

# **Multivariate methods for environmental space-time processes; a comparison with a dynamic linear model based alternative**

Nhu Le

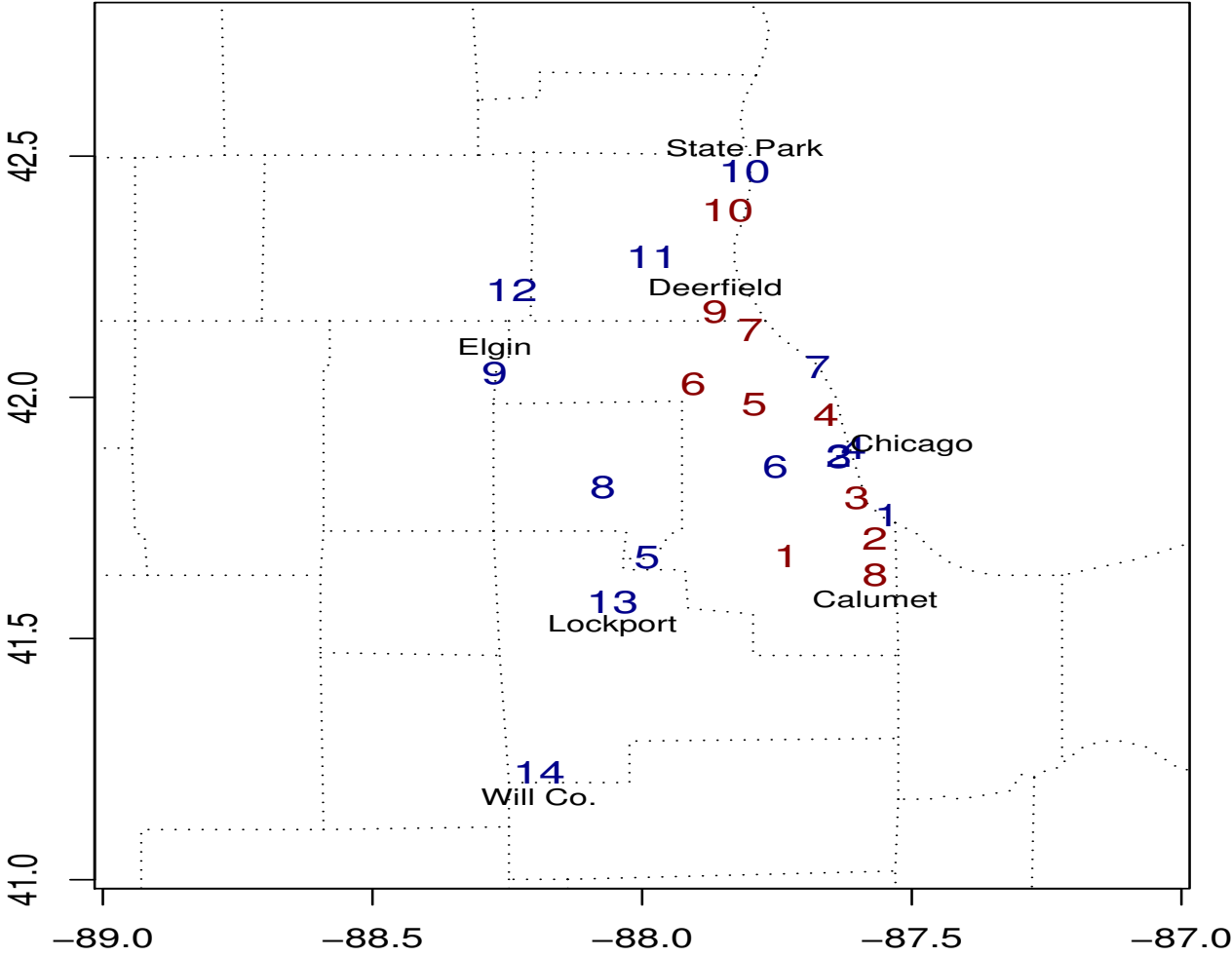
UBC & BC Cancer Research Centre

(Joint work with Yiping Dou and Jim Zidek)

# Outline

- Data: Hourly Ozone Levels (EPA-AIRS Database)
- Spatial Correlation Leakage
- Multivariate Bayesian Spatial Prediction (BSP) Approach
- Dynamic Linear Model-based (DLM) Approach
- Results
- Concluding Remarks

