NRCSE

Misalignment and use of deterministic models

Work with

Veronica Berrocal
Peter Craigmile
Wendy Meiring
Paul Sampson
Gregory Nikulin

The choice of spatial scale– some questions

1. Which spatial scale is correct?
2. What if there is *spatial misalignment*?
3. How do we change from one spatial scale to another?
4. What if we have different spatial datasets that come to us on different spatial scales?
5. How do we combine data sources?

We need to be careful not to be misled in our inferences.

Changes of support

Observed	Inference	Method
point	point	kriging
point	line	contouring
point	area	block kriging
area	point	ecological inference
area	area	misalignment

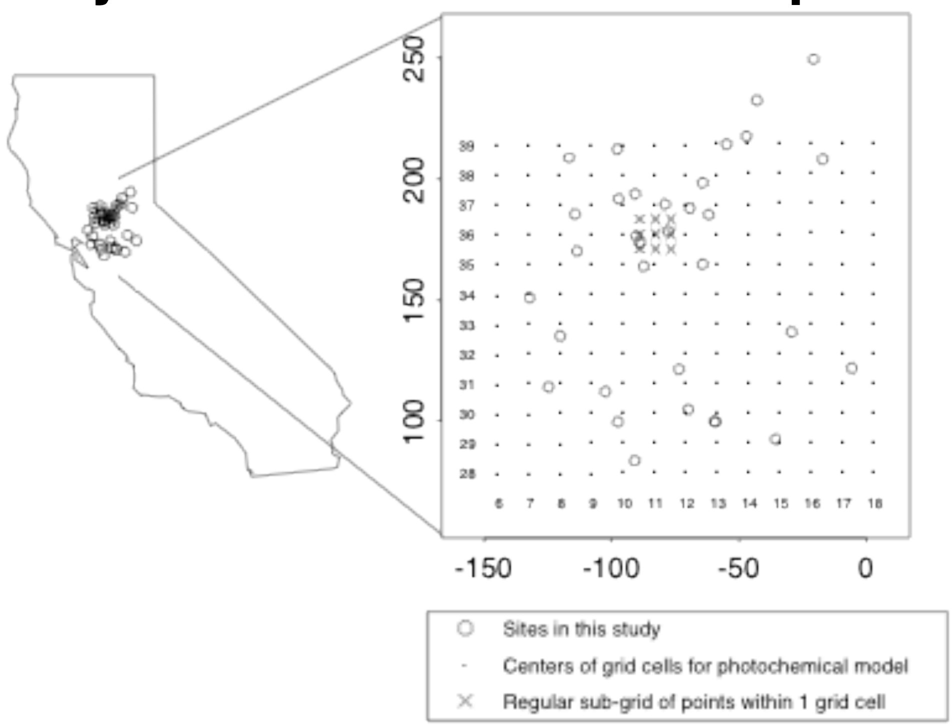
Some issues in model assessment

- Spatiotemporal misalignment
  - Grid boxes vs observations
- Types of error
  - Measurement error and bias
  - Model error
  - Approximation error
- Manipulate data or model output?

Assessing the SARMAP model

60 days of hourly observations at 32 sites in Sacramento region

Hourly model runs for three "episodes"



Task

Estimate from data the ozone level in a grid square.

Issues:

- Transformation
- Diurnal cycle
- Temporal dependence
- Spatial dependence
- Space-time interaction

