My research interests

Peter F. Craigmile

The Ohio State University
department of STATISTICS

http://www.stat.osu.edu/~pfc/

Stat 8010

Department of Statistics

The Ohio State University

22 January 2016

Being a graduate student

- "Notes on the PhD Degree" by Douglas Comer.

  http://www.cs.purdue.edu/homes/dec/essay.phd.html

- "How to Be a Good Graduate Student" by Marie desJardins.

  http://www.cs.indiana.edu/how.2b/how.2b.html

- "Writing and Presenting your Thesis or Dissertation" by S. Joseph Levine.

  http://www.learnerassociates.net/dissthes/

- Stasny, Elizabeth (2001), "How to get a job in academics", The American Statistician, 55 (1), 35-40. http://www.jstor.org/stable/2685527

- N. Altman, D. Banks, J. Hardwick, K. Roeder, P. Craigmile, J. Hardin, and M. Gupta (2005), "The Institute of Mathematical Statistics New Researchers' Survival Guide"

  http://imstat.org/publications/books/NewResearchersGuide.pdf

Choosing a Ph.D. topic

- Question: which statistics classes do you enjoy?
- Read through some statistics journals. A good list is at

  http://www.statsci.org/jourlist.html
- Read through the OSU statistics web site, especially

  http://www.stat.osu.edu/people/faculty
- Talk to people (professors and students).
  - Try a reading course (not too many at once!)
- Your chosen Ph.D. topic should be **interesting** and **do-able**, but most importantly **fun**!

Some useful computing skills for your Ph.D.

- R (https://www.r-project.org) and R Studio (https://www.rstudio.com)
- R markdown (http://rmarkdown.rstudio.com)
- LaTeX and BibTeX (Look at "The Not So Short Introduction to LATEX 2ϵ" at https://tobi.oetiker.ch/lshort/lshort.pdf)
- Learn how to make presentations in LaTeX (Beamer, https://www.ctan.org/pkg/beamer?lang=en, is popular).
- If you have not already learned a programming language: C++ and Python (https://docs.python.org/2/tutorial/) are popular.
  - You can connect R and C++ with Rcpp (http://www.rcpp.org).
- Learn how to run jobs on a server (Hint: learn about the `ssh` program and some basic linux (http://www.tldp.org/LDP/Bash-Beginners-Guide/html/)

Some of my research interests

- Time series analysis, spatial processes, and space-time (spatio-temporal) modeling
- Longitudinal generalized linear modeling
- Long memory processes
- Frequency and time-frequency methods (useful for modeling nonstationary processes)
- Volatility modeling
- Methods for extremes
- Varied applications

  e.g., environmental sciences, psychology, climate science, paleoclimate, space weather.

Space-time modeling of climatic trends

- Some of this work is joint research with Peter Guttorp,

  University of Washington, Seattle and Norwegian Computing Center, Oslo.

  – Later work in collaboration with Veronica Berrocal, University of Michigan.

- Funded by the National Science Foundation.

Space-time modeling

- There is a growing interest in being able to model phenomena that vary **both in time and across space**, in the presence of uncertainty.
- Many interesting scientific questions involve investigating the **interaction** between time and space.
  - Modeling the processes marginally (i.e., solely in time, or solely over space), is not sufficient to answer these questions.
- Excellent **reviews** of space-time literature can be found in Gneiting and Schlather [2002] and Gneiting et al. [2007].
- Good **books**: Le and Zidek [2006] and Cressie and Wikle [2011].

Thinking about space-time data and modeling

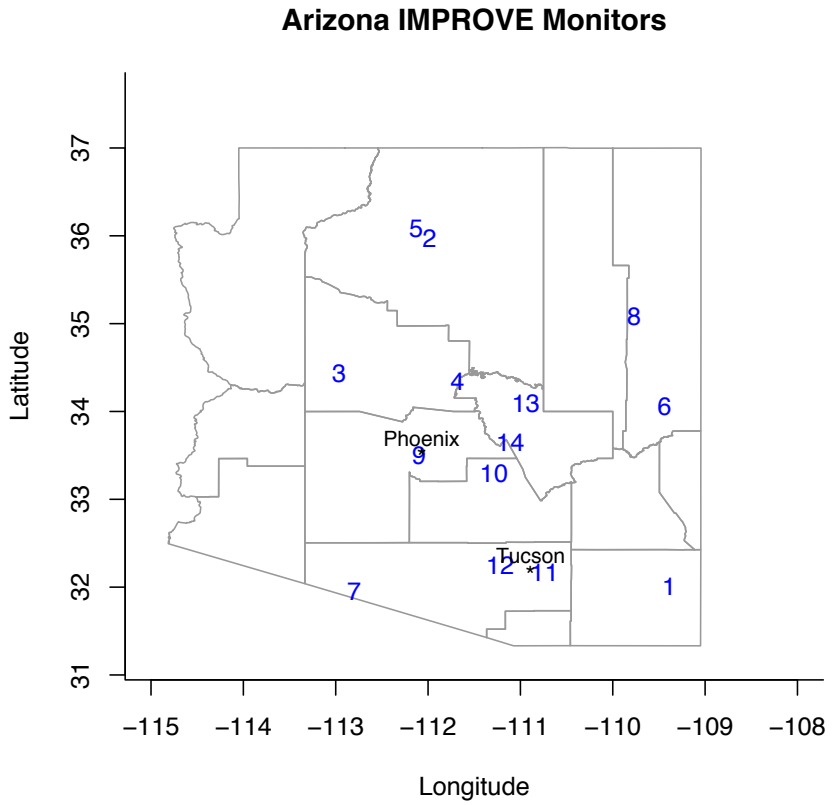
- The **format** of the data matters.
  - What are the spatial and temporal domains?
  - Is the data regular or irregular in some way?
- What **exploratory** data analysis techniques are appropriate?
- How can we write down statistical models that capture the uncertainty in space-time data?
- Often space-temporal datasets can be large or massive – this can limit our methods of analysis.

A good rule of thumb

- A space-time **statistical methodology** should be faithful to:
  - time series analysis *and*
  - spatial methods such as geostatistical process or areal unit modeling.
- Think of it this way:
  - In the absence of spatial data, inference should proceed via standard time series methods (e.g., Brockwell and Davis [2002]).
  - On the other hand, in the absence of multiple observations in time, use standard methods from geostatistics (e.g., Cressie [1993]; Stein [1999]; Banerjee et al. [2004]).

