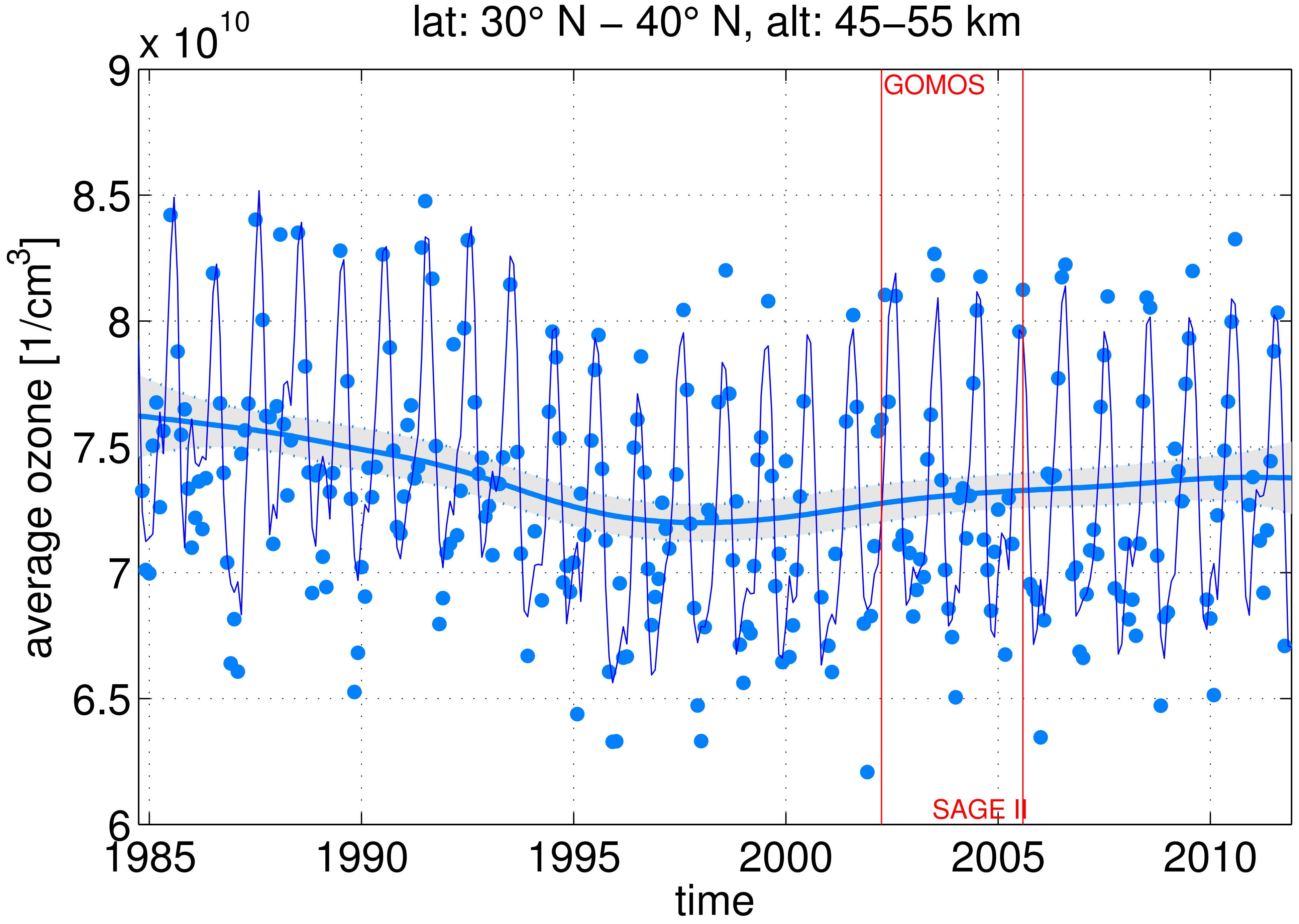
- **Goal: a statistical model consistent with the observed variability.**



Dynamic linear model (DLM)

- General framework for studying dynamical changes in time series data by local regression analysis.
- Uses a state space process description of the model components (trends, seasonality, proxies).
- Suitable for univariate and multivariate time series analysis.

- Includes hierarchical statistical model for uncertainties in data, process, and parameters.
- Verifiable statistical assumptions.



Dynamic linear model (DLM) as a hierarchical statistical model

state space model

$$y_t = H_t x_t + \varepsilon_t \qquad \varepsilon_t \sim N(0,R_t)$$
$$x_t = M_t x_{t-1} + E_t \qquad E_t \sim N(0,Q_t)$$

hierarchical statistical model

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

- y_t: observations,
- x_t: model states,
- H_t: observation operator,
- M_t: model operator,
- ε_t :observation uncertainty,
- E_t :model uncertainty.

- θ: structural and variance parameters in H_t, M_t, R_t, and Q_t.

- Bayes formula:

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

Statistical estimation and analysis by Kalman filter, Kalman Smoother and Markov chain Monte Carlo (MCMC).

Simple example: spline smoothing

$y_t = H_t x_t + \varepsilon_t$ $\quad \varepsilon_t \sim N(0,R_t)$

$x_t = M_t x_{t-1} + E_t$ $\quad E_t \sim N(0,Q_t)$

$y_t = \mu_t + \varepsilon_{obs}$,

$\mu_t = \mu_{t-1} + \alpha_{t-1} + \varepsilon_{level}$,

$\alpha_t = \alpha_{t-1} + \varepsilon_{trend}$,

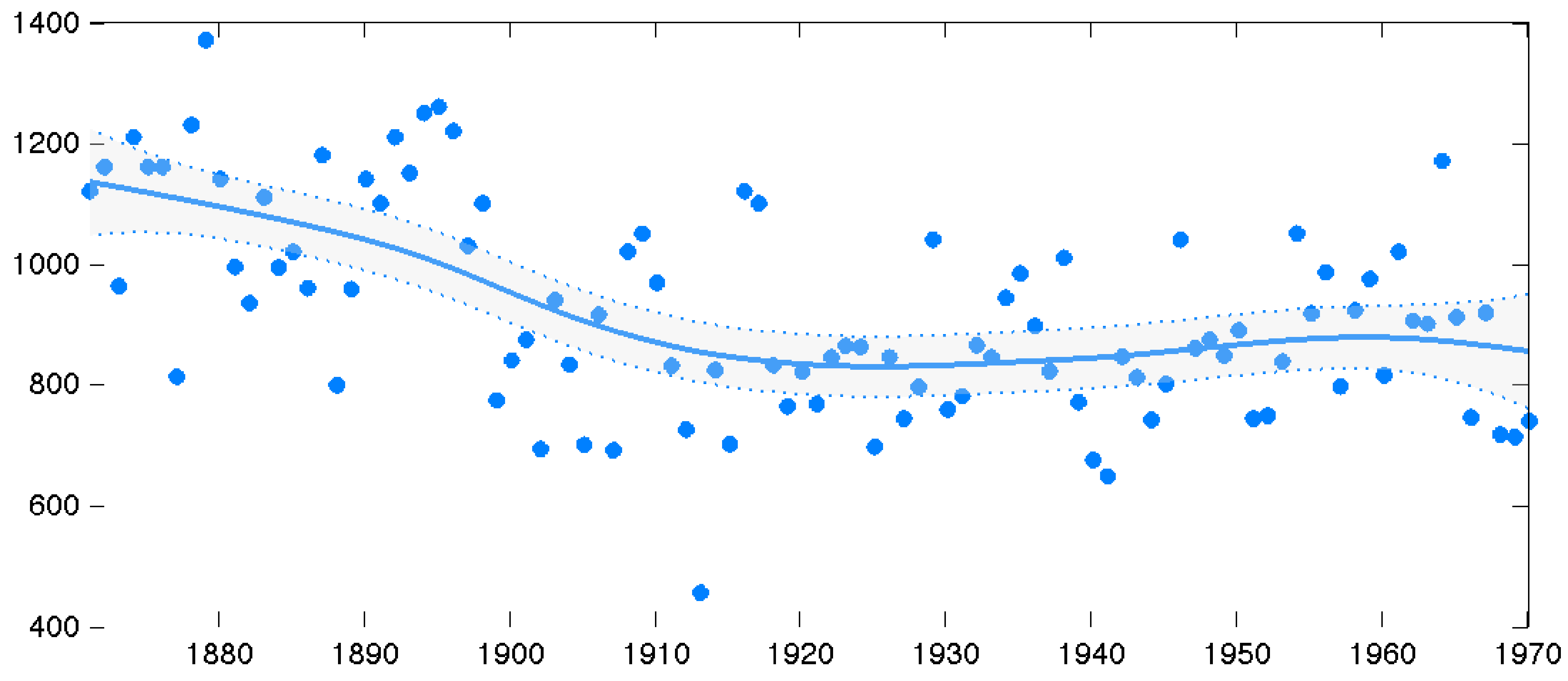
$\varepsilon_{obs} \sim N(0, \sigma^2_{obs})$, observations