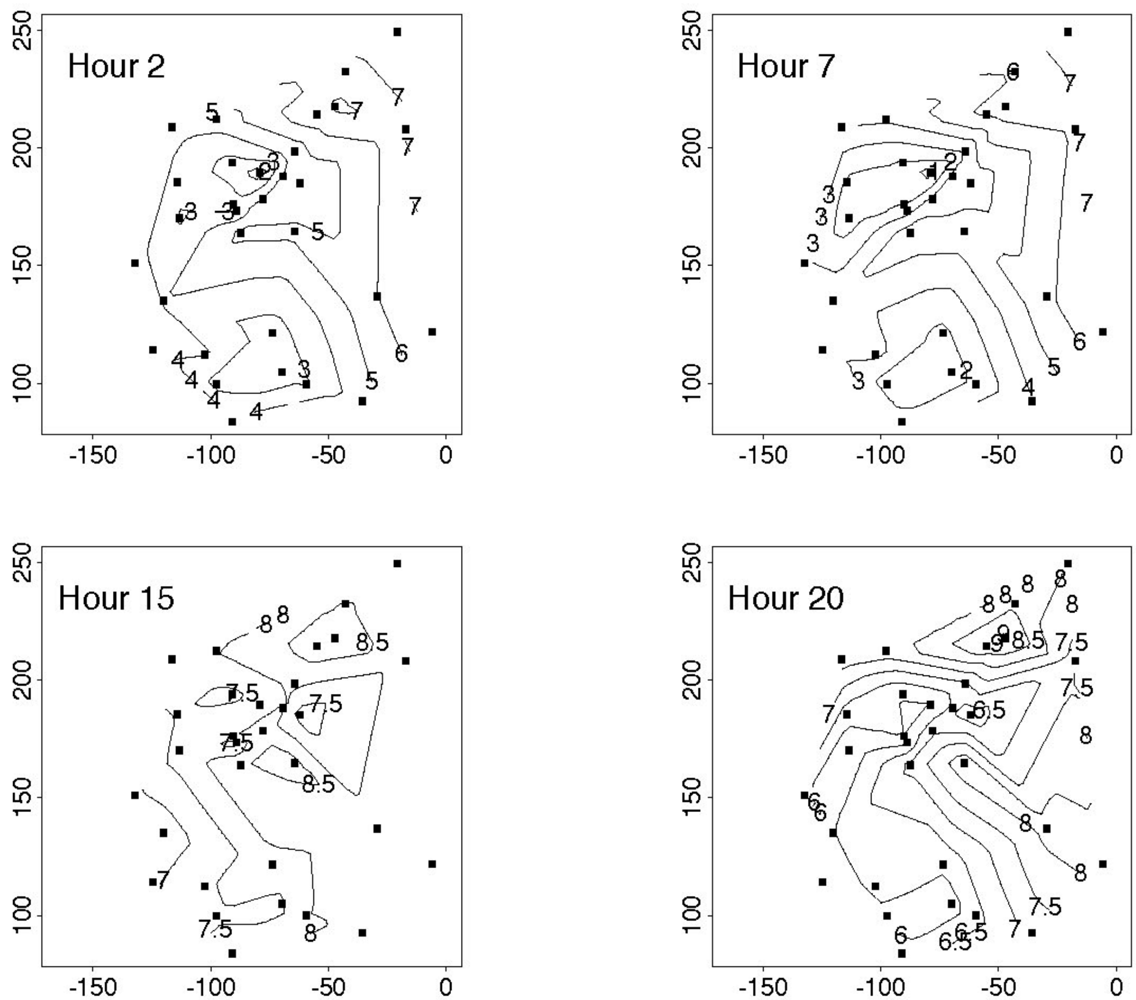
Transformation

Heterogeneous variability–mean and variance positively related

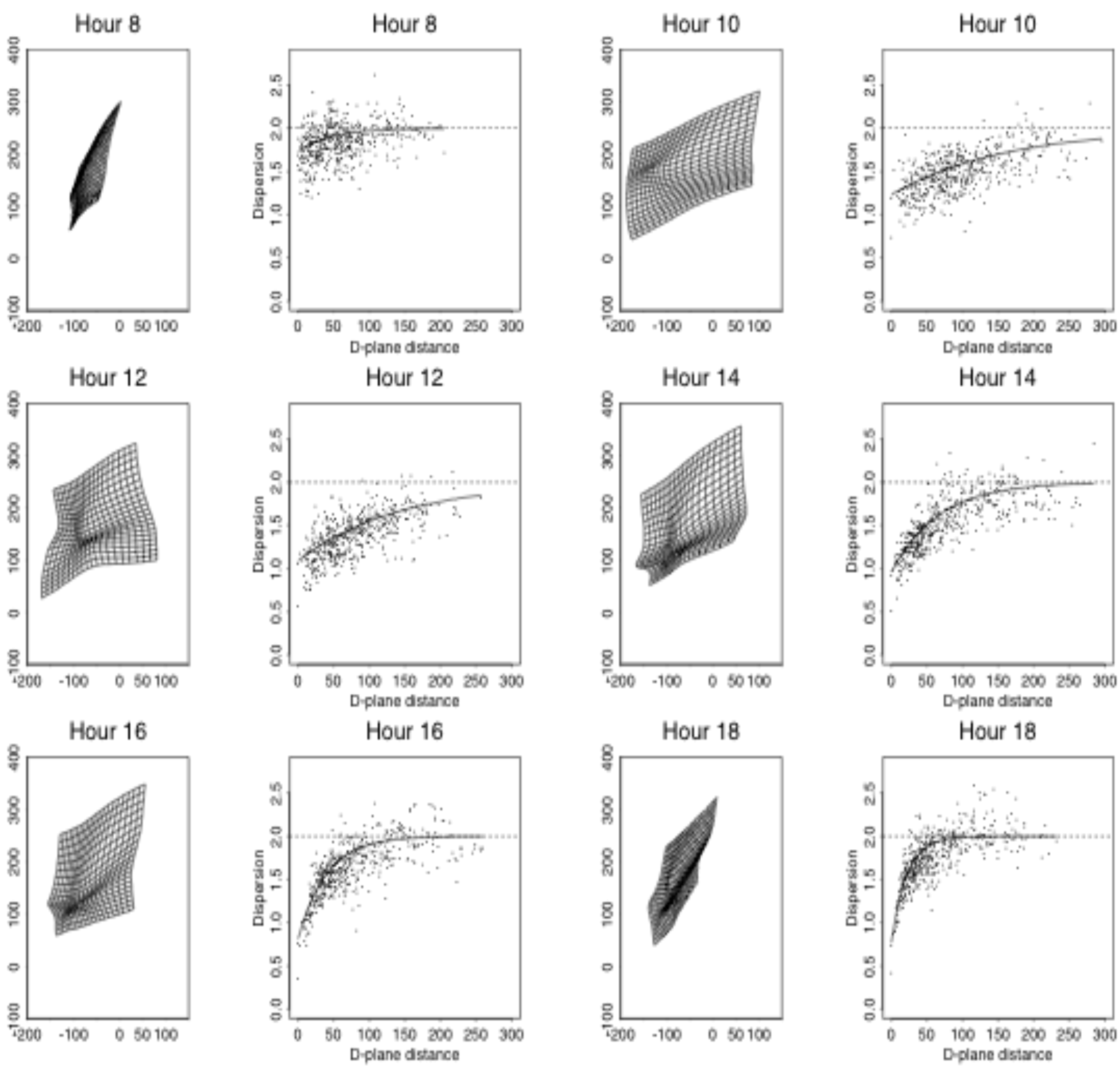
Square root transformation

All modeling now on square root scale–approximately normal

Diurnal cycle



Spatial dependence



Estimating a grid square average

$$V_t(s) = \sqrt{Z_t(s)}$$

$$V_t(s) = \mu_t(s) + W_t(s)$$

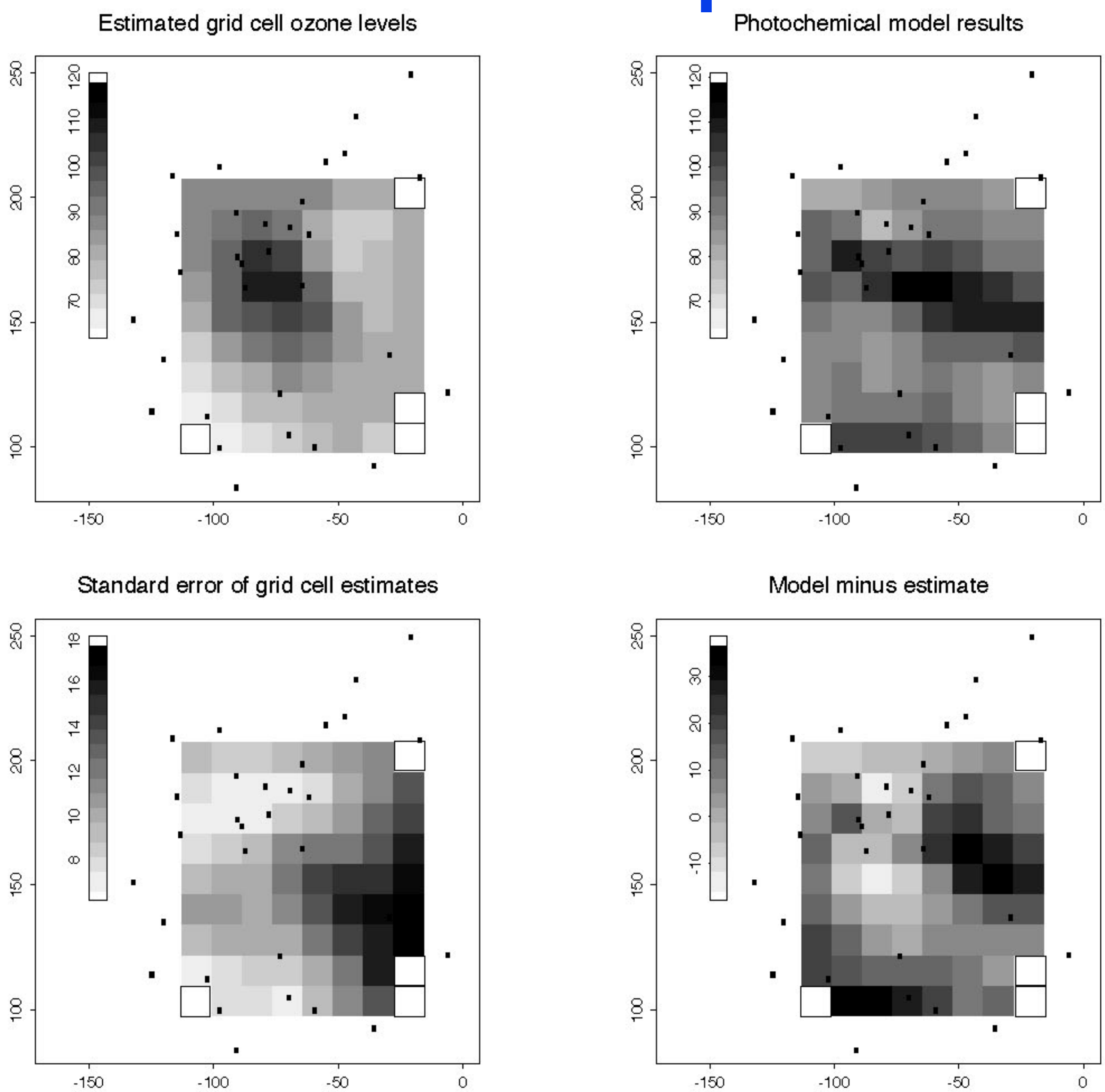
$$W_t(s) = \alpha_1(s)W_{t-1}(s) + \alpha_2(s)W_{t-2}(s) + Y_t(s)$$