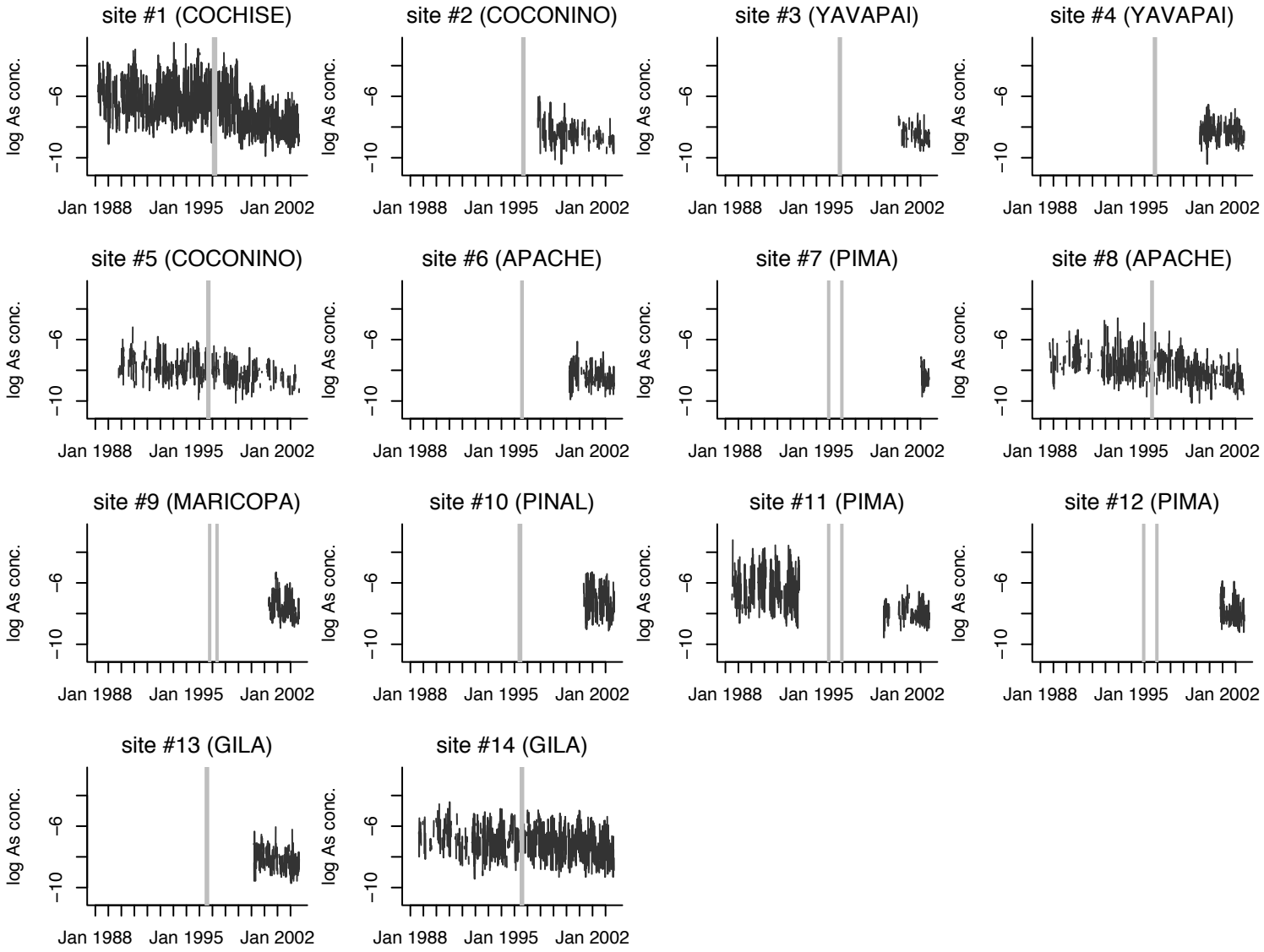
Ex: Interagency Monitoring of Protected Visual Environments

- Established 1985.
- Aim: to provide long-term air-quality records for U.S. national parks and wilderness areas.
- Readings (units of $\mu g/m^3$) are collected every 3-4 days
- Certain observations $<$ MDL.



- Q: Is it an issue that the data is only observed in rural regions?

Site-by-site time series



Spatio-temporal data

- Let $y(\boldsymbol{s}, t)$ denote the **data value** at location $\boldsymbol{s} \in D$ and time $t \in T$.
- If we have more than one observation at each location and time it could be **vector-valued**: $\boldsymbol{y}(\boldsymbol{s}, t)$.
- The temporal domain $T \subset \mathbb{Z}$ can be **discrete-time**.
  - Currently working with **Radu Herbei** on Stochastic Differential Equation (SDE) models in which the temporal domain $T \subset \mathbb{R}$ is **continuous-time**.
- The spatial domain D can be:
  1. a subset of $\mathbb{R}^d$: **geostatistical** in space, or
  2. it can index areas: **areal**/**lattice** in space.

Spatio-temporal processes

(Similar to with time series or a spatial process)

- We imagine $\{y(\boldsymbol{s}, t) : \boldsymbol{s} \in D, t \in T\}$ as being drawn from a stochastic **spatio-temporal process**

$$\{Y(\boldsymbol{s}, t) : \boldsymbol{s} \in D, t \in T\}$$

  (the "what could have beens").

- We model our data by writing down a statistical model for this **spatio-temporal process**.

A model for spatio-temporal processes

- Is the process **Gaussian** (do the data follow a joint normal distribution?)
- If not, can we **transform** to Gaussianity?
  - Sometimes, this is not possible (e.g., we have count data), and we need to consider more complicated processes.
- In the Gaussian case, as with space and time, we might consider the decomposition

$$Y(\boldsymbol{s}, t) = \mu(\boldsymbol{s}, t) + Z(\boldsymbol{s}, t)$$

$$\text{Process} = \text{Process mean} + \text{Noise}$$

What goes in the process mean?

- Typically, what we can account for. Anything left over goes into the noise. (this is a simplification).
- If we have other variables recorded in space and/or time, we might try to relate these **covariates** to the mean.
  - More simply, the mean could be a function of time (e.g., year) and space (e.g., latitude and longitude).
- We could use a linear regression to relate p **covariates** $x_j(\boldsymbol{s}, t)$, $j = 1, \ldots, p$ to the mean:

$$\mu(\boldsymbol{s}, t) \;=\; \beta_0 + \sum_{j=1}^{p} \beta_j \; x_j(\boldsymbol{s}, t).$$

What is left over?

- After we remove our estimate of the mean, we are left with our estimate of $Z(\boldsymbol{s}, t)$ (the noise).
- We usually assume that $Z(\boldsymbol{s}, t)$ is a **mean zero stationary process**.
- Does the estimate of the noise have mean zero? (if not, we may need to rethink the model for the mean)
- Is the estimated noise process stationary?

Stationary processes

- A **stationary spatio-temporal process** has

  1. A constant mean that does not depend on space and time.

  2. A covariance that only depends on spatial lag $\boldsymbol{h}$ and the temporal lag u, not the time and spatial points; i.e.,

$$\text{cov}(Z(\boldsymbol{s}, t), Z(\boldsymbol{s} + \boldsymbol{h}, t + u)) \;=\; C_Z(\boldsymbol{h}, u).$$

- As with spatial processes, we could further simplify by assuming **isotropy** in space (only depends on the distance between locations, not the direction).