$\varepsilon_{level} \sim N(0, \sigma^2_{level})$, local level

$\varepsilon_{trend} \sim N(0, \sigma^2_{trend})$, local trend

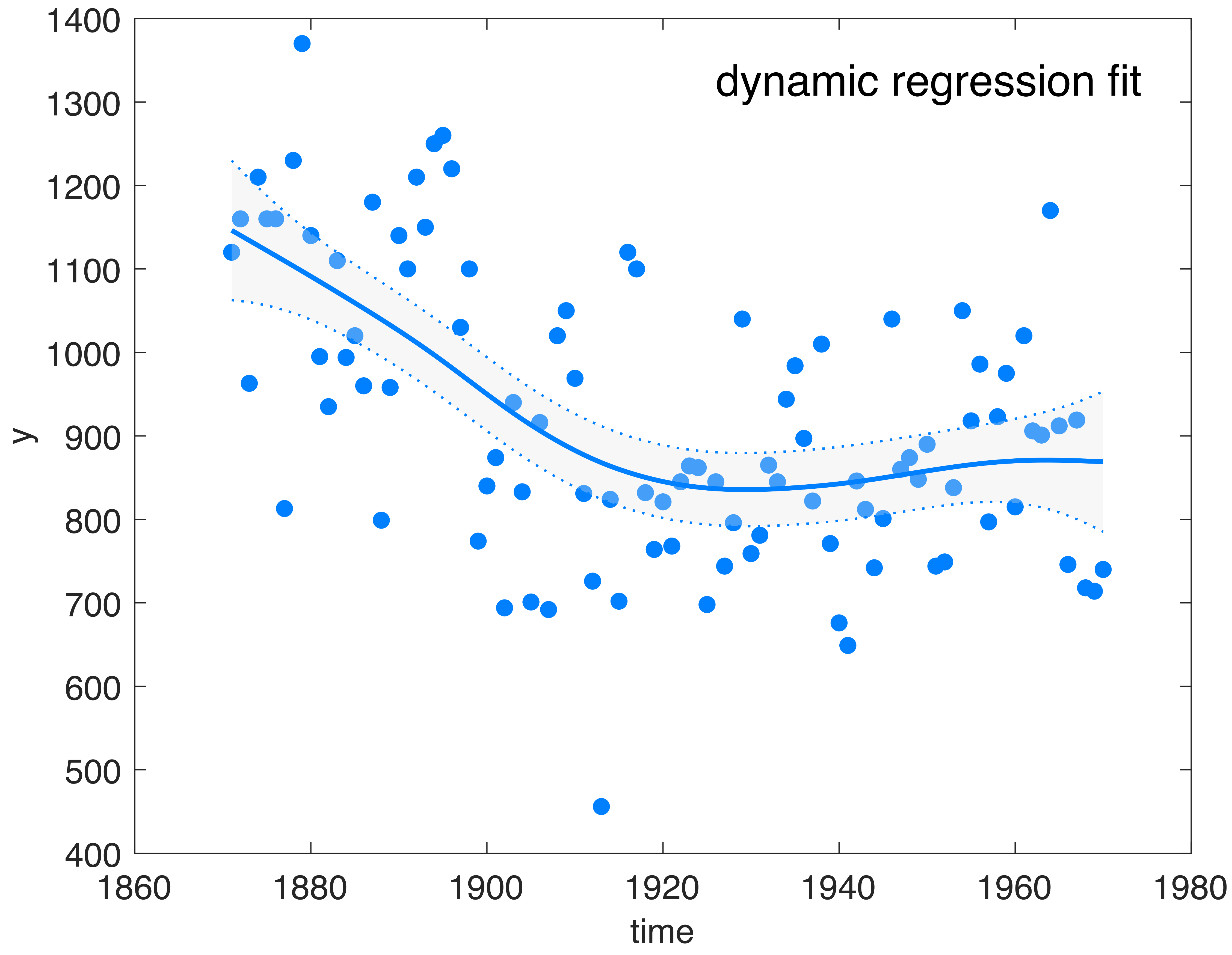
$$M_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad H_t = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad x_t = \begin{bmatrix} \mu_t & \alpha_t \end{bmatrix}^T, \theta = \begin{bmatrix} \sigma^2_{\text{obs}} & \sigma^2_{\text{level}} & \sigma^2_{\text{trend}} \end{bmatrix}^T.$$

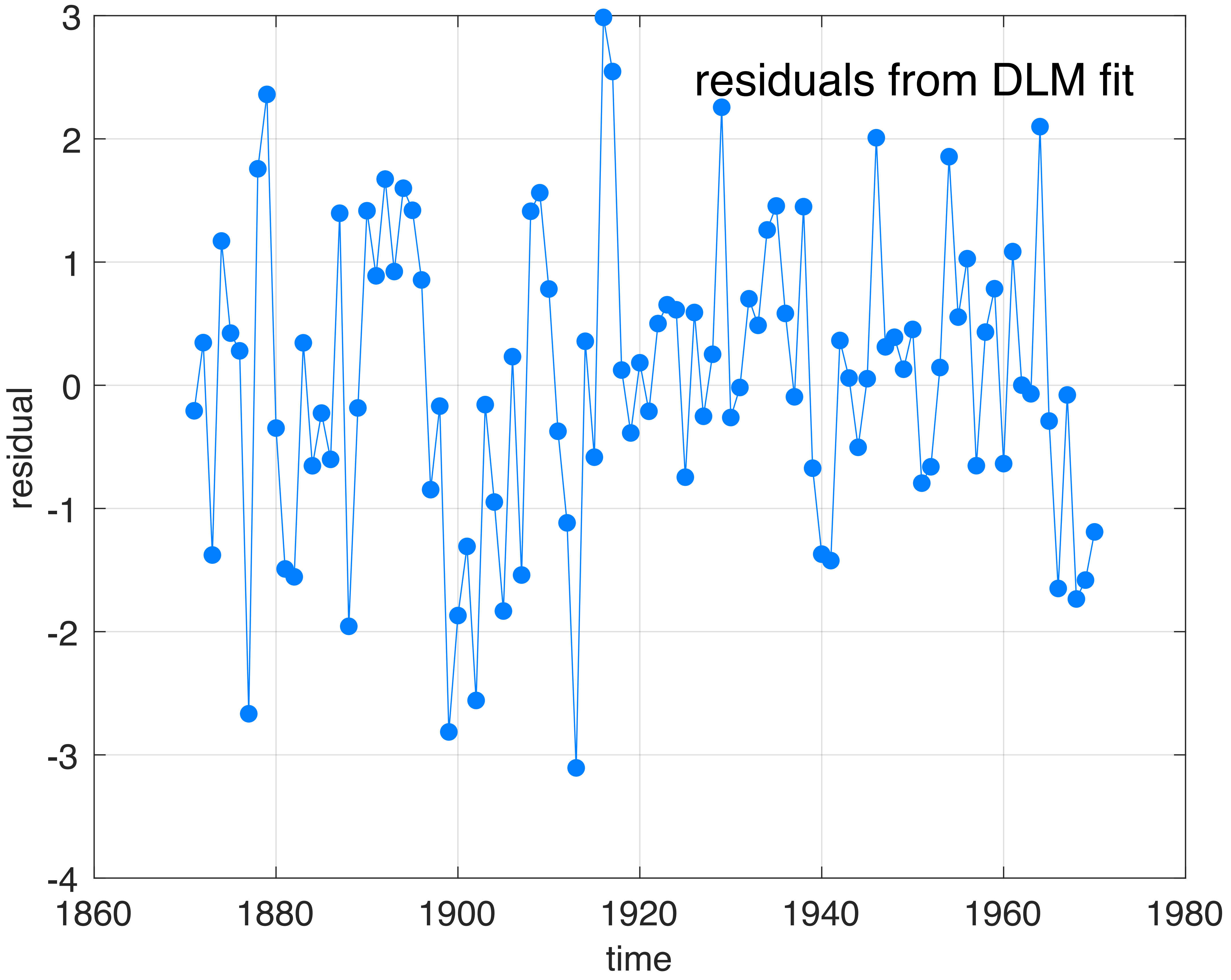
When $\sigma_{level} = 0$, this is cubic spline smoothing with smoothness parameter $\lambda = \sigma^2_{trend}/\sigma^2_{obs}$.