Estimate $\frac{1}{|A|}\int_A V_t^2(s)ds$ using

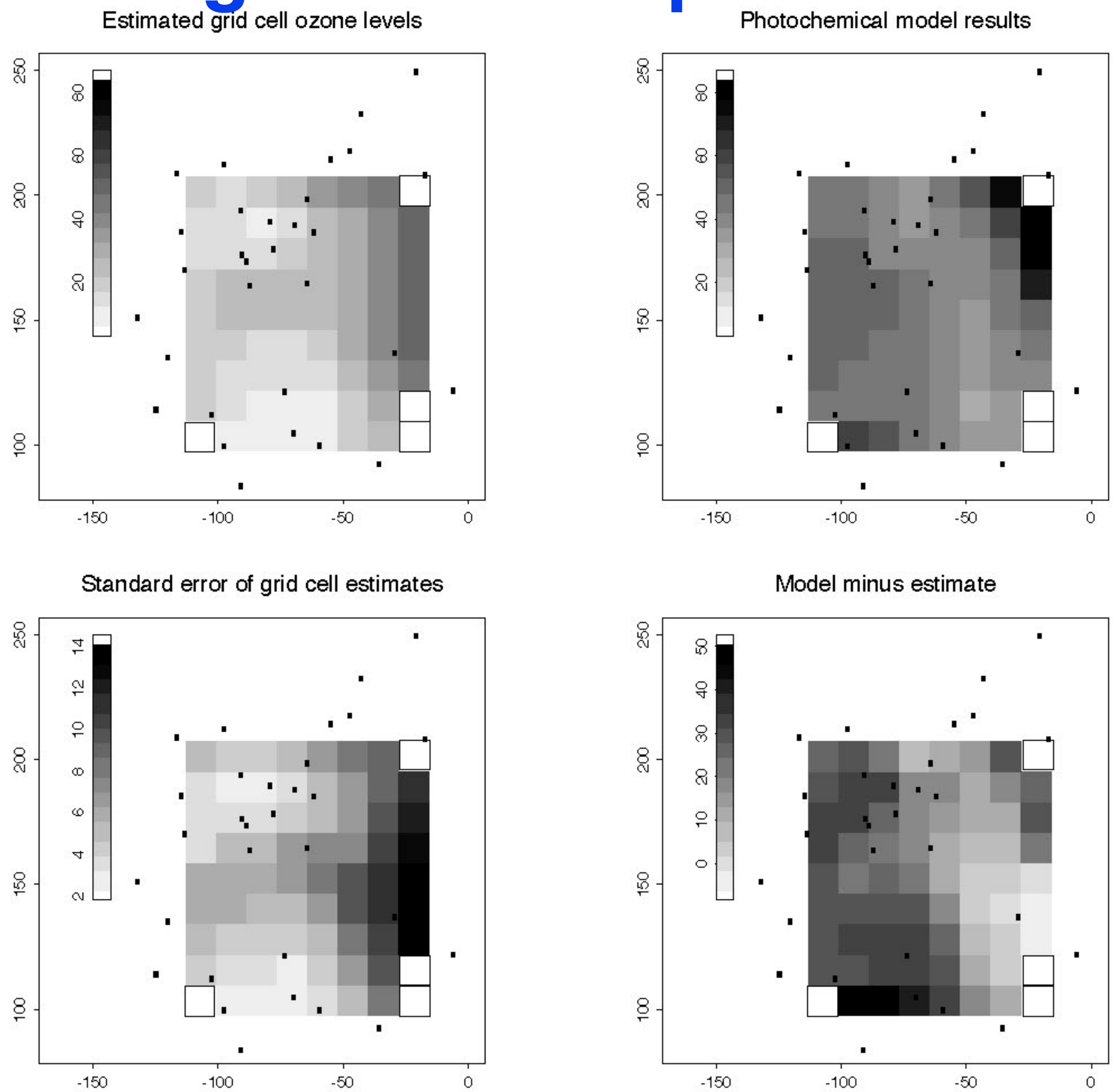
$$\frac{1}{M}\sum E\left\{V_t(s_j)^2 \,\middle|\, \text{data from } 1,\ldots,t\right\}$$

(*not* averages of squares of kriging estimates on the square root scale)

Afternoon comparison



Nighttime comparison



Regional climate models

- Not possible to do long runs of global models at fine resolution
- Regional models (dynamic downscaling) use global model as boundary conditions and runs on finer resolution
- Output is averaged over land use classes
- “Weather prediction mode” uses reanalysis as boundary conditions

Comparison of model to data

Model output daily averaged 3hr predictions on $(12.5\ km)^2$ grid

Use open air predictions only

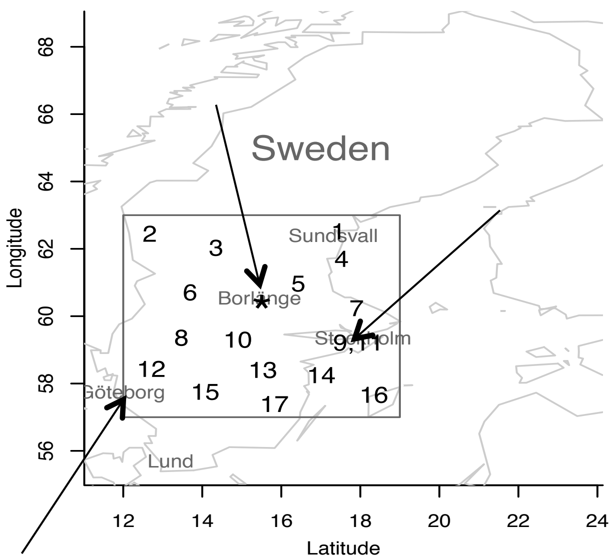
RCA3 driven by ERA 40/ERA Interim

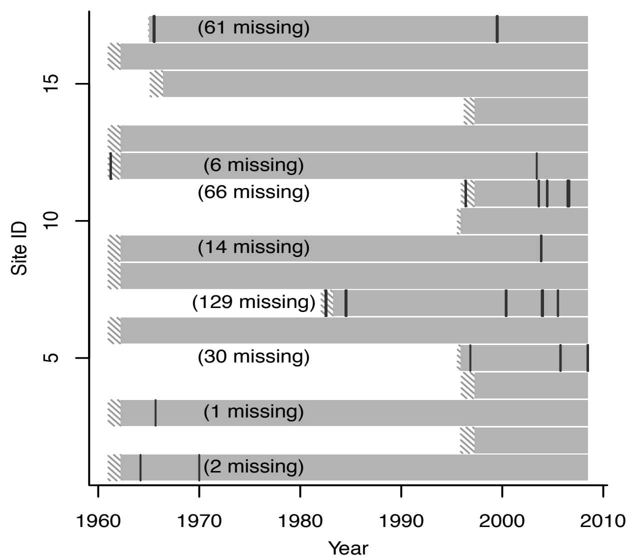
Data daily averages point measurements (actually weighted average of three hourly measurements, min and max)