- We could also assume **separability** – the covariance in space is independent of the covariance in time (usually an unrealistic assumption).

Covariance models

- Research continues into **covariance models**, $C_Z(\boldsymbol{h}, u)$, for spatio-temporal processes.

  (variograms do not tend to be used for spatio-temporal modeling)

- Many models that have been proposed share features of spatial and temporal models, but allow for interactions between space and time.

- Many (approximate) covariance models have been proposed that can be used for large datasets.

  References: Gneiting and Schlather [2002], Le and Zidek [2006], Gneiting et al. [2007], and Cressie and Wikle [2011].

Fitting space-time models

- The methods of analysis is the same as for time series and spatial models:
  - Least squares methods;
  - Maximum likelihood estimation;
  - Bayesian inference.
- But, there is a lack of general purpose software for fitting space-time models.

  (There are a few R packages, but more commonly you need to write your own code!)

Spatio-temporal climate modeling

- Climate data is collected over **space and time**.
- Usually our scientific questions of interest (e.g., studying climate change) must acknowledge the fact that:
  - **Changes** of climate phenomena over time vary spatially

    (and conversely, changes over space can vary in time).
- **Statistical models** that we build to investigate scientific hypotheses must allow for these **space-time interactions**, while accounting for the **uncertainty** in space-time climate data.

Statistical features of climate data

- Can be **Gaussian**- or **non-Gaussian**-distributed.

  (e.g., mean temperatures are often Gaussian,

  whereas precipitation or maximum temperatures are not).

- **Smooth** changes of the **mean** over long temporal and spatial scales

  (**the trend**) are often non-linear.

- Certain features of climate data are **seasonal**.

- Climate data exhibit **non-constant variances**, often over seasons.

- Even after accounting for trend and seasonalities,

  long-range correlations remain.

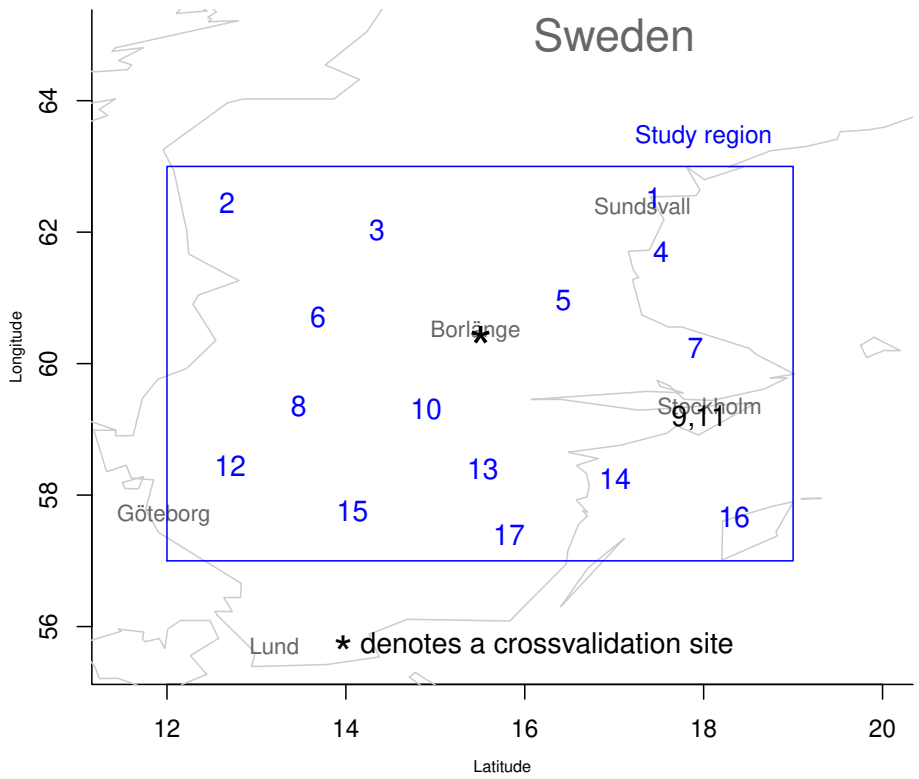
Space-time modeling of temperature trends

- Analyses of **temperature** data in **space** and **time** play a crucial role in studying climate change.
- The estimation of **global mean temperature** has been one of the most common indicators of global warming, although it is not one of the most sensitive.
- The World Meteorological Organization Commission on Climatology in Feb 2010 endorsed a proposal for an international collaborative effort to produce a new generation of land surface air temperature datasets.

Our motivation: temperatures in central Sweden

[For further details see Craigmile and Guttorp, 2011]

- **Daily average temperatures** for 1961–2008 from Swedish Meteorological and Hydrological Institute historical network database of synoptic temperatures.
- Data not homogenized, but undergone quality control.
- Series vary in length (13-48 years).



- We need to account for a small amount of missing data.