DLM vs. linear regression

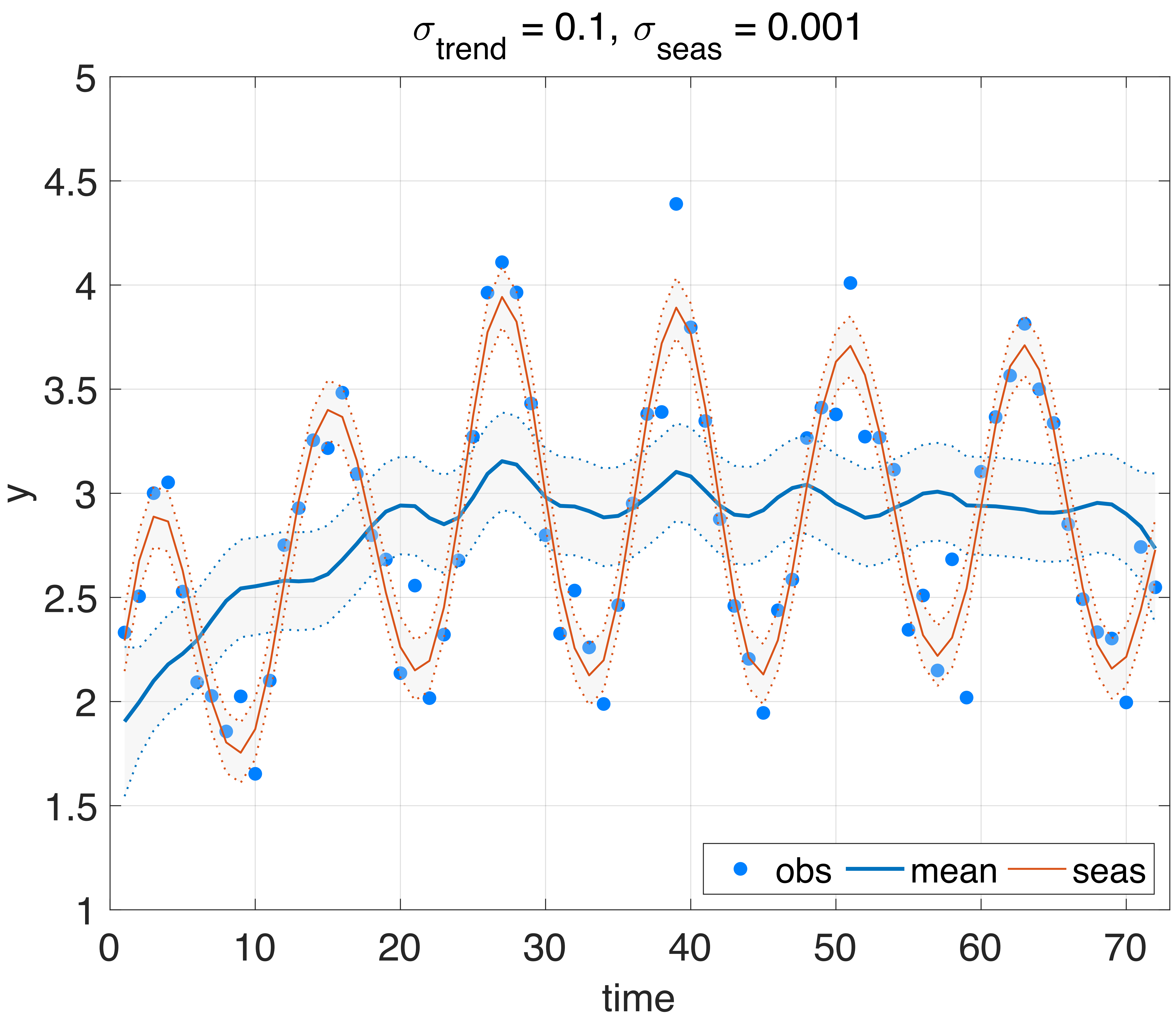
Linear regression is a special case of DLM, with $\sigma^2_{trend} = \sigma^2_{level} = 0$.





Estimating smoothness

- Flexibility comes with a price to pay.
- The extra variance parameters control the smoothness of the fit.
- We look for results consistent with the observations and prior $p(\theta)$, $\theta = [\sigma^2_{trend}, \sigma^2_{seas}]^T$
- Bayesian hierarchical modelling allows estimation by optimization or by MCMC.



Bayesian data analysis

- Bayesian model forces you to think how the observations are generated.
- This involves both prior and the likelihood, jointly.

- Observations simulated from the (prior or posterior) model should look plausible.

- Hierarchy: data model, process model, parameter model.

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

- DLM is a model for the systematic part and the prior, not just for the noise.

- The state space descriptions is closely related to data assimilation in, e.g., numerical weather forecasting.

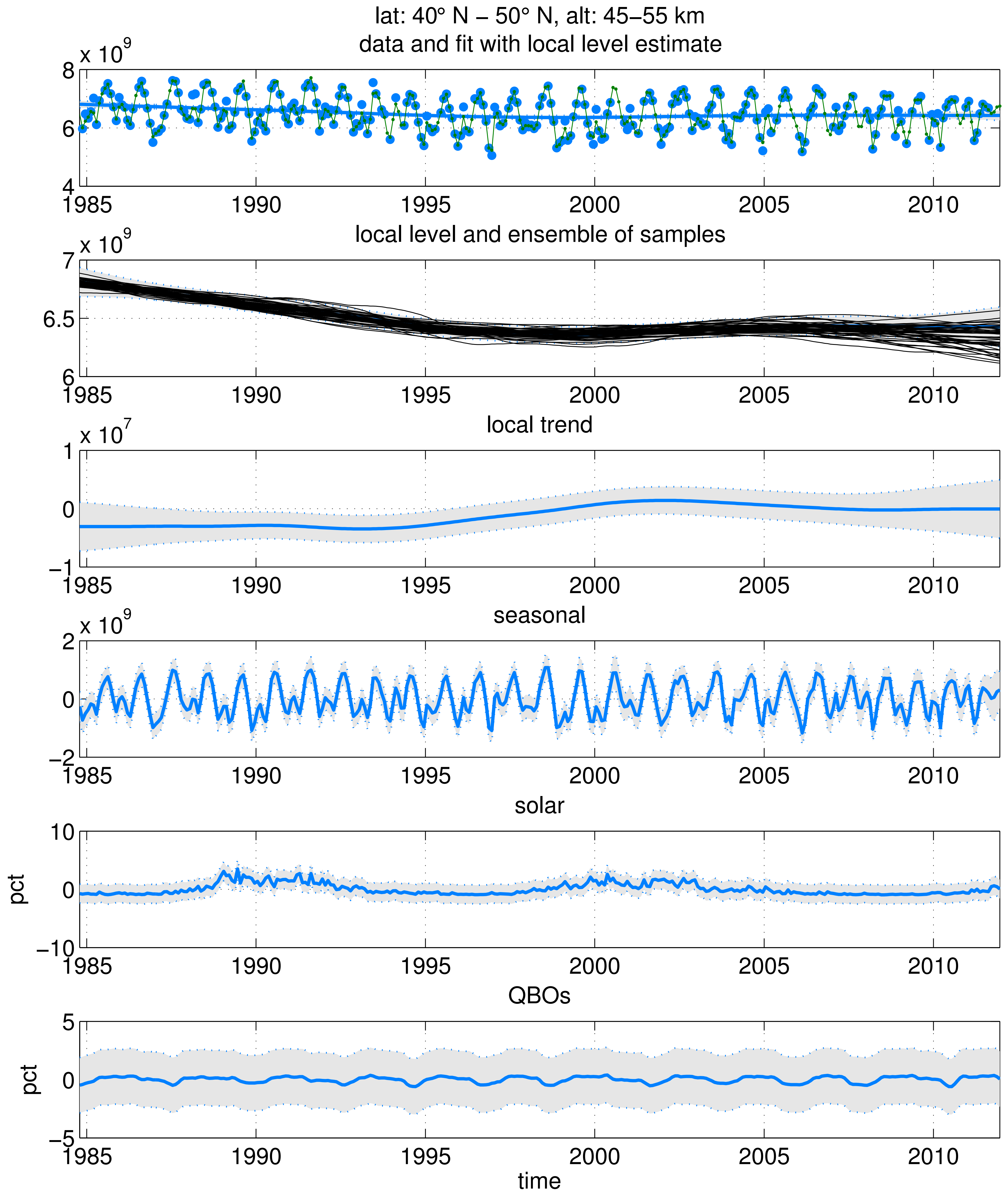
$$y_t = H_t x_t + \varepsilon_t \qquad \varepsilon_t \sim N(0,R_t)$$
$$x_t = M_t x_{t-1} + E_t \qquad E_t \sim N(0,Q_t)$$

General model for trend, seasonality, AR error, proxies

$$y_t = \mu_t + \gamma_t + \beta_t X_t + \eta_t + \varepsilon_{obs,t}$$

μ_t: background level, the trend,
γ_t: seasonal effect,
β_t: coefficient for proxy covariates X_t,
η_t: autoregressive error term,
$\varepsilon_{obs,t}$: observation uncertainty.

All model components are defined by suitable model operator M_t and can depend on time index t.



The system matrices involved

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

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

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

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

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

Model for stratospheric ozone with local level, two harmonic seasonal components, and solar and QBO proxies.

The model ”state” x_t has 9 elements and we have 2 variance parameters in θ.

How to do it in practice?

- The R statistical program has a **dlm** package.
- For python there is **pydlm** and some DLM models in package **statsmodels**.
- Matlab **dlm** toolbox is used for the examples in this presentation.