Aggregate model and data to seasonal averages

Data

SMHI synoptic stations in south central Sweden, 1961-2008





Upscaling

Geostatistics: predicting grid square averages from data

Difficulties:

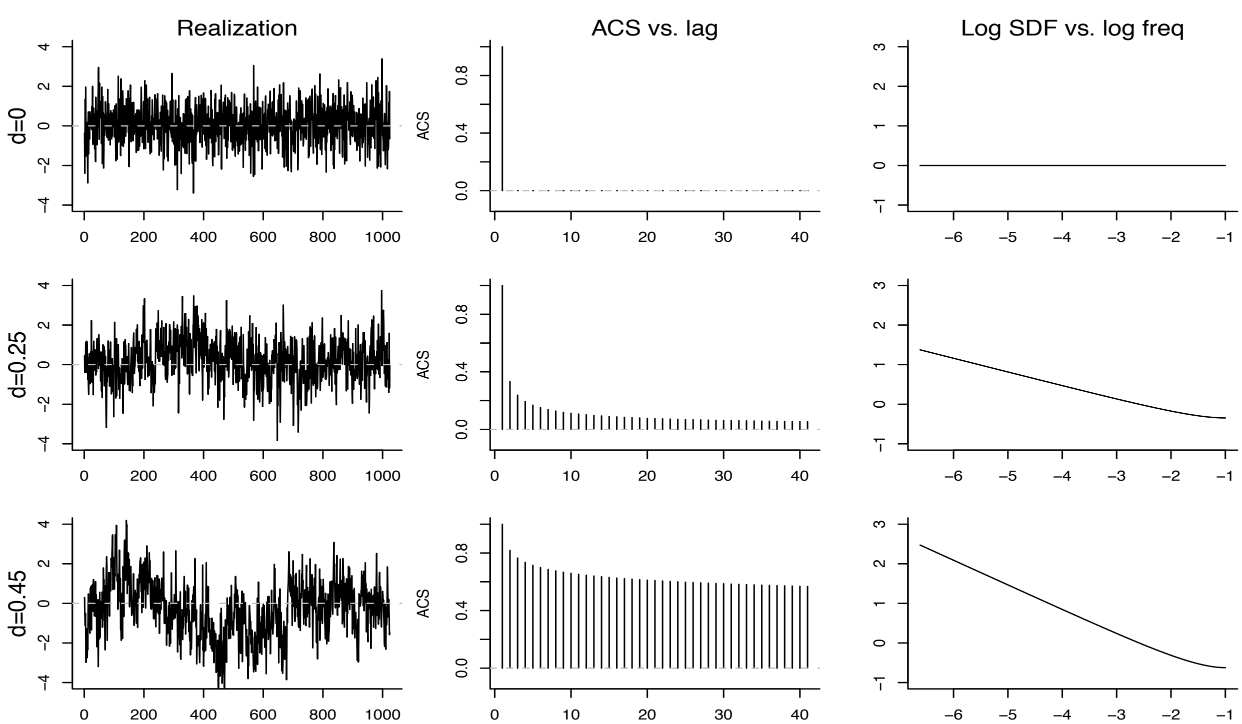
Trends

Seasonal variation

Long term memory features

Short term memory features

Long term memory models



A “simple” model

$$Y_t(s) = \mu_t(s) + \varphi_t(s) + \exp(\alpha_t(s))\eta_t(s)$$

space-time trend

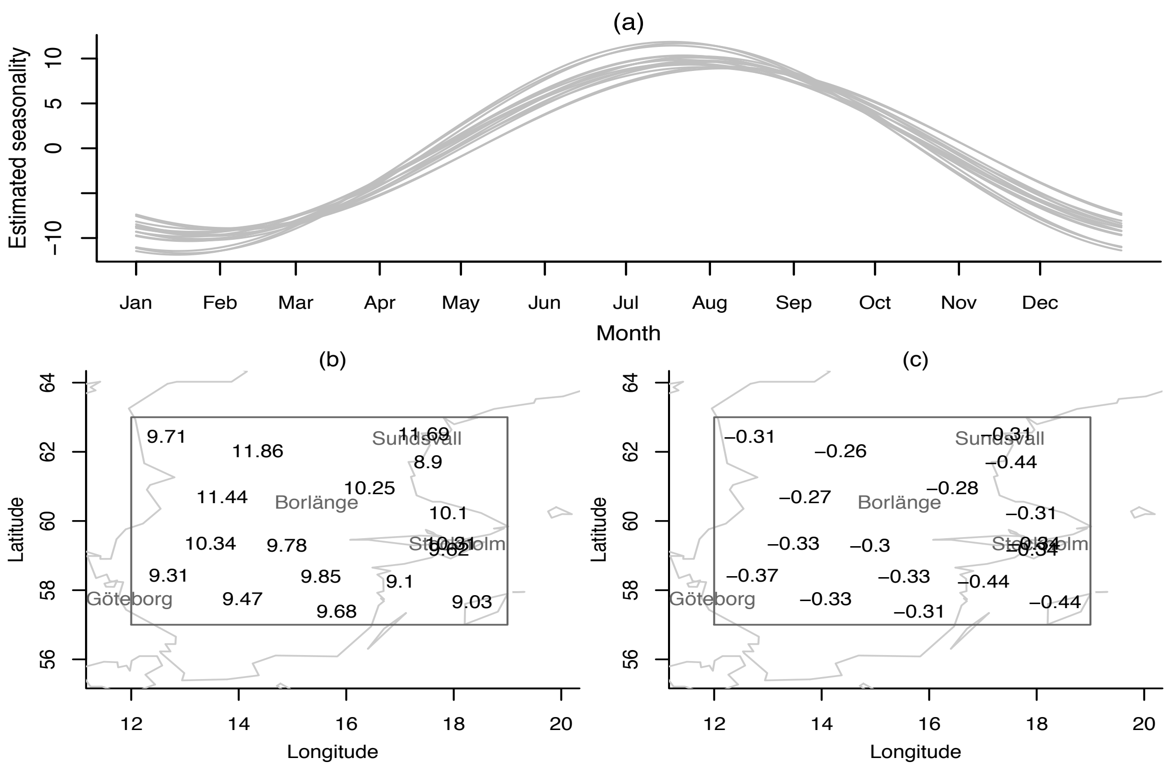
periodic seasonal component

seasonal variability

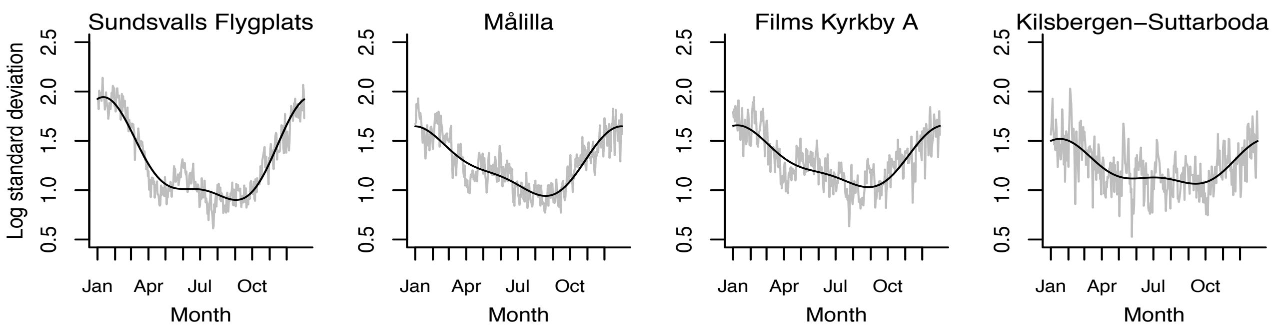
noise

Seasonal part

$$\phi_t(s) = A(s)\cos(2\pi t / 365.25 + \theta(s))$$



Seasonal variability



Modulate noise $\zeta_t(s) = \exp(\alpha_t(s))\eta_t(s)$

$\alpha_t(s)$ **two term Fourier series**