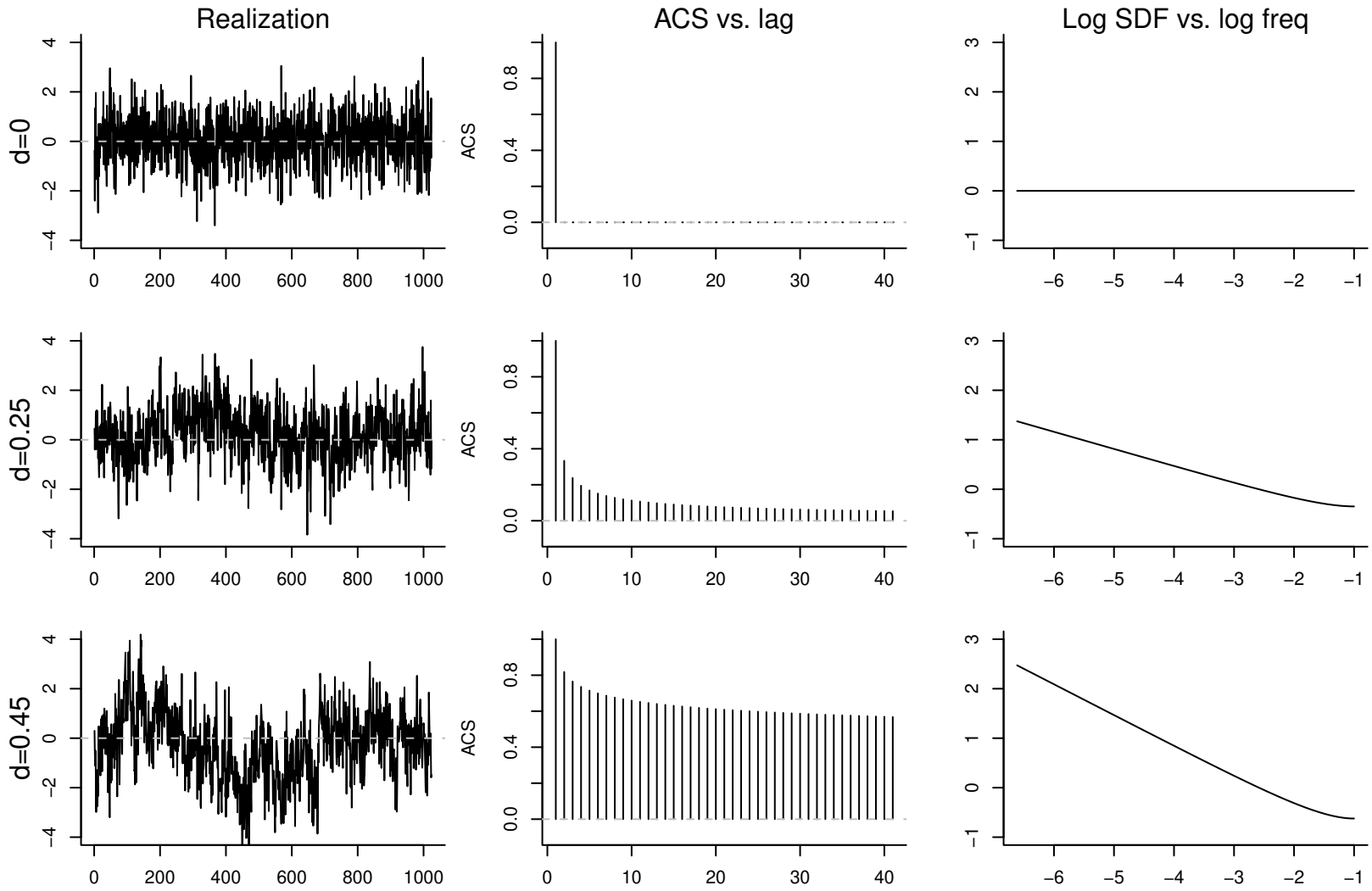
Our modeling aim

- To characterize any **space-time trends** observed in the temperatures over central Sweden.
  - We must also be able to characterize the seasonal and irregular noise components inherent in these series.
  - Due to the self-similar behavior of climate processes over long time scales a commonly used noise model is the **long range dependent process** [e.g., Hussain and Elbergali, 1999, Koutsoyiannis, 2003, Craigmile et al., 2004, Cohn and Lins, 2005].

What is long range dependence?

- Long range dependent processes are time series models in which the autocorrelations **decay slowly** with increasing lags.
  - Equivalently, the spectral density function (SDF) is **unbounded** at zero frequency.
- Visual features of the data:
  - Relatively long periods of large and small values.
  - Over short periods of time there is evidence of trends and seasonality. They disappear as the period length increases.
  - BUT, the time series **looks stationary**.

Long range dependence, continued



Long range dependence and trend

- In the context of a single time series analysis, Craigmile et al. [2004] provided **wavelet**-based estimates of trend, confidence regions, and significance tests when a trend is observed in the presence of long range dependence.
  - Appropriate for modeling **monthly** or **yearly** average temperatures.
  - Not for **daily** mean temperatures [See also Caballero et al., 2002].
  - This advocates the use of an error process that can capture both **short** and **long** range dependence.
- We consider a **space-time** long range dependence model.
  - Not the first example – see Haslett and Raftery [1989].

A spatially varying additive classical decomposition

- At each spatial location $\boldsymbol{s}$ and for each time t we model the temperature process by

$$Y_t(\boldsymbol{s}) \;=\; \mu_t(\boldsymbol{s}) + s_t(\boldsymbol{s}) + \zeta_t(\boldsymbol{s}),$$

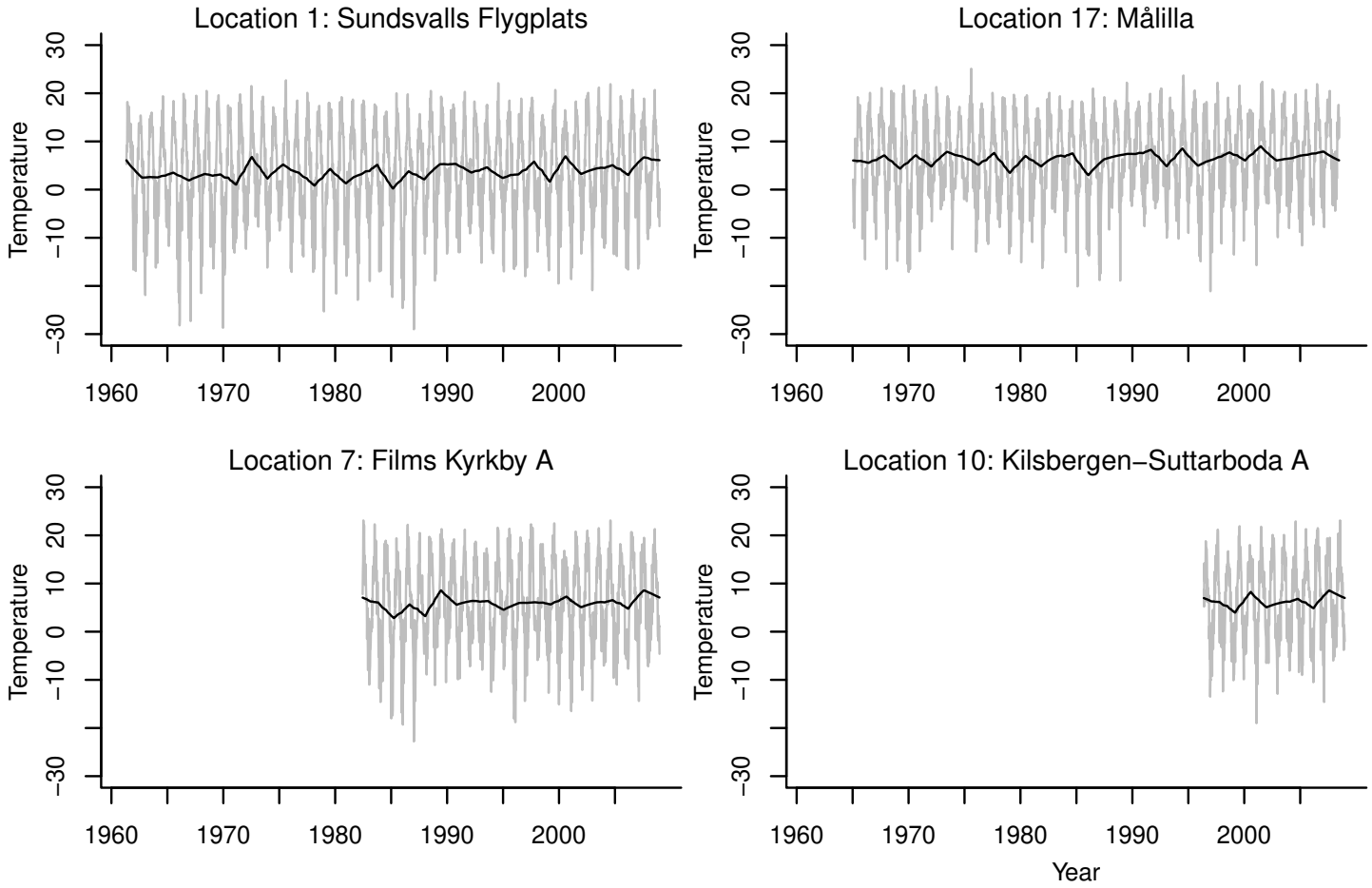
  where

  $\mu_t(\boldsymbol{s})$ is the trend component,

  $s_t(\boldsymbol{s})$ is a seasonal component with period d (with average zero), and

  $\zeta_t(\boldsymbol{s})$ is a mean zero random (noise) component.