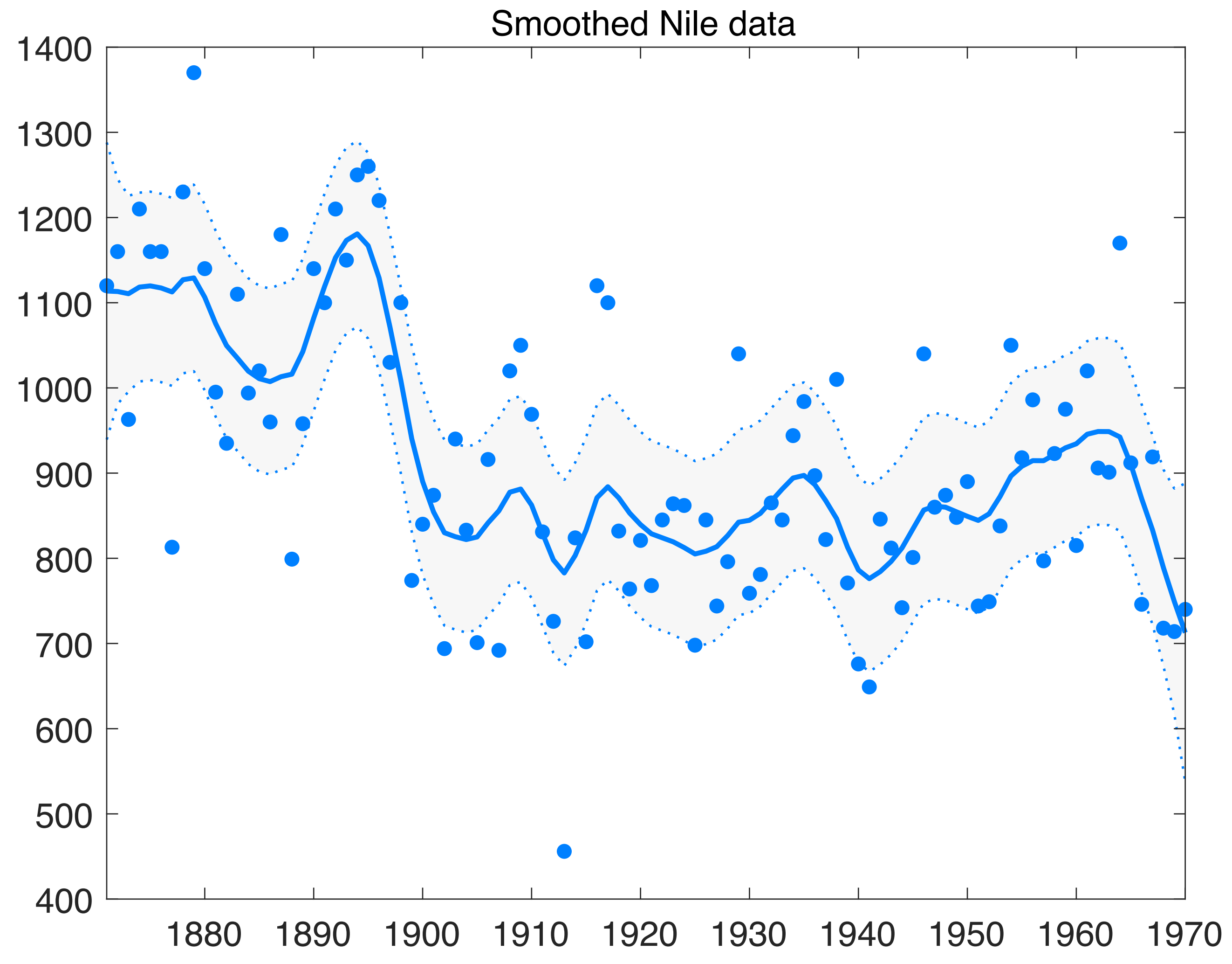
- Some programming skills are needed, as in most data analysis tasks.
- Key aspects in any statistical modelling: visualization, model building, parameter estimation, residual analysis, uncertainty quantification.

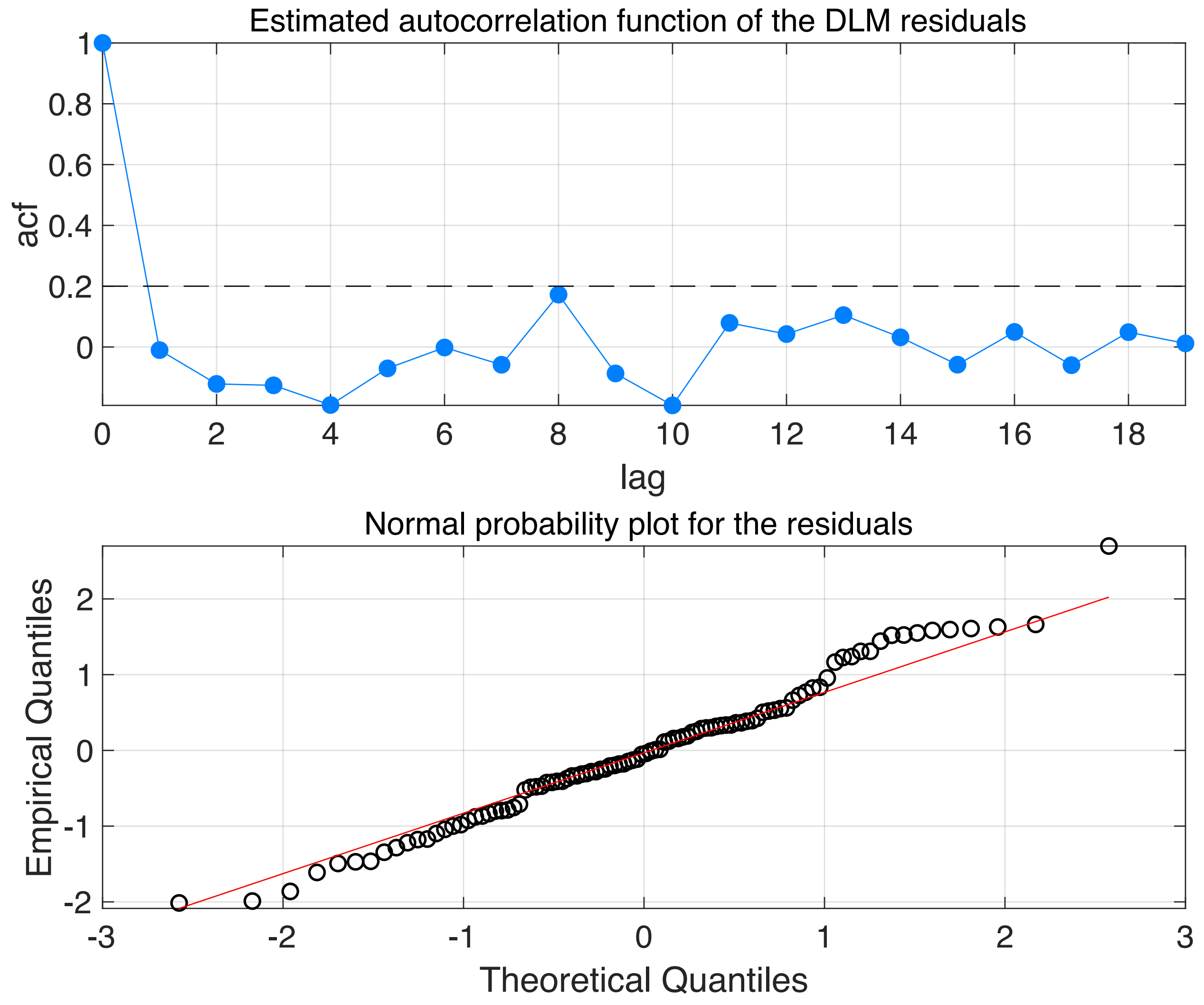
[1] G. Petris, S. Petrone, P. Campagnoli: *Dynamic Linear Models with R*. Springer, 2009.
[2] T.J. Durbin, S.J. Koopman: *Time Series Analysis by State Space Methods*. Oxford University Press, second edition, 2012.
[3] A.C. Harvey: *Forecasting, structural time series and the Kalman filter*. Cambridge University Press, 1990.

Dynamic linear model Matlab toolbox

- DLM toolbox at http://helios.fmi.fi/~lainema/dlm, https://github.com/mjlaine/dlm