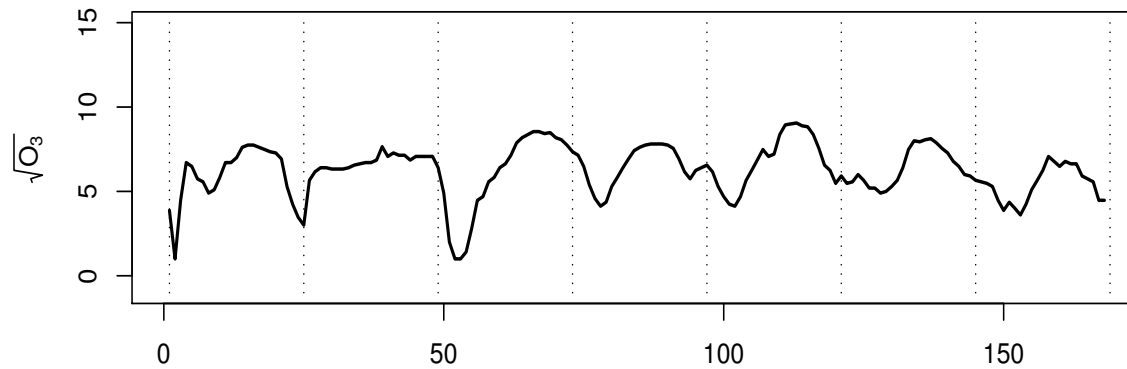
# AIRS locations: Gauged and Ungauged



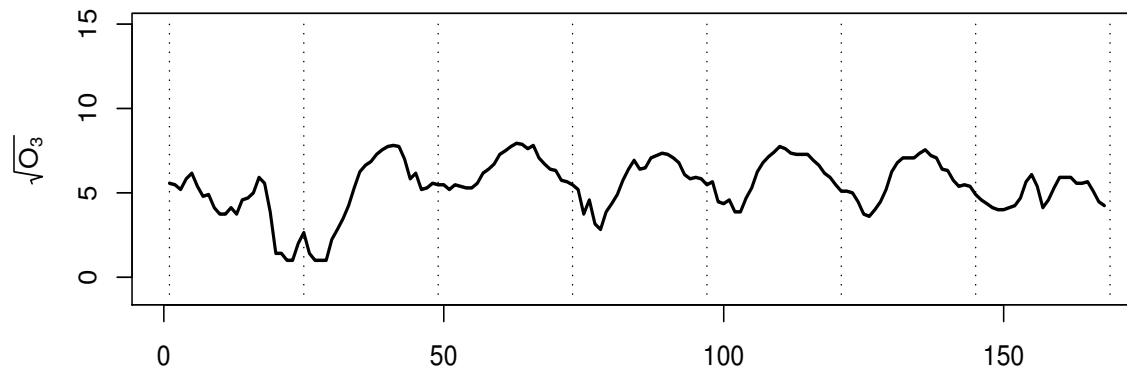
# Hourly Observations: May 1-Aug 31, 2000