Both long and short memory

Consider a stationary Gaussian process with spectral density

$$S_\eta(f) = B(f)\left|4\sin^2(\pi f)\right|^{-\delta}$$

Short term memory → $B(f)$

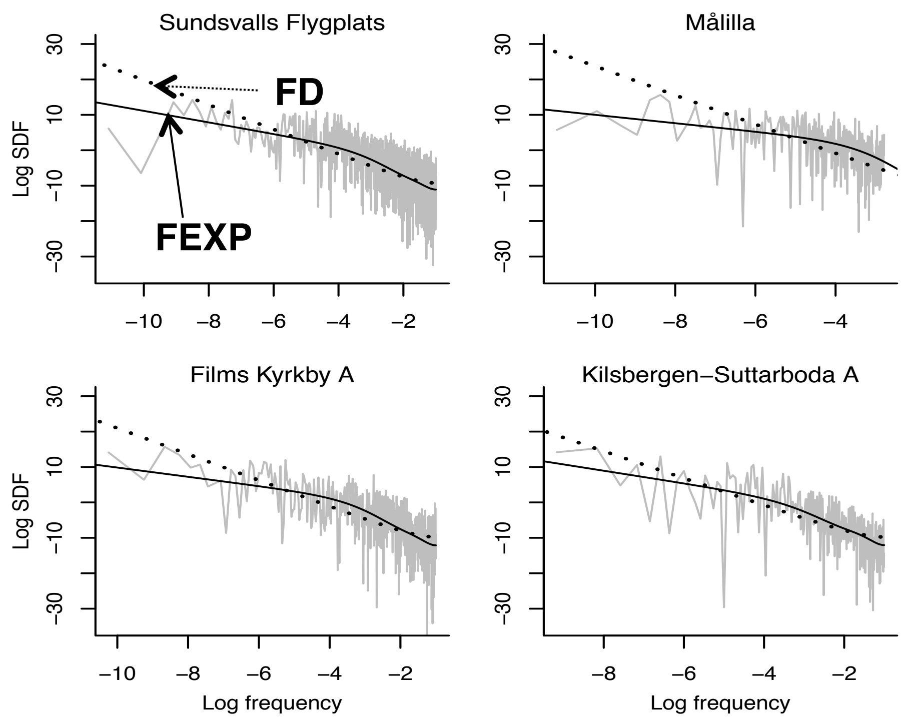
Long term memory → $\left|4\sin^2(\pi f)\right|^{-\delta}$

Examples:

B(f) constant: fractionally differenced process (FD)

B(f) exponential: fractional exponential process (FEXP) (log B truncated Fourier series)

Estimated SDFs of standardized noise



Clear evidence of both short and long memory parts

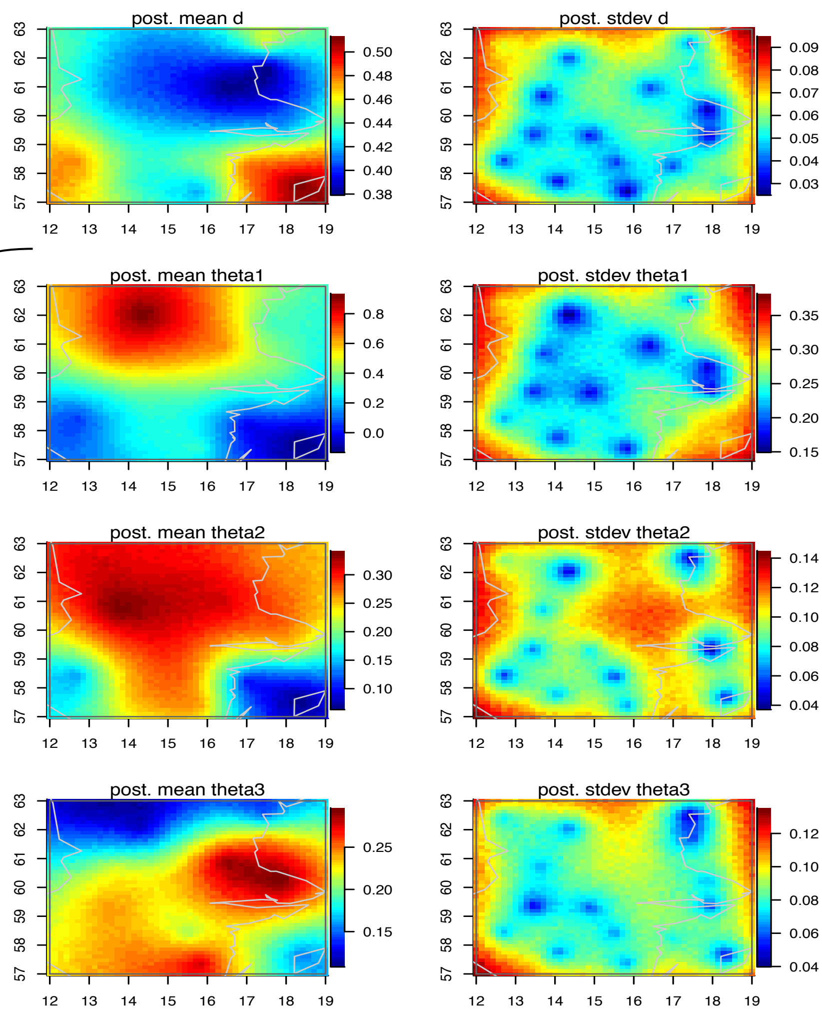
Space-time model

- Gaussian white measurement error
- Process model in wavelet space
  - scaling coefficients have mean linear in time and latitude
  - separable space-time covariance
- Gaussian spatially varying parameters