Site-by-site analysis

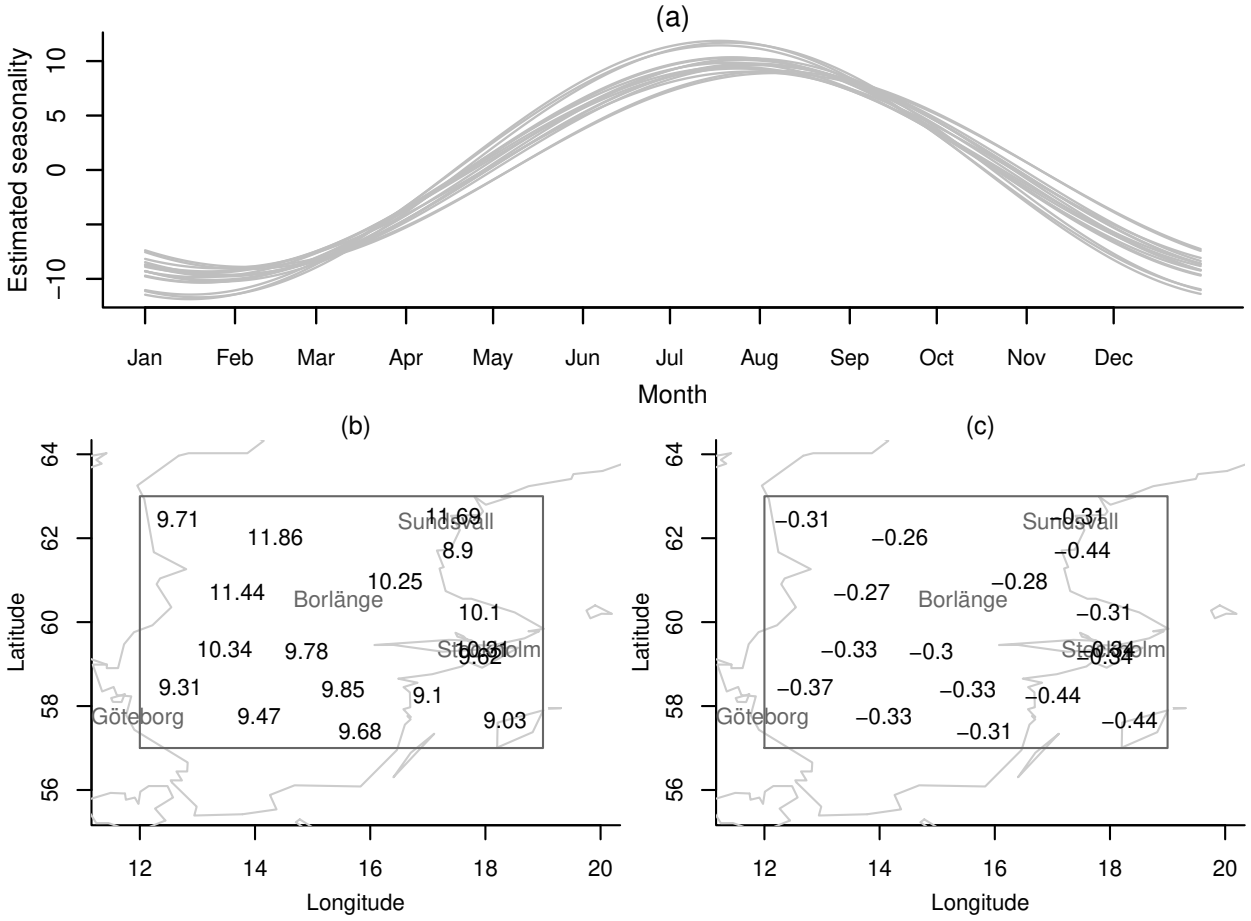


Daily average temperature

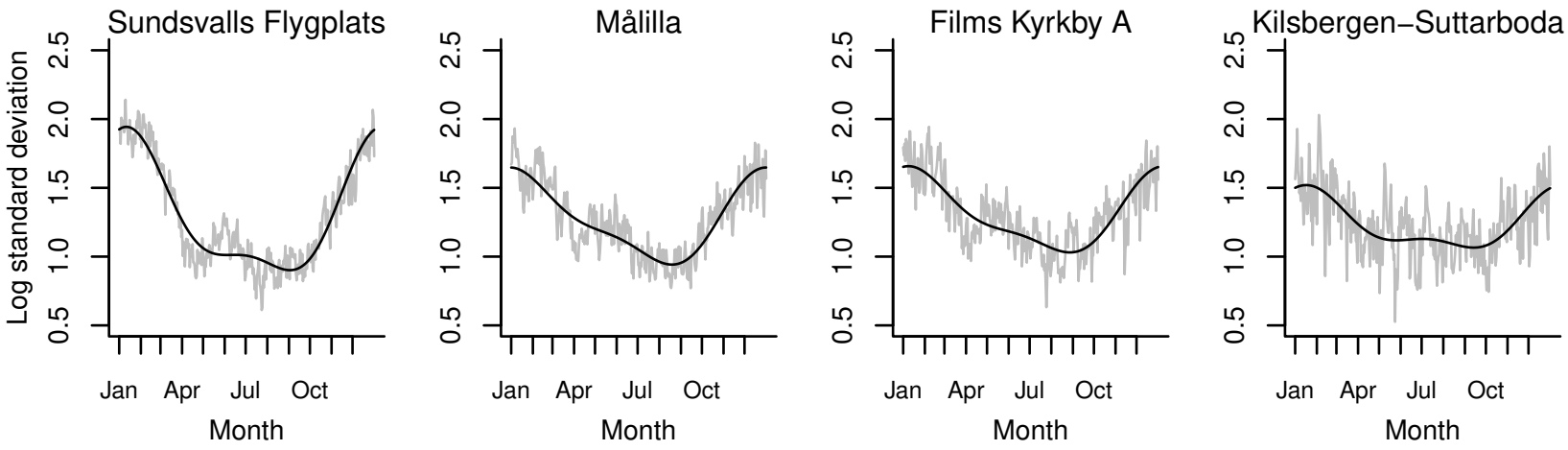
Naive wavelet-based estimates of the trend, $\mu_t(\boldsymbol{s})$.