## Week 1: Ozone levels (ppb)

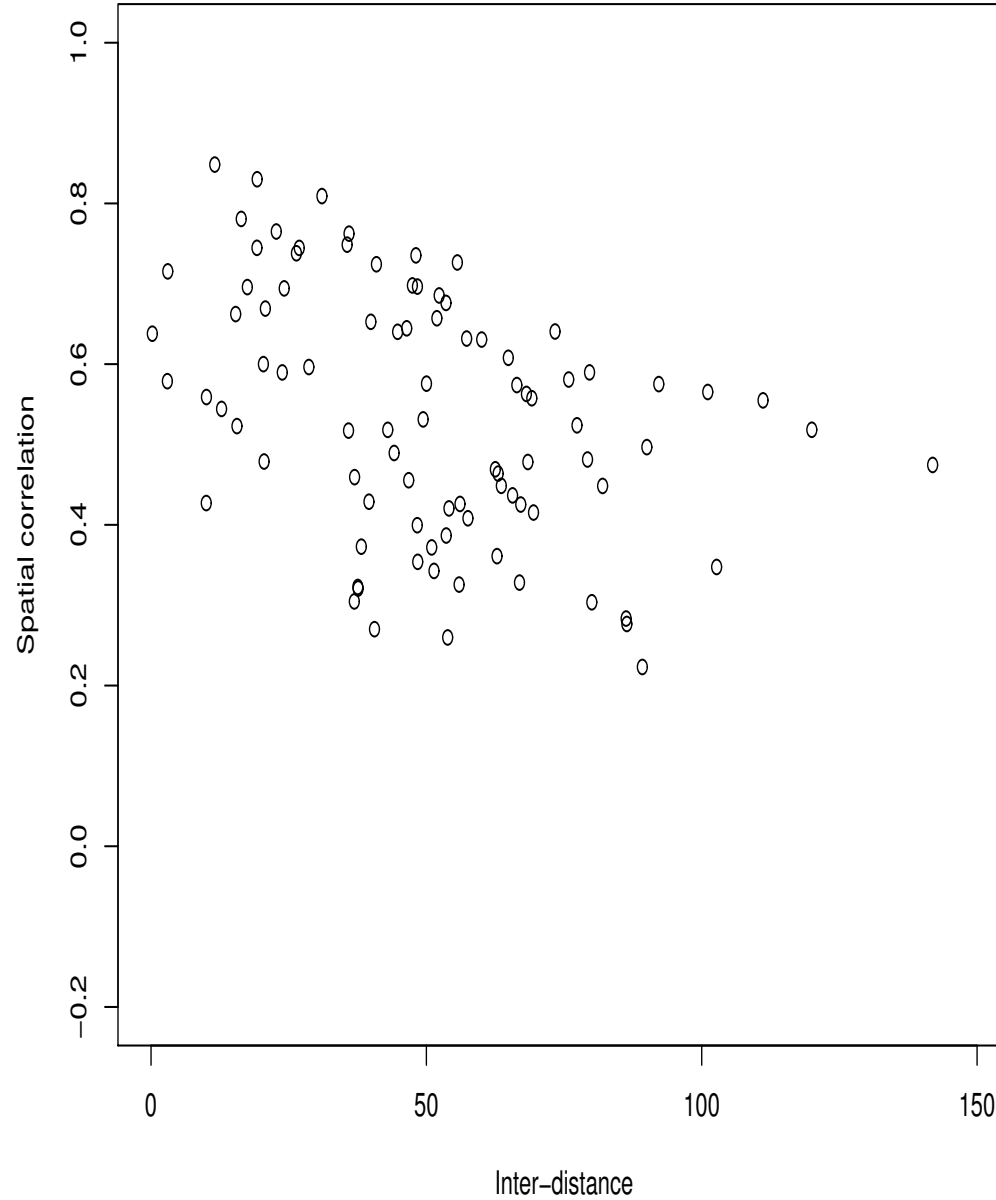
Station 10



Station 13

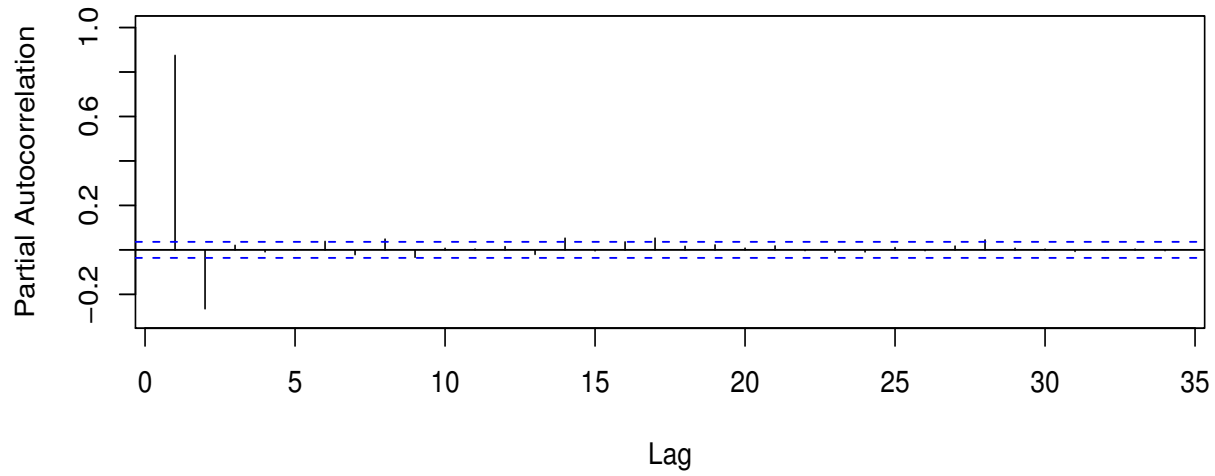


# Spatial Correlation - detrended series

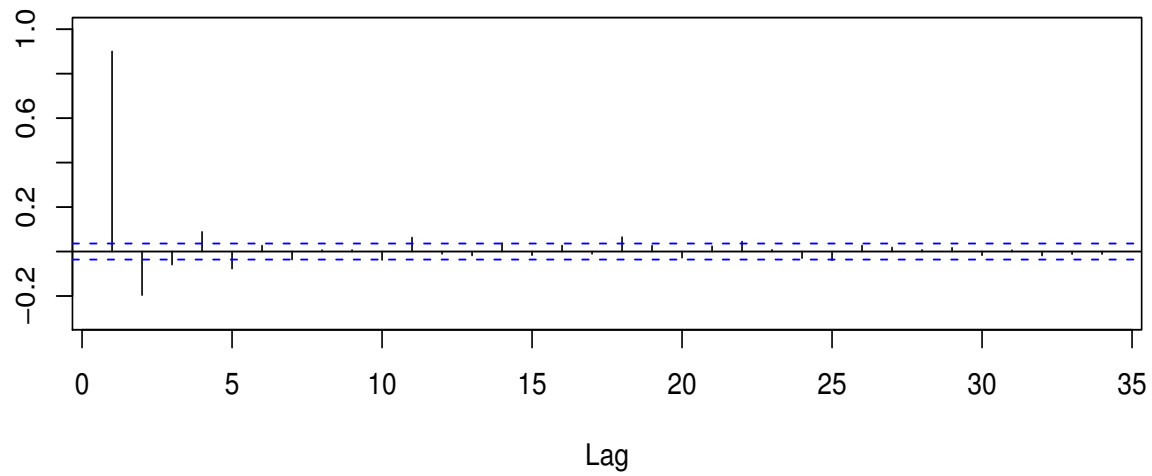


# Partial Autocorrelation - detrended series

Station 7