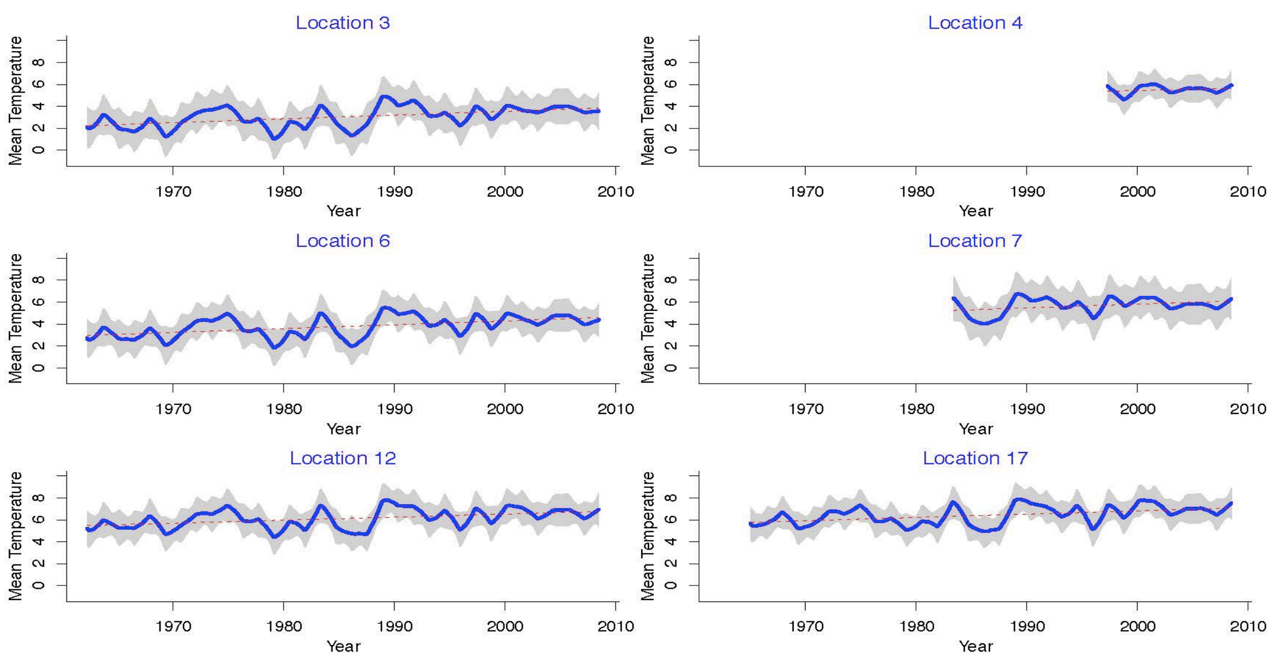
Dependence parameters

LTM

Short term



Trend estimates



Estimating grid squares

Pick q locations systematically in the grid square

Draw sample from posterior distribution of Y(s,t) for s in the locations and t in the season

Compute seasonal average

Compute grid square average

Downscaling

Climatology terms:
Dynamic downscaling
Stochastic downscaling
Statistical downscaling

Here we are using the term to allow
- **data assimilation for RCM**
- **point prediction using RCM**

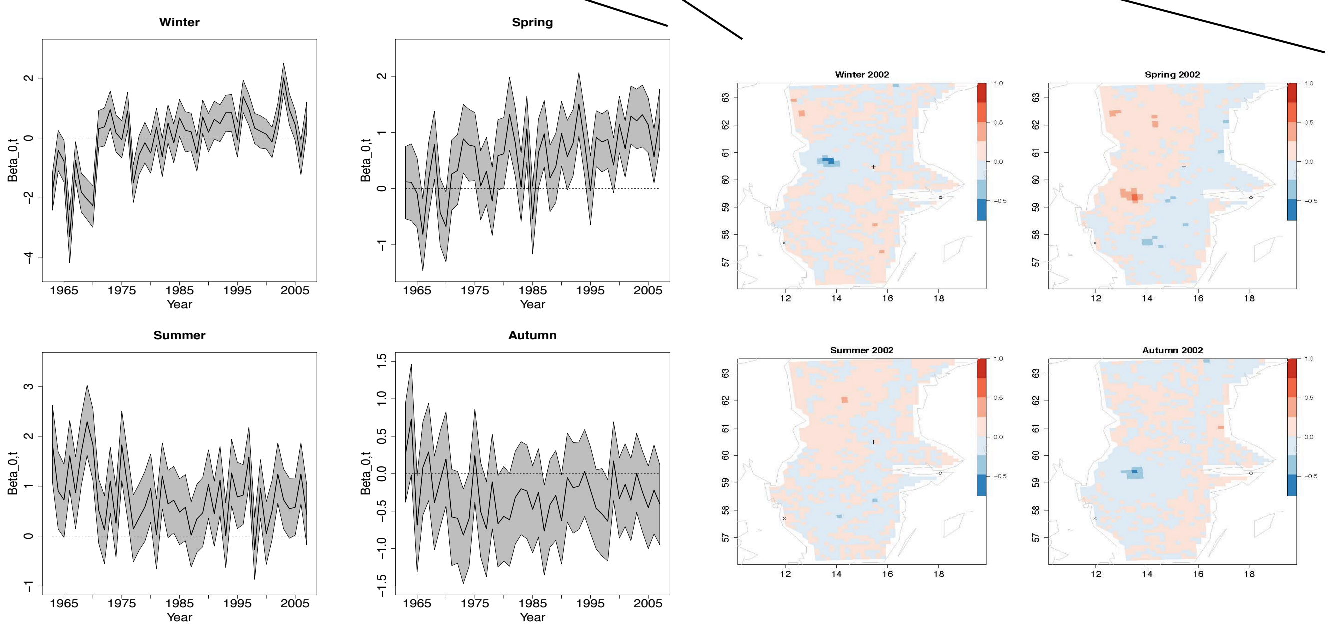
Downscaling model

(0.91,0.95)

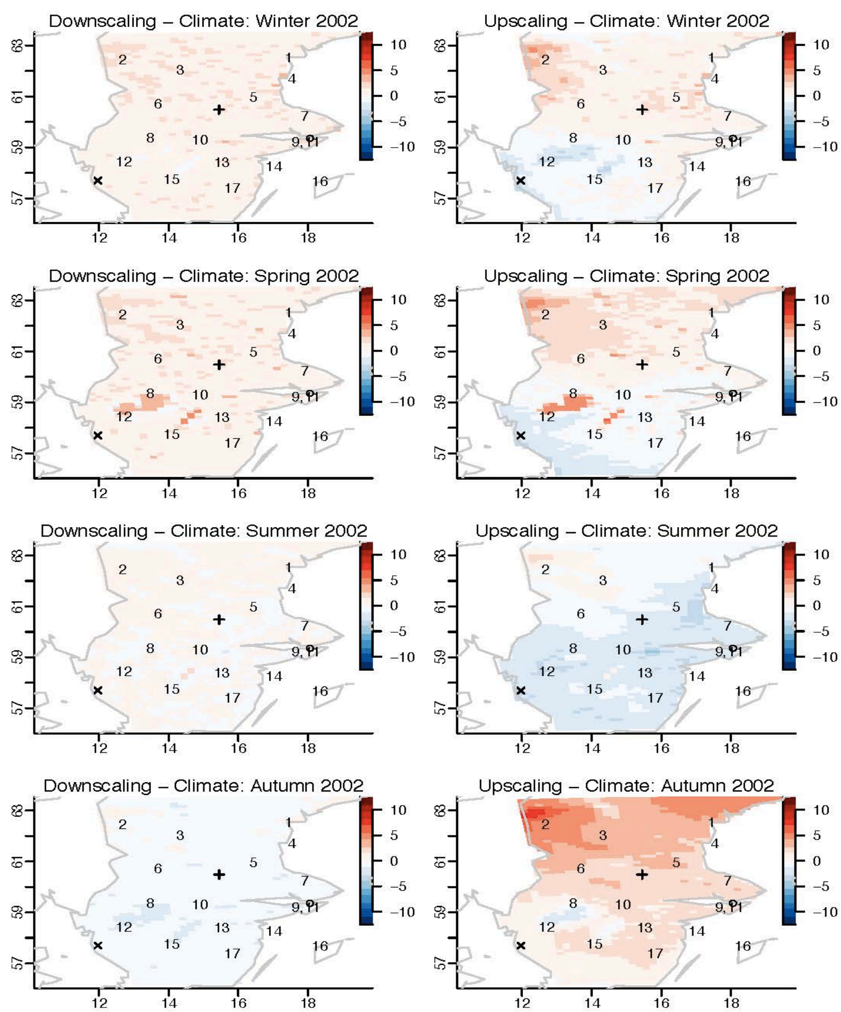
$$Y(s,t) = \tilde{\beta}_0(s,t) + \beta_1 \tilde{x}(s,t) + \varepsilon(s,t)$$