Seasonal patterns

$$s_t(\boldsymbol{s}) \;=\; A(\boldsymbol{s})\cos(2\pi t/365.25 + \phi(\boldsymbol{s}))$$



Seasonal patterns, continued



- To capture the strong **seasonal** patterns in the **variance** let

$$\zeta_t(\boldsymbol{s}) = \sigma_t(\boldsymbol{s})\ \eta_t(\boldsymbol{s})$$

$$\begin{aligned}\log \sigma_t(\boldsymbol{s}) \;\; = \;\; & \alpha_0(\boldsymbol{s}) + \alpha_1(\boldsymbol{s}) \sin(2\pi t/365.25) + \alpha_2(\boldsymbol{s}) \cos(2\pi t/365.25) + \\ & \alpha_3(\boldsymbol{s}) \sin(2\pi t/182.625) + \alpha_4(\boldsymbol{s}) \cos(2\pi t/182.625).\end{aligned}$$

Models incorporating long and short range dependence

- $\{\eta_t : t \in \mathbb{Z}\}$ is a **stationary Gaussian** process with zero mean and spectral density function (SDF)

$$S_\eta(f) \;=\; B(f)\,|4\sin^2(\pi f)|^{-\delta}.$$

- Usually $|4\sin^2(\pi f)|^{-\delta}$ controls the **long range** dependence:

  $\delta \in (-1/2, 1/2)$ is the **difference parameter**.

  (extending to higher values of δ leads to nonstationary processes).

- $B(f)$ is some model of the **short range** dependence (almost!)

Examples of $B(\cdot)$

1. When $B(f) = \sigma^2$ for all f:

   $\{\eta_t\}$ is a fractionally differenced (FD) process [Granger and Joyeux, 1980, Hosking, 1981].

2. $B(\cdot)$ is the SDF of an autoregressive moving average process:

   $\{\eta_t\}$ is an autoregressive fractionally integrated moving average (ARFIMA) process

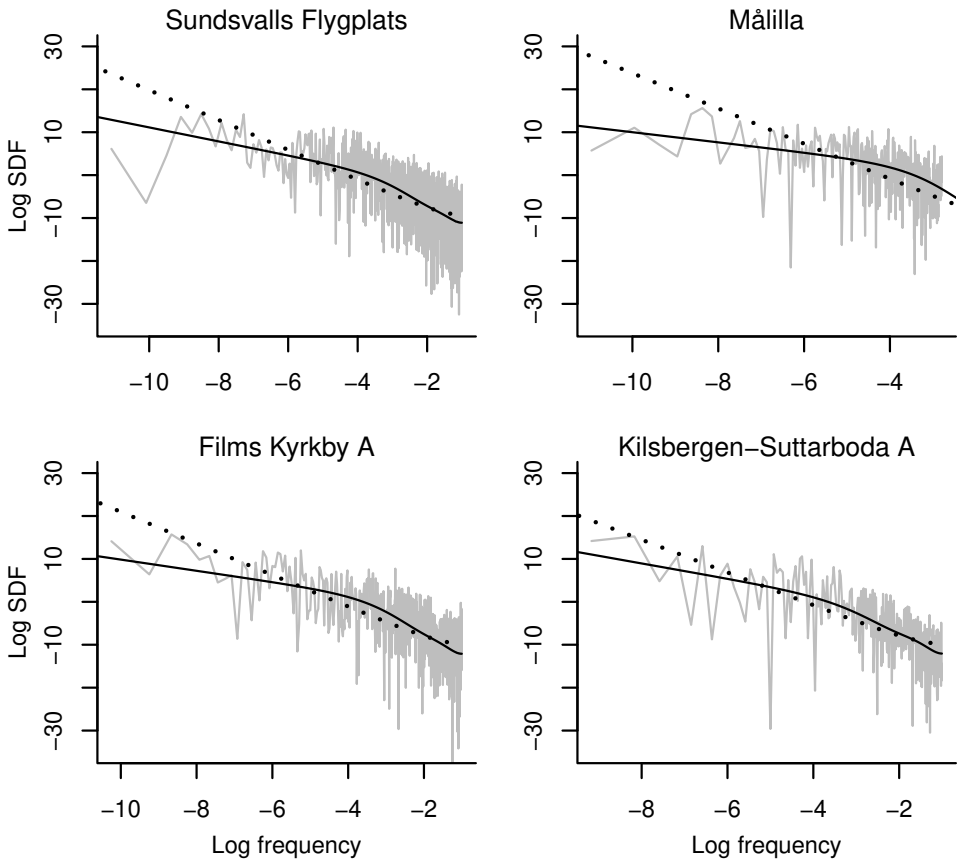
3. $B(\cdot)$ is the SDF of an exponential process [Bloomfield, 1973] (a truncated Fourier basis):

   $\{\eta_t\}$ is called the fractional exponential (FEXP) process [Beran, 1993].

Estimated SDFs of standardized noise

Dotted: FD process

Solid: FEXP process, $p=3$



- Strong evidence of both **short** and **long** range dependence.
- **Spatial** patterns in the parameter estimates.

- Negative association between the long range dependent parameter and either latitude or log10 of elevation [See also Király and Jánosi, 2005]