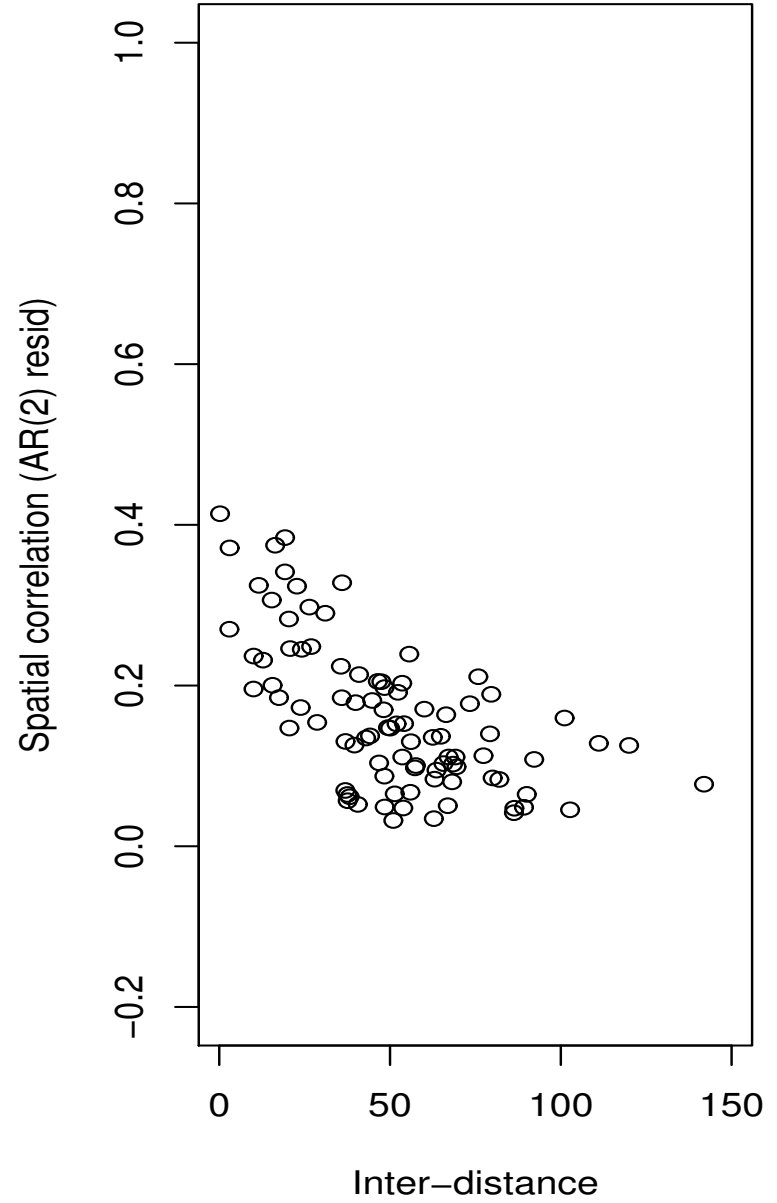
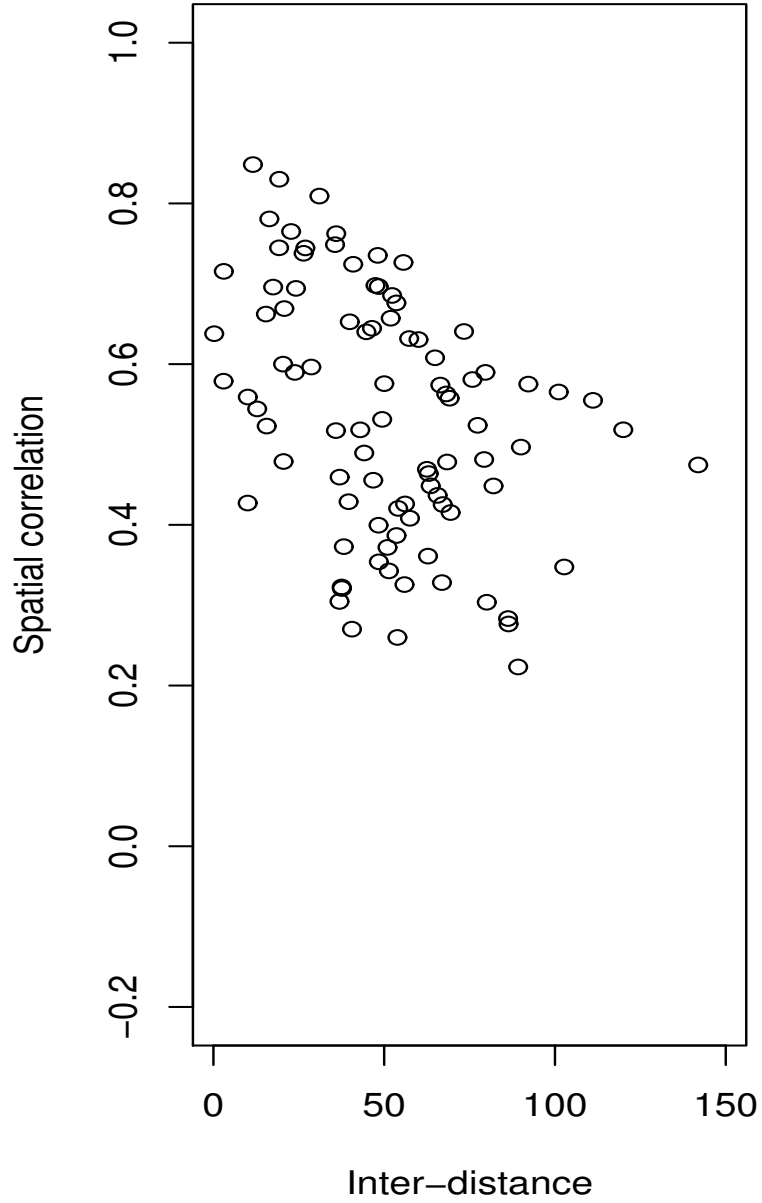


Station 9





# Spatial Correlation Leakage



## Spatial Correlation Leakage - Simple Case (Zidek et al 2002)

AR(1) Model:  $y_t(s) = \alpha y_{t-1}(s) + \epsilon_t(s)$

$\epsilon_t(s)$  – time independent with spatial correlation

Spatial correlation:

$$\text{cor}(\epsilon_t(s), \epsilon_t(s')) = \text{cor}(y_t(s), y_t(s')) -$$

$$\frac{\alpha}{\sqrt{(1-\alpha^2)}} [\text{cor}(y_{t-1}(s), \epsilon_t(s')) + \text{cor}(y_{t-1}(s'), \epsilon_t(s))]$$