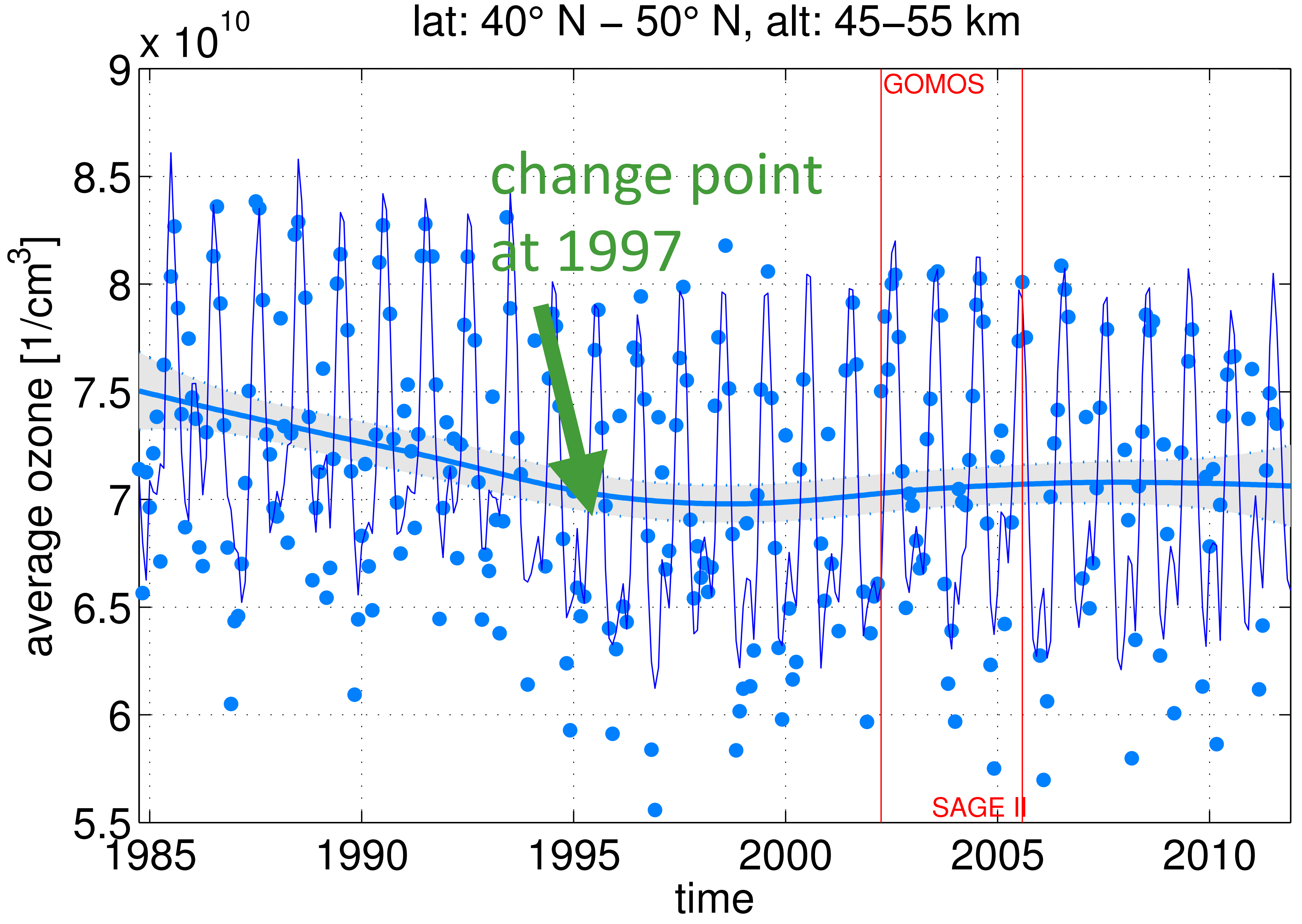
```
out = dlmfit(y,s,w);
dlmplotfit(out,t); title('Smoothed Nile data')
dlmplotdiag(out,t);
```

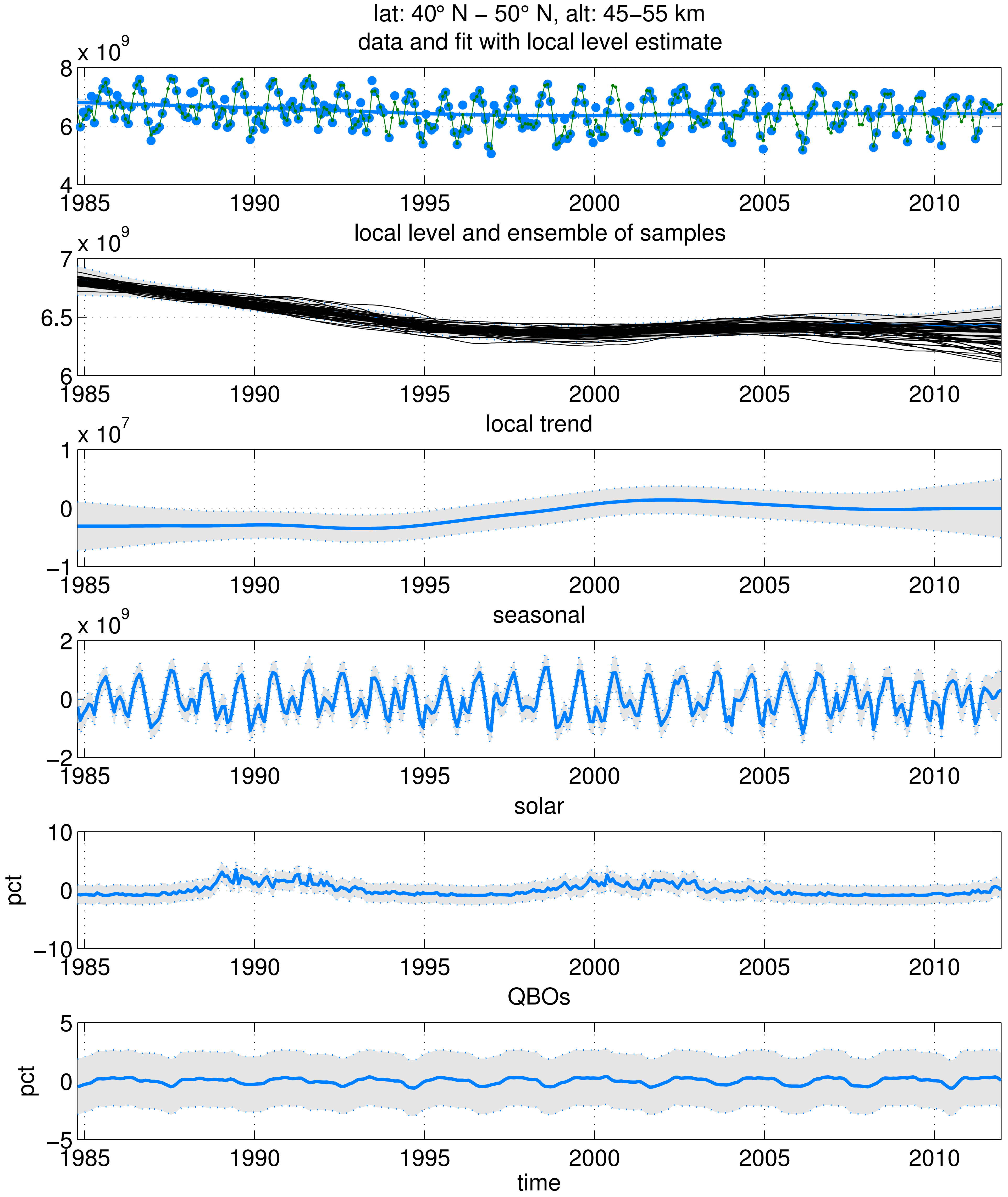




Stratospheric ozone from satellite observations

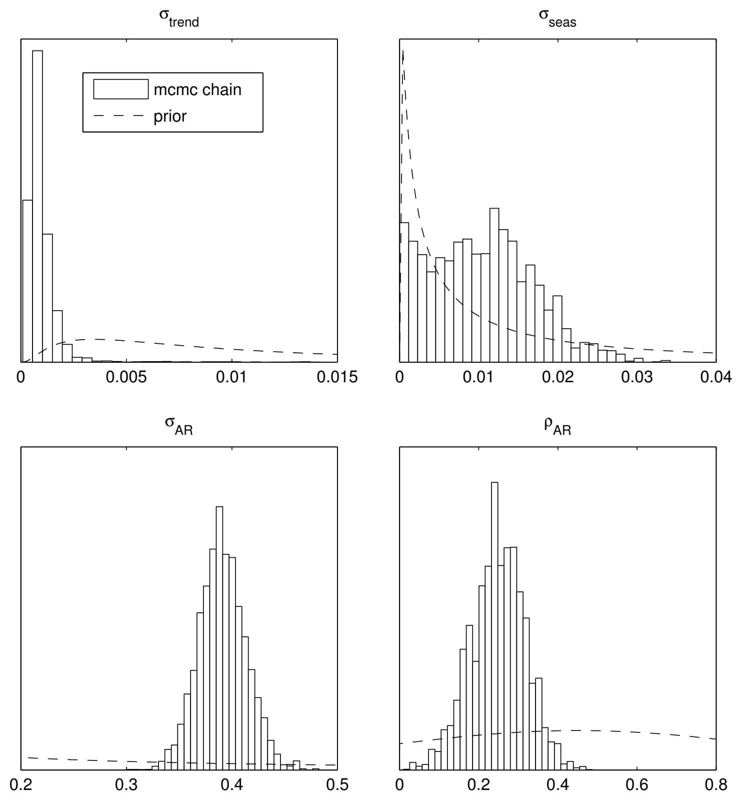
Merged monthly SAGE II - GOMOS observations for one latitude band and altitude region.