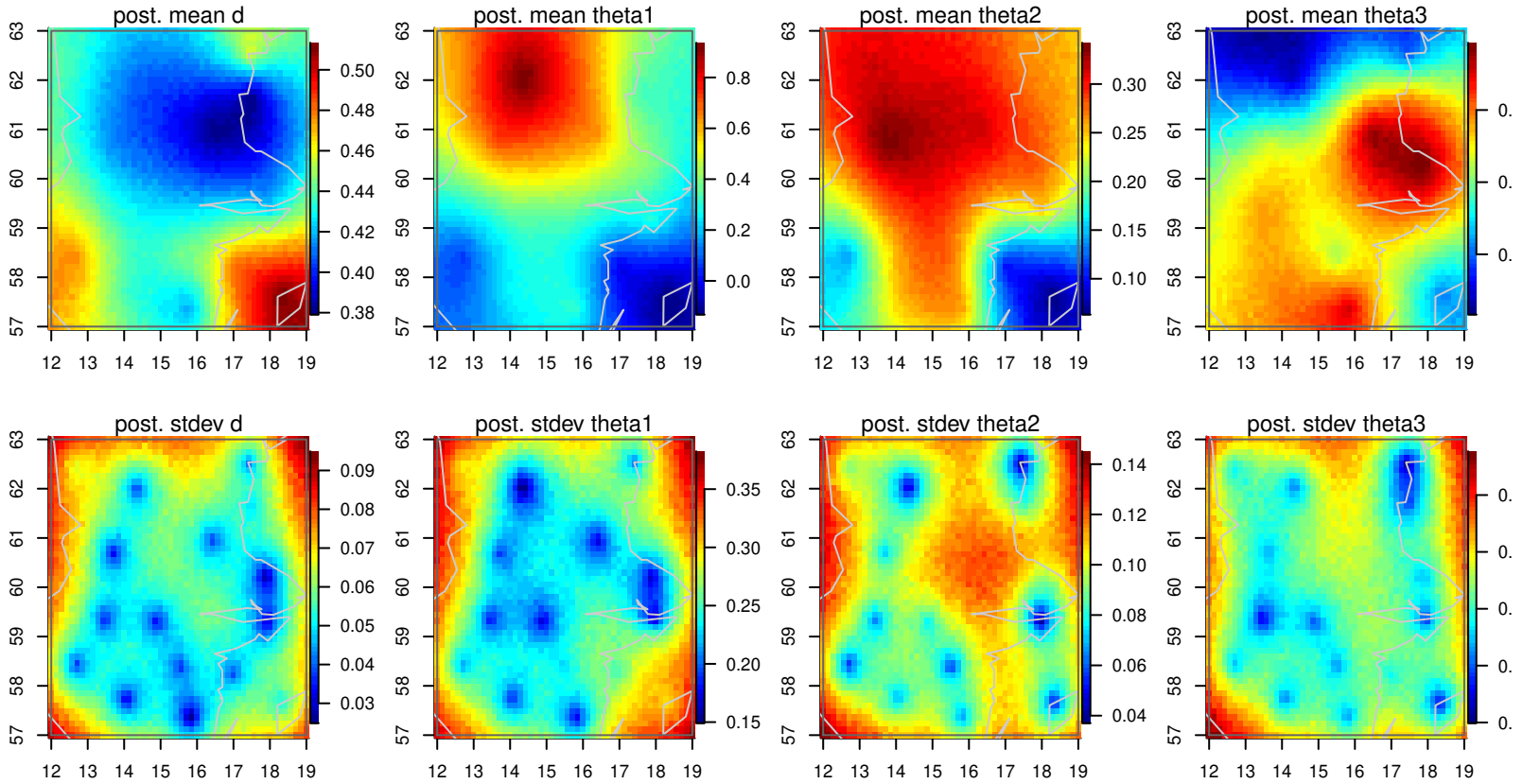
Completing the space-time model

- We include a Gaussian measurement process.
- We use a **wavelet**-based space-time process for the trend to captures **smooth** changes in the mean temperature over scales greater than a year.
  - The space-time covariance is assumed separable; AR(1)-in-time, and exponential over space.
- We assume **Gaussian spatial processes** for the spatially-varying parameters (or transformations of the parameters).

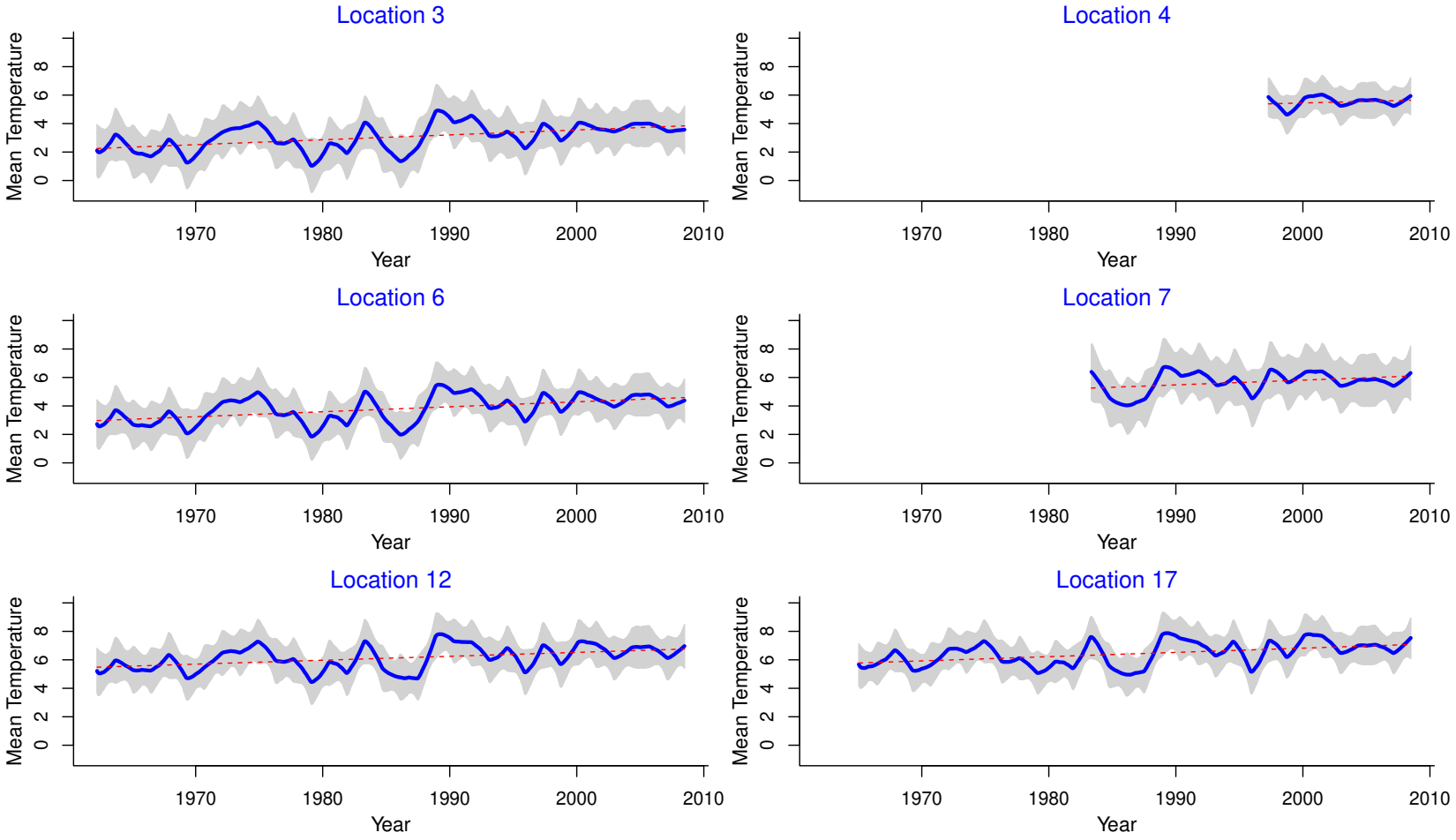
Bayesian inference via Markov chain Monte Carlo

- We carry out inference using a wavelet transform of the temperature series at each location.
- Fix measurement error stdev at 0.2^oC [Folland et al., 2001].
- Use MCMC to draw samples from posterior distribution.
- Discard first 2,000 draws, sample next 20,000 discarding every 10.