Cross-corr = 0  $\rightarrow$   $\text{cor}(y_t(s), y_t(s')) = \text{cor}(\epsilon_t(s), \epsilon_t(s'))$

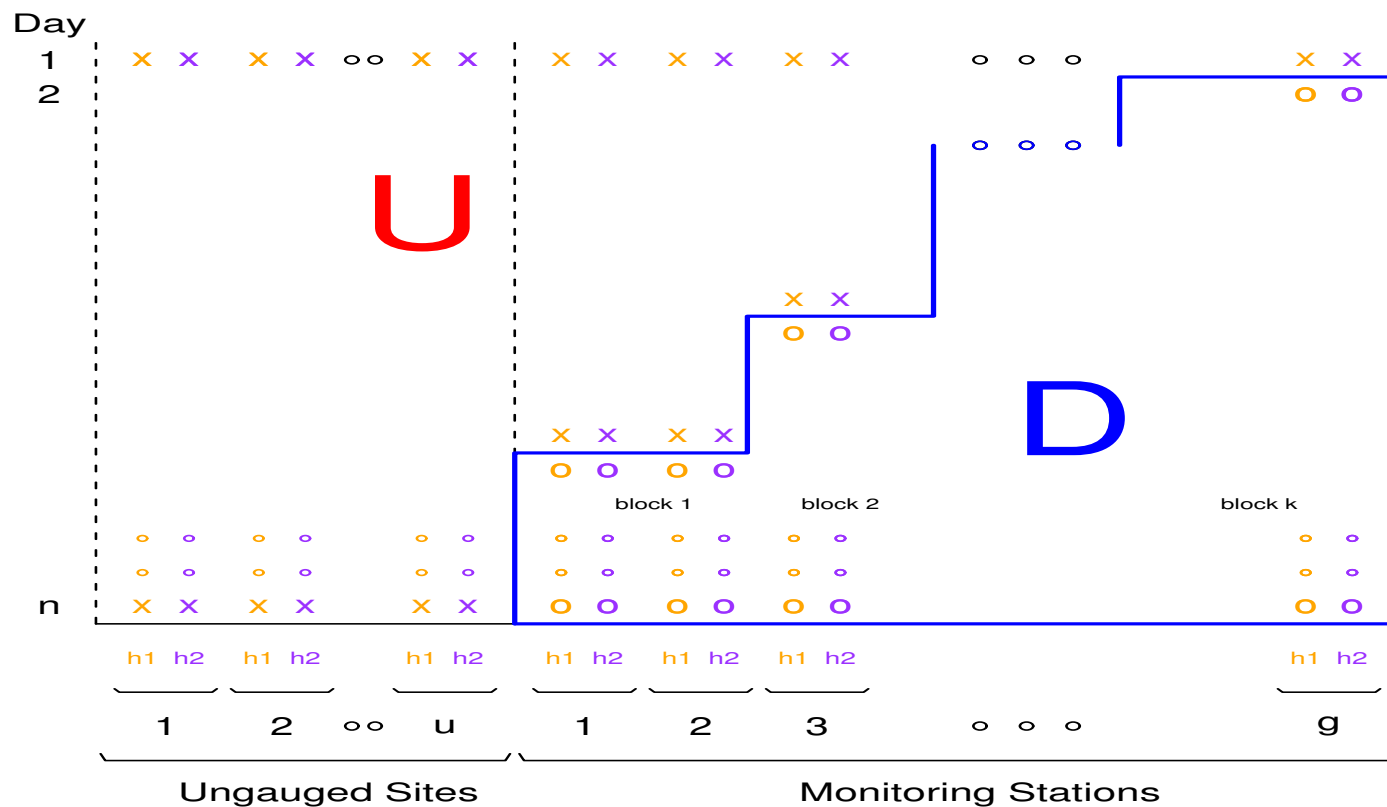
Correlation leakage occurs since sample ones  $\neq 0$

- substantial when  $\alpha$  is large.