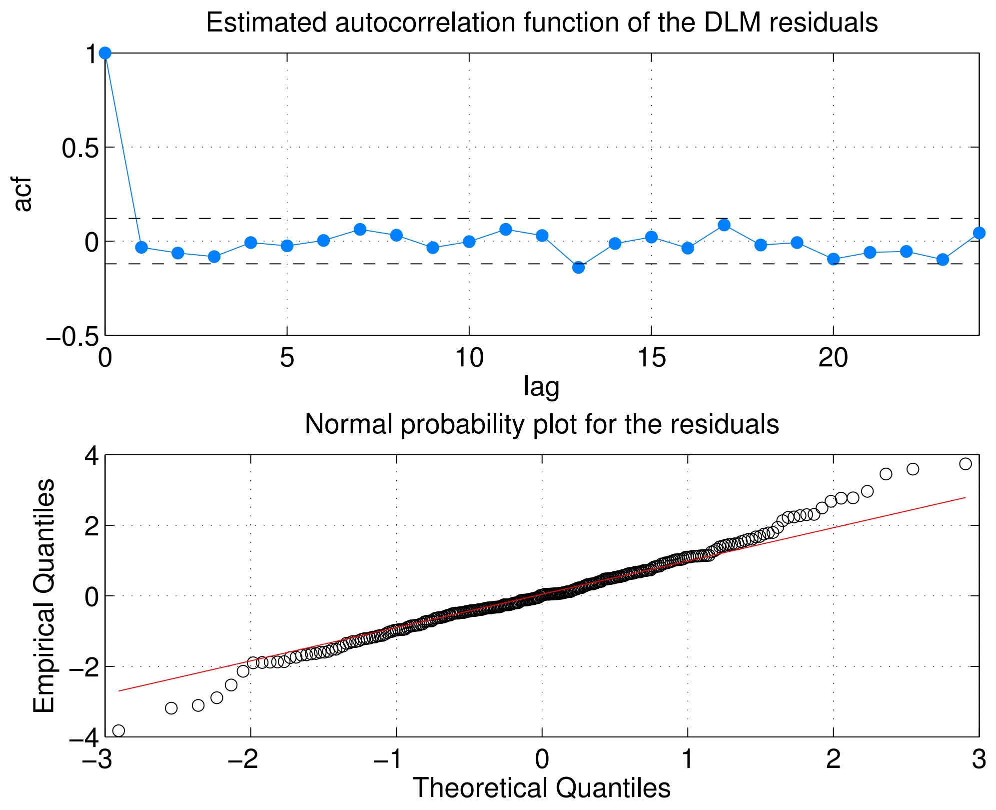




Time series components and their uncertainties by DLM analysis.

Parameter estimation by MCMC. Posterior distributions and residual analysis.





M. Laine et al.: Analysing time-varying trends in stratospheric ozone time series using the state space approach, ACP 14, 2014

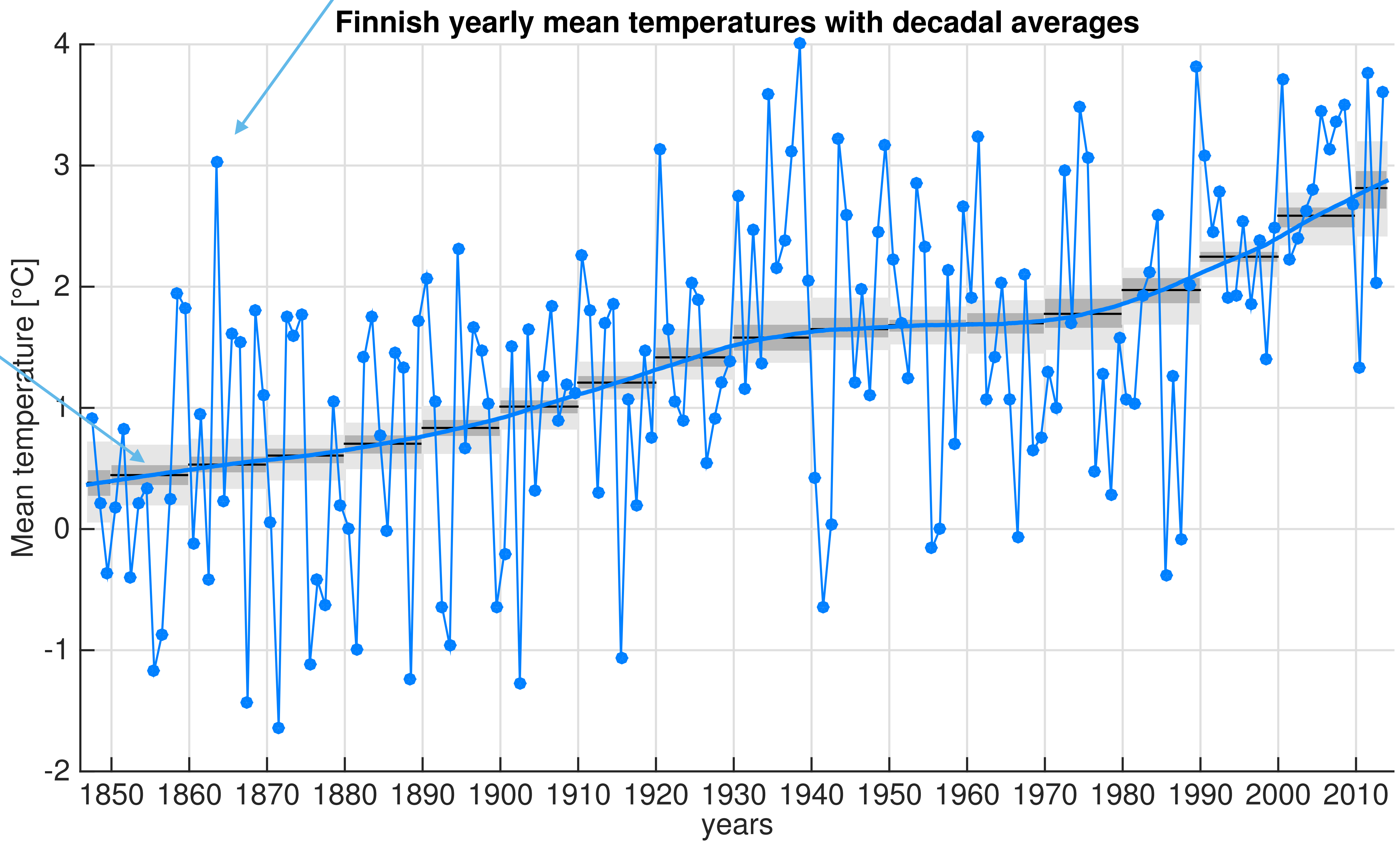
Finnish station temperatures 1847 - 2013

- Local trend
- Seasonality
- AR(1) error

The data are monthly means, here we show yearly averages, only.

Temperature raise 2,3°C (±0,4) 1850-2010.

Estimated decadal means



Mikkonen, Laine, et al., 2015

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

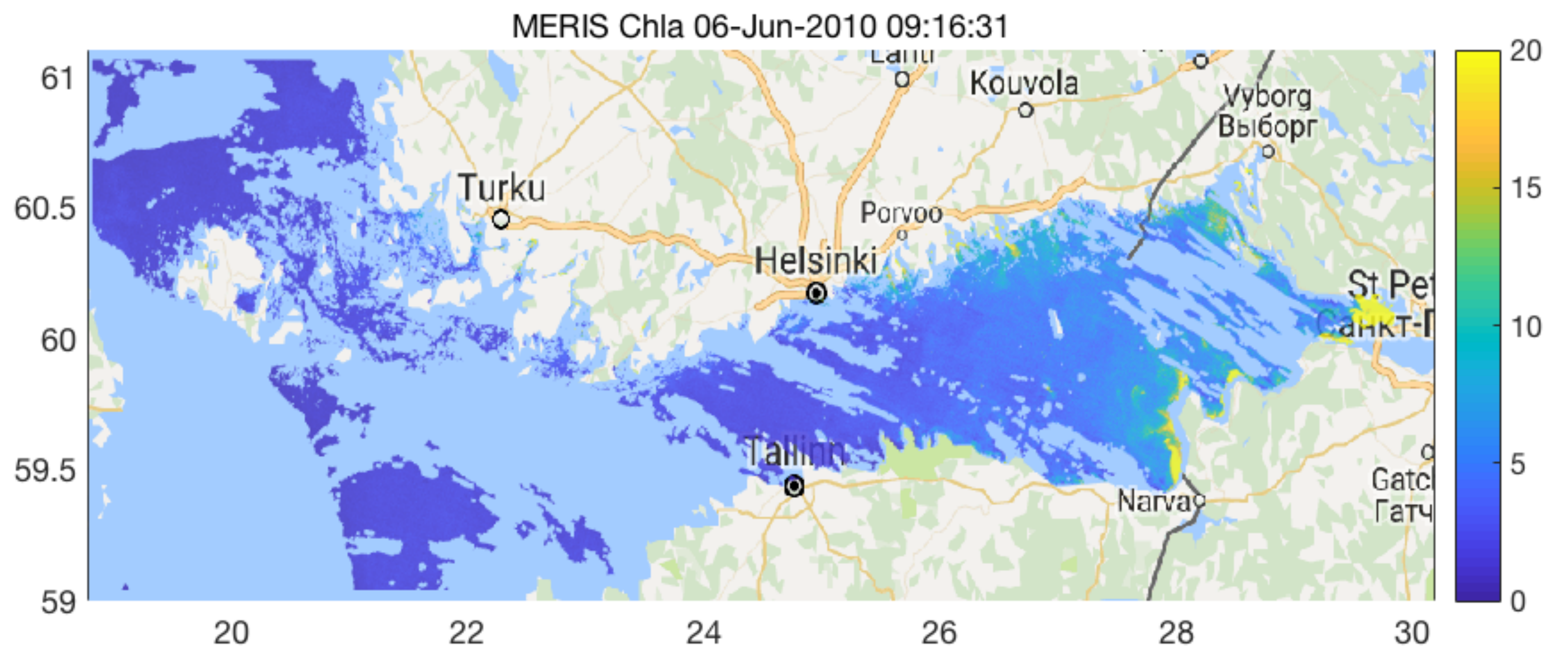
Data fusion as multivariate time series analysis by DLM