Posterior summaries of the long range dependent parameters

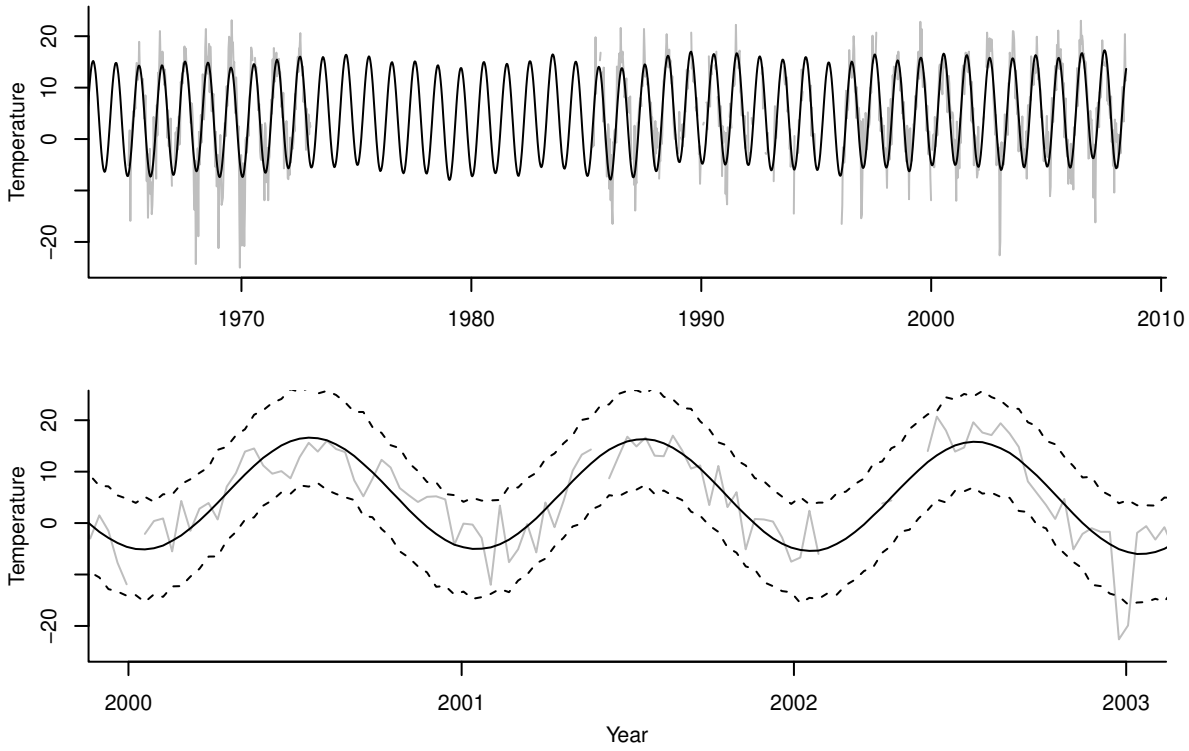


Posterior summaries of the trends



- Average temperatures are non-constant and oscillate over decades.
  - Over longer scales, there is weak evidence of a linear increase.

Validation using temperatures at Borlänge Flygplats



- Our model validates well – 96% of the observed values lie within the posterior predictive bands.

Discussion

- We carried out various sensitivity studies.
- Space-time models can be defined via a **classical decomposition** to simultaneously model trend, seasonality, and noise. Once we model these terms accurately, then we assess the **significance** of trends.
- Growing interest in modeling **nonstationarities** in space and time.
  - **Wavelet** transforms of each time series have the ability to naturally capture such non-stationary-in-time behavior.
  - Wavelet transformations in space may be less applicable in practice.
- We have also compared to Norwegian Computer Center climate model output [Berrocal et al., 2012].

References

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall, Boca Raton, FL., 2004.

J. Beran. Fitting long-memory models by generalized linear regression. *Biometrika*, 80:817–822, 1993.

V. Berrocal, P. Craigmile, and P. Guttorp. Regional climate model assessment using statistical upscaling and downscaling techniques. *Environmetrics*, 23:482–492, 2012.

P. Bloomfield. An exponential model for the spectrum of a scalar time series. *Biometrika*, 60:217–226, 1973.

P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting (Second Edition)*. Springer-Verlag, New York, 2002.

R. Caballero, S. Jewson, and A. Brix. Long memory in surface air temperature: detection, modeling, and application to weather derivative valuation. *Climate Research*, 21:127–140, 2002.

T. A. Cohn and H. F. Lins. Nature's style: Naturally trendy. *Geophysical Research Letters*, 32:5, 2005.

P. Craigmile, P. Guttorp, and D. B. Percival. Assessing nonlinear trends using the discrete wavelet transform. *Environmetrics*, 15(4):313–335, 2004.

P. F. Craigmile and P. Guttorp. Space-time modeling of trends in temperature series. *Journal of Time Series Analysis*, 2011. To appear.

N. A. C. Cressie. *Statistics for Spatial Data (Revised edition)*. Wiley-Interscience, New York, 1993.

N. A. C. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, New York, 2011.

C. K. Folland, N. A. Rayner, S. J. Brown, T. M. Smith, S. S. P. Shen, D. E. Parker, I. Macadam, P. D. Jones, R. N. Jones, N. Nicholls, and S. M. H. Global temperature change and its uncertainties since 1861. *Geophysical Research Letters*, 28:2621–2624, 2001.

T. Gneiting and M. Schlather. Space-time covariance models. In A. H. El-Shaarawi and W. W. Piegorsch, editors, *Encyclopedia of Environmetrics*, volume 4, pages 2041–2045. John Wiley & Sons, Chichester, 2002.

T. Gneiting, M. G. Genton, and P. Guttorp. Geostatistical space-time models, stationarity, separability and full symmetry. In B. Finkenstadt, L. Held, and V. Isham, editors, *Statistical Methods for Spatio-Temporal Systems*, pages 151–175. Chapman & Hall/CRC, Boca Raton, FL., 2007.

C. W. J. Granger and R. Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1:15–29, 1980.

J. Haslett and A. E. Raftery. Space-time modeling with long-memory dependence – assessing Ireland's wind power resource. *Journal of the Royal Statistical Society, Series C*, 38:1–50, 1989.

J. R. M. Hosking. Fractional differencing. *Biometrika*, 68:165–176, 1981.

S. Hussain and A. Elbergali. Fractional order estimation and testing, application to Swedish temperature data. *Environmetrics*, 10:339–349, 1999.

A. Király and I. M. Jánosi. Detrended fluctuation analysis of daily temperature records: Geographic dependence over Australia. *Meteorology and Atmospheric*

Physics, 88:119–128, 2005.

D. Koutsoyiannis. Climate change, the hurst phenomenon, and hydrologic statistics. *Hydrological Sciences Journal*, 48(1):3–24, 2003.

N. D. Le and J. V. Zidek. *Statistical Analysis Of Environmental Space-Time Processes*. Springer, New York, 2006.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, 1999.