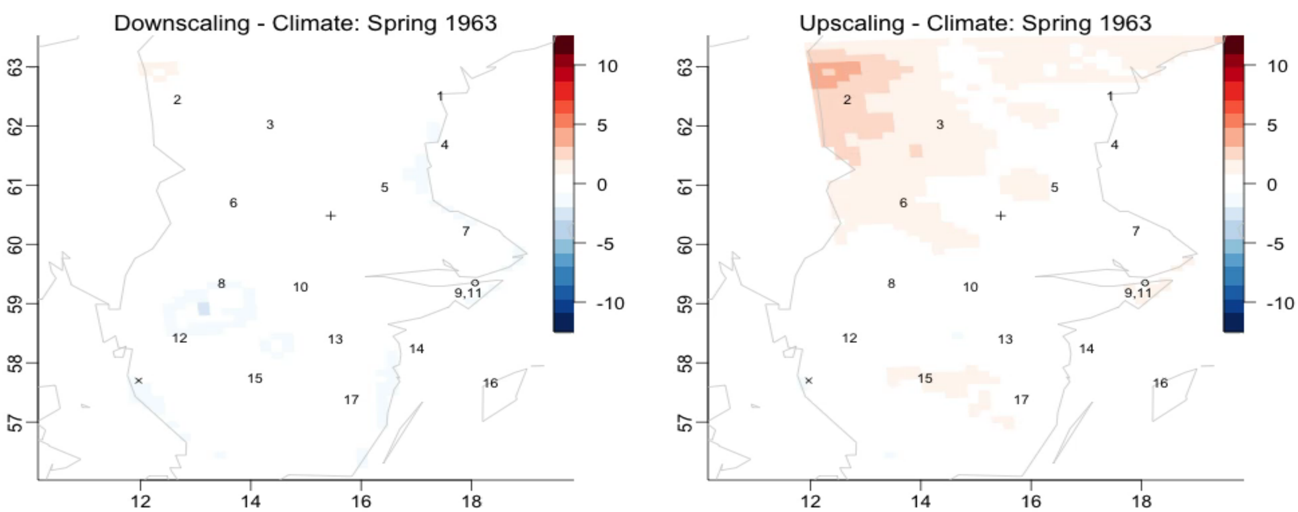
$$\tilde{\beta}_0(s,t) = \beta_0(t) + \beta(s,t)$$

smoothed RCM



Comparisons




Downscaling - Climate: Spring 1963
Upscaling - Climate: Spring 1963
63
62
61
60
59
58
57
12
14
16
18
10
5
0
-5
-10
1
2
3
4
5
6
7
8
9,11
10
12
13
14
15
16
17