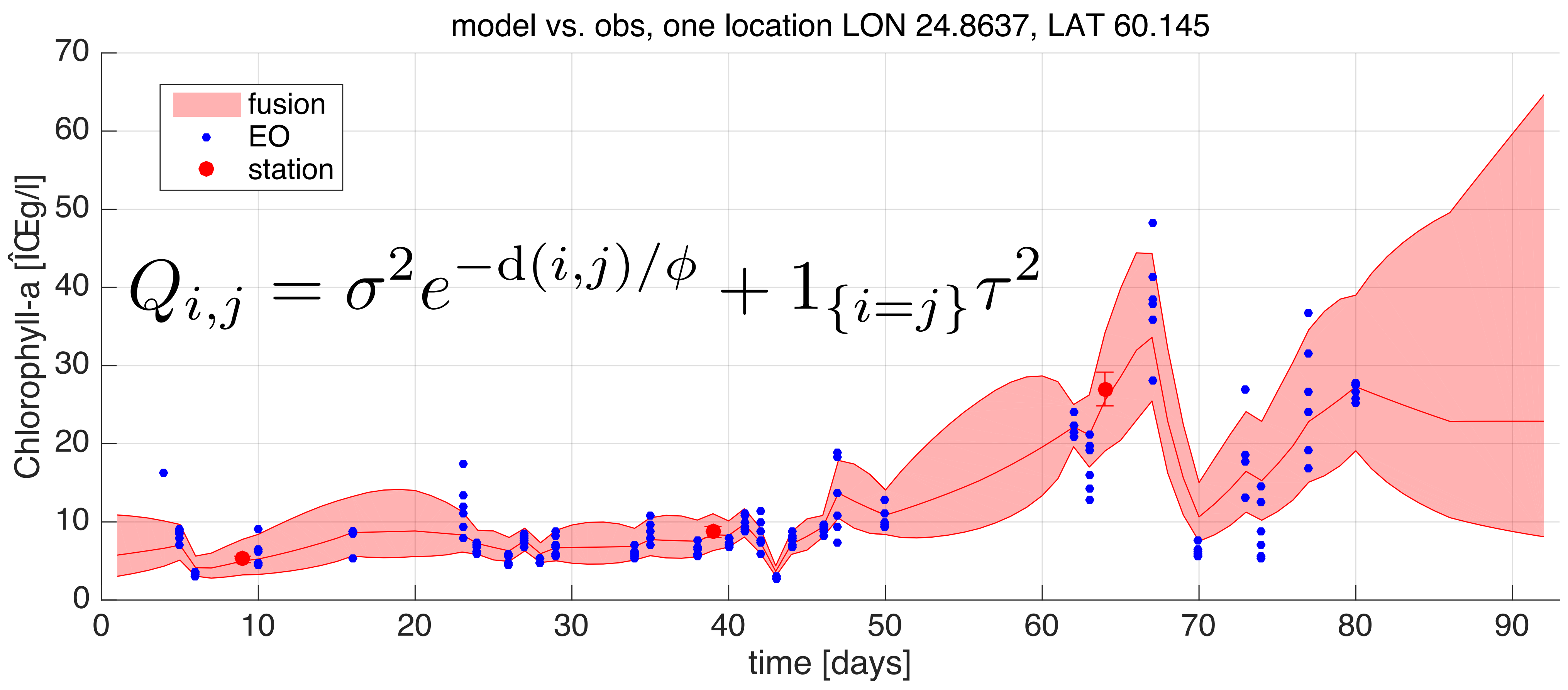
$y_t = H_t x_t + \varepsilon_t$ $\varepsilon_t \sim N(0,R_t)$

$x_t = M_t x_{t-1} + E_t$ $E_t \sim N(0,Q_t)$

- x_t is 2-3D regular grid of the modelled variable.
- M_t can be a trivial random walk model.
- Q_t is the assumed background spatial covariance structure.
- H_t maps model grid to observation locations.

- Data fusion of MERIS/ENVISAT satellite data and in-situ observations of Chlorophyll-a in Gulf of Finland.





*Computational tools

For dynamic **linear** models we have efficient computational tools for all the relevant statistical distributions in the hierarchical model.

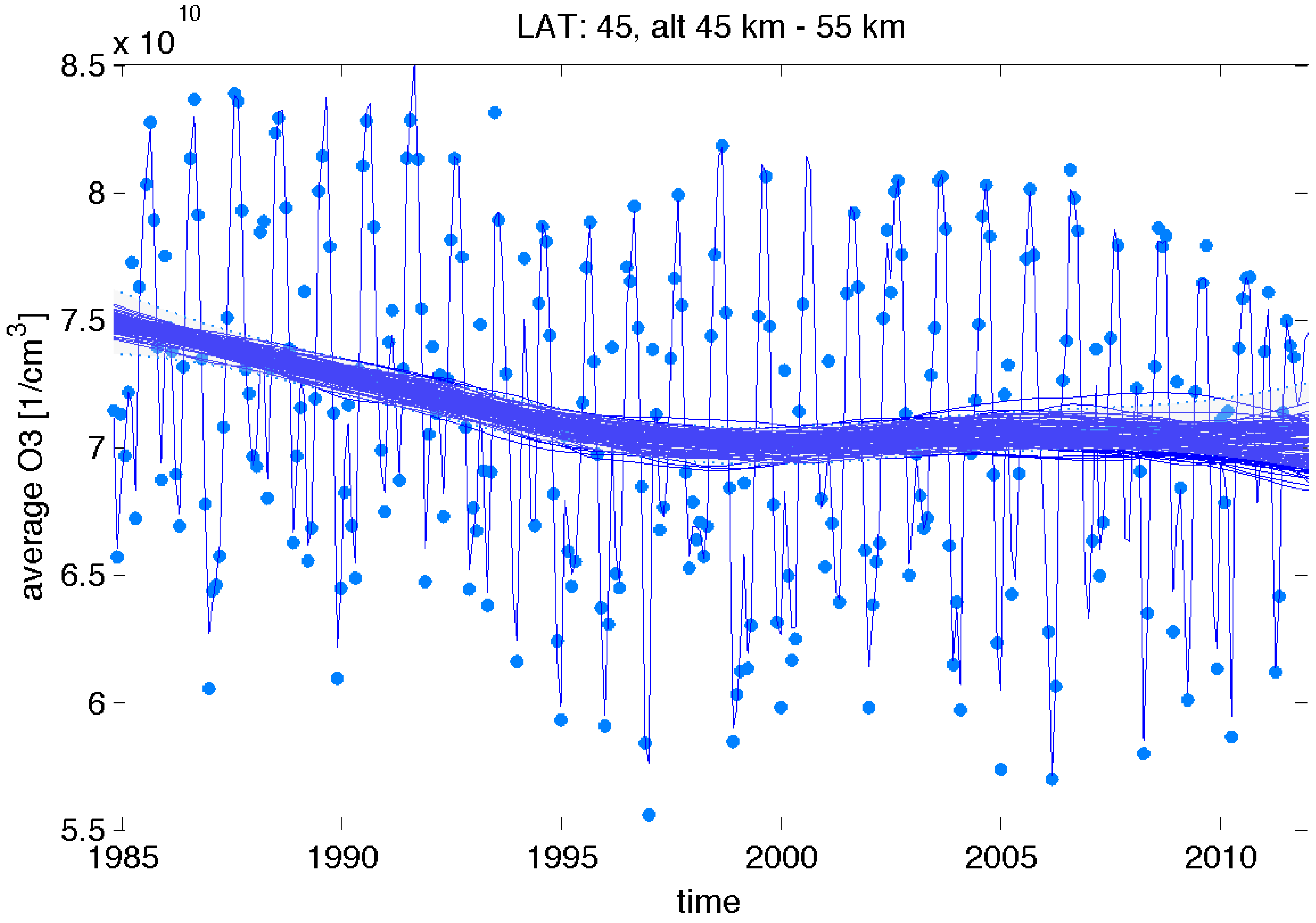
Distribution	method
$p(x_t \mid y_{1:t}, \theta)$	Kalman filter
$p(x_t \mid y_{1:n}, \theta)$	Kalman smoother
$p(x_{1:n} \mid y_{1:n}, \theta)$	Simulation smoother/RTO
$p(y_{1:n} \mid \theta)$	Kalman filter likelihood
$p(x_{1:n}, \theta \mid y_{1:n})$	MCMC
$p(x_{1:n} \mid y_{1:n})$	MCMC
$p(\mathrm{trend}(x_{1:n}) \mid y_{1:n})$	MCMC

However, for large state x_t some approximative methods or dimension reduction is needed.

*DLM with MCMC, full sampling for trend statistics

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

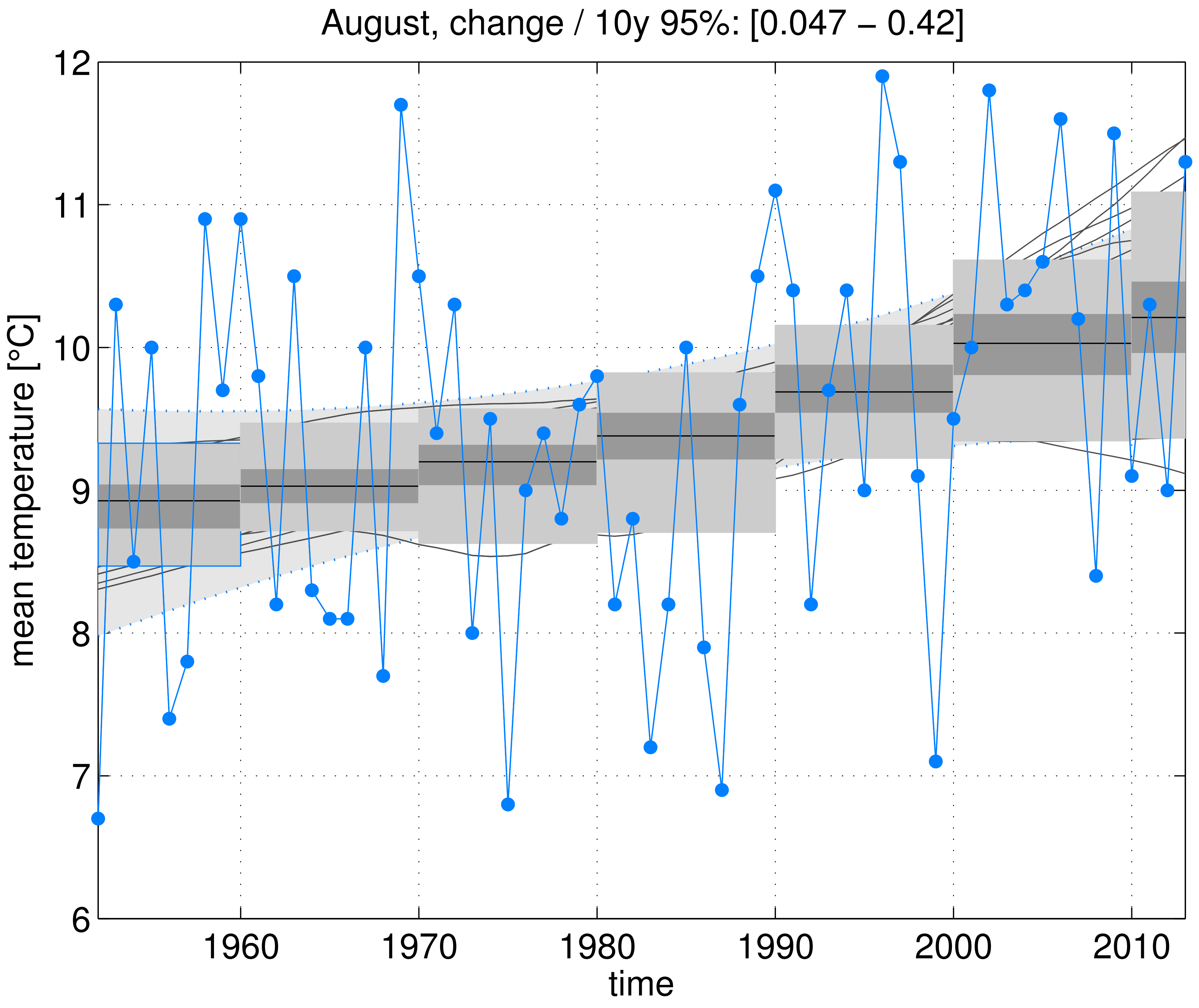
$$p(x_{1:n}|y_{1:n}) = \int p(x_{1:n}|y_{1:n}, \theta)\, d\theta.$$



- We get uncertainty distribution for trend related statistics.

Kilpisjärvi (69°2’54”N, 20°47’42”E) temperatures

- Monthly mean temperatures in August at Kilpisjärvi.
- Fitted DLM model.
- Sample from the background level.
- Estimated decadal averages.

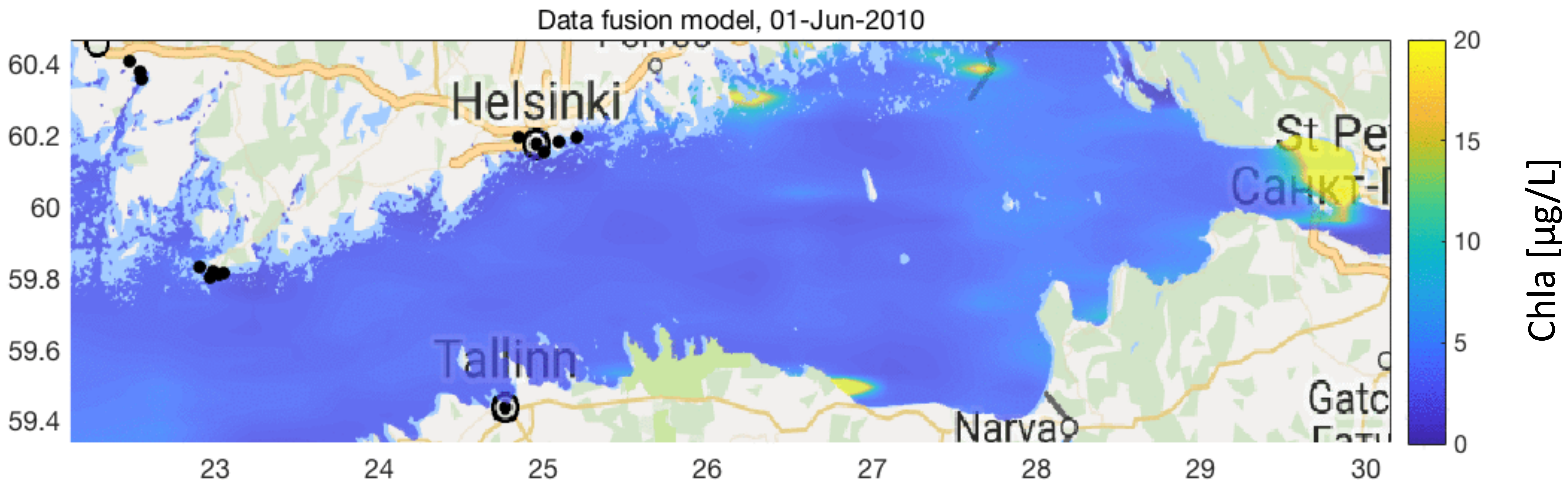


*Data fusion with dimension reduction

- Combine data from different sources to a common regular spatio-temporal grid.
- Reduced dimension smoother and a multivariate DLM time series model.
- Needs sparse model error precision matrix Q^{-1}.
- Needs basis P to form reduced state $x_t = \mu_t + P\alpha_t$ and covariance $C_t = PC^{\alpha}P^{T}$.
- Hierarchical models for hyper parameters in Q and P possible.
- Non-linear models by EKF and EnKF.

Data fusion of Chla in Baltic Sea

- Chlorophyll-a from satellite (Meris/ENVISAT, later Sentinel-2) with in-situ observation from stations and commercial vessels.
- EO data in 3774x674 (0.003° lat-lon) resolution, state dimension ~2.5 10^6. Using 30 principle components (empirical orthogonal functions) to describe the state.
- P is 2 543 676 x 30, C^{α} is 30 x 30.



Thank You!

- S. Tukiainen, J. Railo, M. Laine, et al.: Retrieval of atmospheric CH_4 profiles from TCCON FTS data using dimension reduction and MCMC, *Journal of Geophysical Research*, 2016.
- A. Solonen, T. Cui, J. Hakkarainen, Y. Marzouk. On dimension reduction in gaussian filters. *Inverse Problems*, 32, 2016.
- J. M. Bardsley, A. Solonen, H. Haario, M. Laine: Randomize-then-optimize: a method for sampling from posterior distributions in nonlinear inverse problems, *SIAM Journal on Scientific Computing*, 36, 2014.
- J. Hakkarainen, et al. On closure parameter estimation in chaotic systems. *Nonlin. Proc. in Geoph.*, 19, 2012.
- T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, A. Spantini: Likelihood-informed dimension reduction for nonlinear inverse problems, *Inverse Problems*, 30, 2014.
- A. Bibov, H. Haario, and A. Solonen. Stabilized BFGS approximate Kalman filter. Inverse Problems and Imaging, 9, 2015.
- S. Mikkonen, M. Laine, et al.: Trends in the average temperature in Finland,1847-2013, *Stoch. Environ. Res. Risk Assess.*, 29, 2015.
- M. Laine, N. Latva-Pukkila, E. Kyrölä: Analysing time-varying trends in stratospheric ozone time series using the state space approach, *Atmos. Chem. Phys.*, 14, 2014.

- Matlab toolbox for MCMC UQ calculations for nonlinear models at http://helios.fmi.fi/~lainema/mcmc.
- Matlab toolbox for DLM calculations for time series at http://helios.fmi.fi/~lainema/dlm.