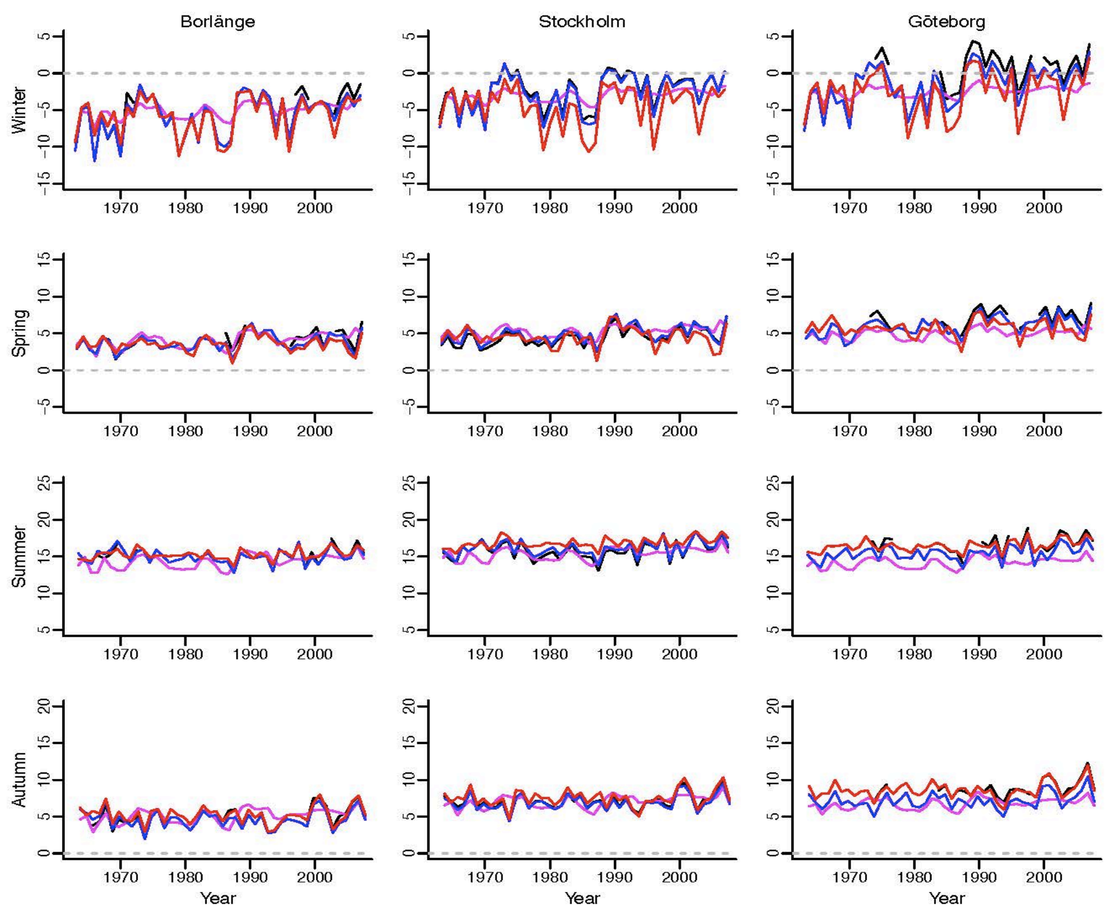
Reserved stations

Borlänge: Airport that has changed ownership, lots of missing data

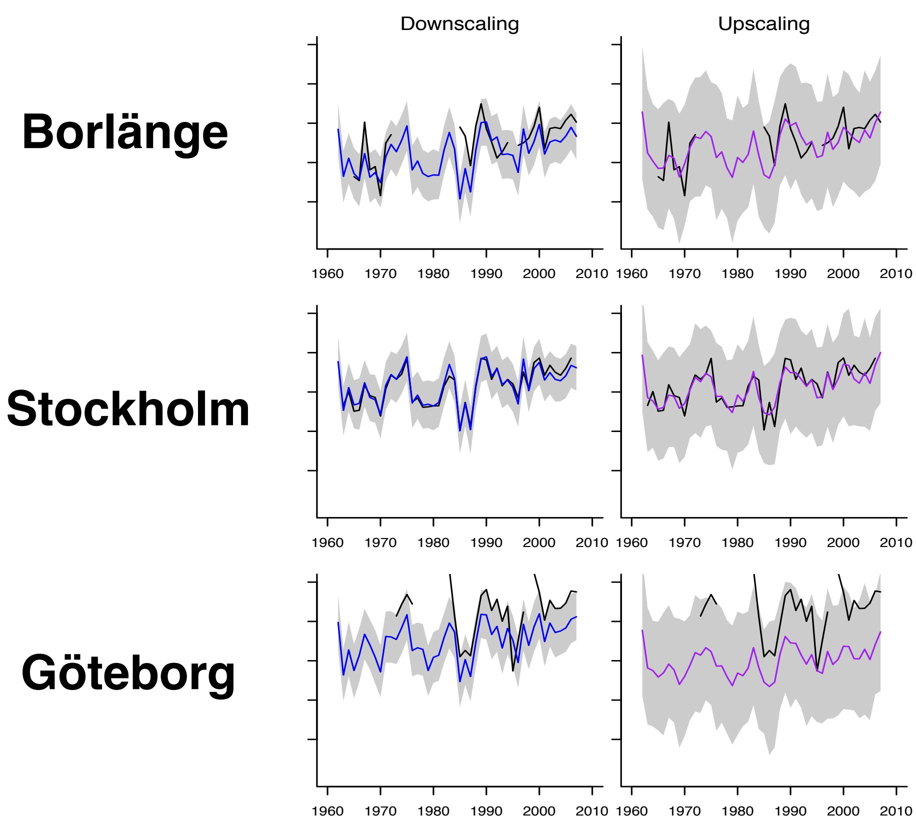
Stockholm: One of the longest temperature series in the world. Located in urban park.

Göteborg: Urban site, located just outside the grid of model output

Predictions and data



Annual scale



Comments

Nonstationarity

in mean

in covariance

Uncertainty in model output

”Extreme seasons” where down-and upscaling agree with each other but not with the model output

Model correction approaches

[redacted]

Statistical analysis of computer code output

Often the process model is expensive to run (in time, at least), especially if different runs needed for MCMC

Need to develop real-time approximation to process model

Kalman filter is a dynamic linear model approximation

SACCO is an alternative Bayesian approach

Basic framework

An emulator is a random (Gaussian) process $\eta(x)$ approximating the process model for input x in R^m.

Prior mean $m(x) = h(x)^T\beta$

Prior covariance $v(x_1, x_2) = \sigma^2 c(x_1, x_2)$

Run the model at n input values to get n output values, so

$$(d \mid \beta, \sigma^2) \sim N(H\beta, \sigma^2 C)$$

$$(\eta(\bullet) \mid \beta, \sigma^2, d) \sim N(m^*, \Sigma^*)$$

The emulator

Integrating out β and σ^2 we get

$$\frac{\eta(x) - m^{**}(x)}{\hat{\sigma} c^{**}(x,x)^{1/2}} \sim t_{n-q}$$

where $q = \dim(\beta)$ and

$$m^{**}(x) = h(x)^T \hat{\beta} + t(x)^T C^{-1}(d - H\hat{\beta})$$

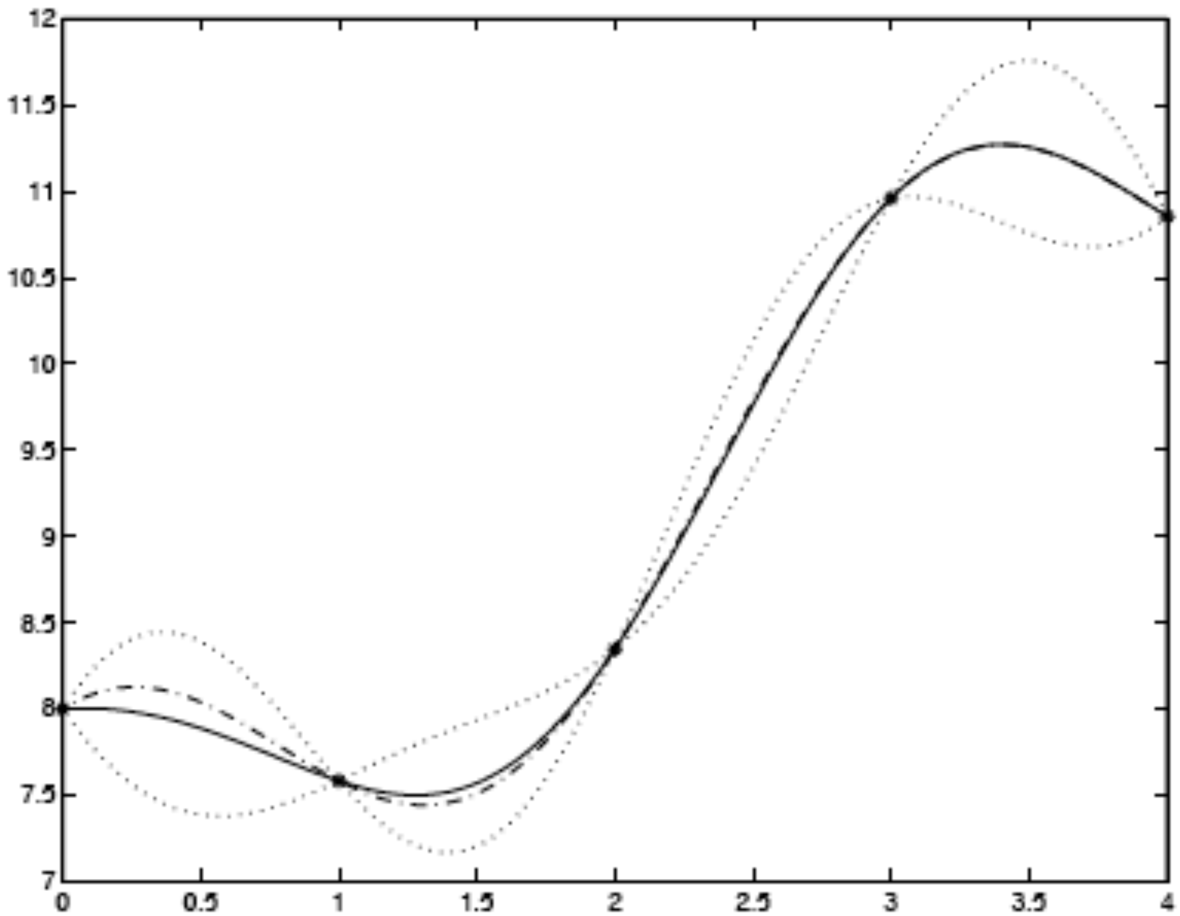
where $t(x)^T = (c(x,x_1),\ldots,c(x,x_n))$

m** is the emulator, and we can also calculate its variance

An example

y=7+x+cos(2x)

q=1, $h^T(x)$=(1 x) n=5



Conclusions

Model assessment constraints:
- **amount of data**
- **data quality**
- **ease of producing model runs**
- **degree of misalignment**

Ideally the model should have
- **similar first and second order properties to the data**
- **similar peaks and troughs to data (or simulations based on the data)**