# Multivariate Bayesian Spatial Prediction (BSP) approach

Multivariate response: hourly values within day

$$Y = [Y_{ij(p)}] = \left[ Y^{[u]}, \left( \begin{array}{c} Y^{[g_1^m]} \\ Y^{[g_1^o]} \end{array} \right), \dots, \left( \begin{array}{c} Y^{[g_k^m]} \\ Y^{[g_k^o]} \end{array} \right) \right]$$



# BSP - Model

Model

$$\left\{ \begin{array}{l} Y \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, A \otimes \Sigma) \\ \boldsymbol{\beta} \mid \Sigma, \boldsymbol{\beta}_0, F \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma) \\ \Sigma \sim GIW(\boldsymbol{\Theta}, \delta) \end{array} \right.$$

$A$ : **Assumed known**

Ex. Multiple pollutants, monthly average:  $A = I_n$

Special case - no staircase

$$\Sigma \sim IW(\Psi, \delta)$$

Le & Zidek (1992, 2006)

## BSP: Predictive distribution

- $(Y_U | D, \mathcal{H}) \sim \left( Y^{[u]} | Y^{[g_1^m, \dots, g_k^m]}, D, \mathcal{H} \right) \times$

$$\prod_{j=1}^{k-1} \left( Y^{[g_j^m]} | Y^{[g_{j+1}^m, \dots, g_k^m]}, D, \mathcal{H} \right) \times \left( Y^{[g_k^m]} | D, \mathcal{H} \right)$$

- Each component follows a **matric-t distribution**  
Mean, covariance, and df: functions of  $\mathcal{H}$  and  $D$
- **Completely characterized given  $\mathcal{H}$**  (all hyper's)

- $\mathcal{H}$ : Empirical Bayes

Computation simpler

Separability -  $\Psi = \Lambda \otimes \Omega$

Non-stationarity (Sampson & Guttorp 1992)

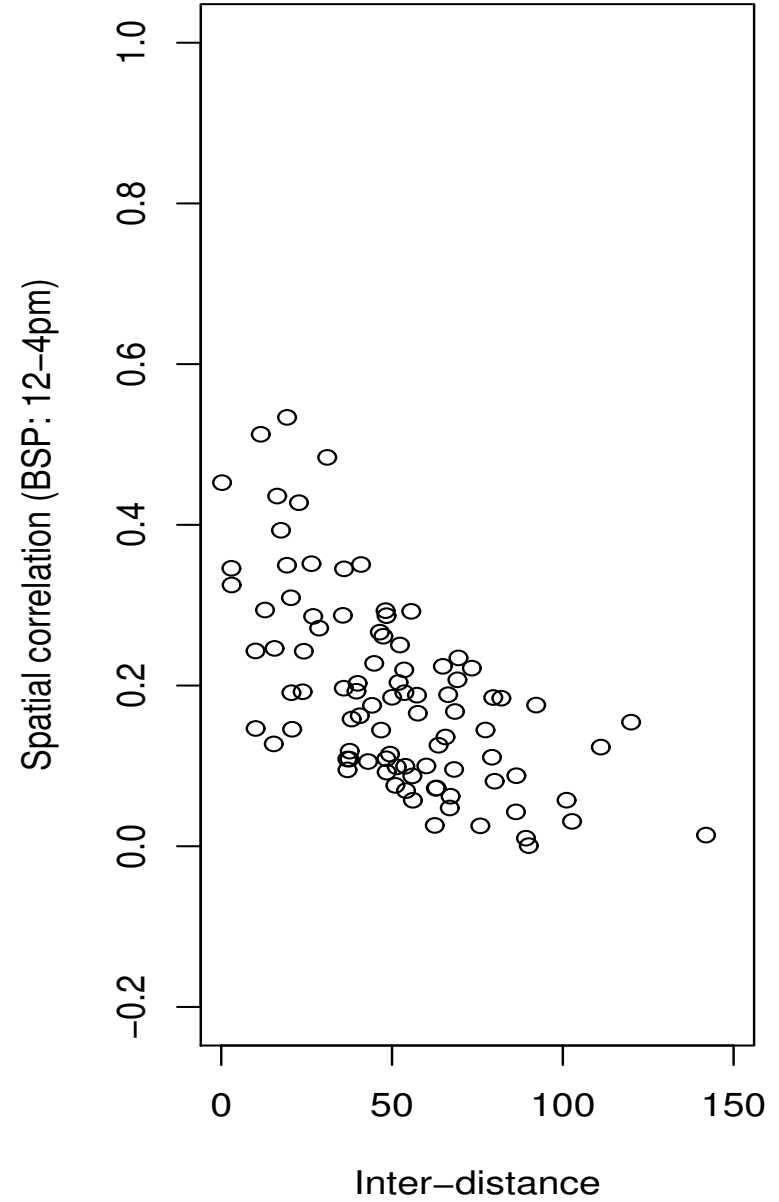
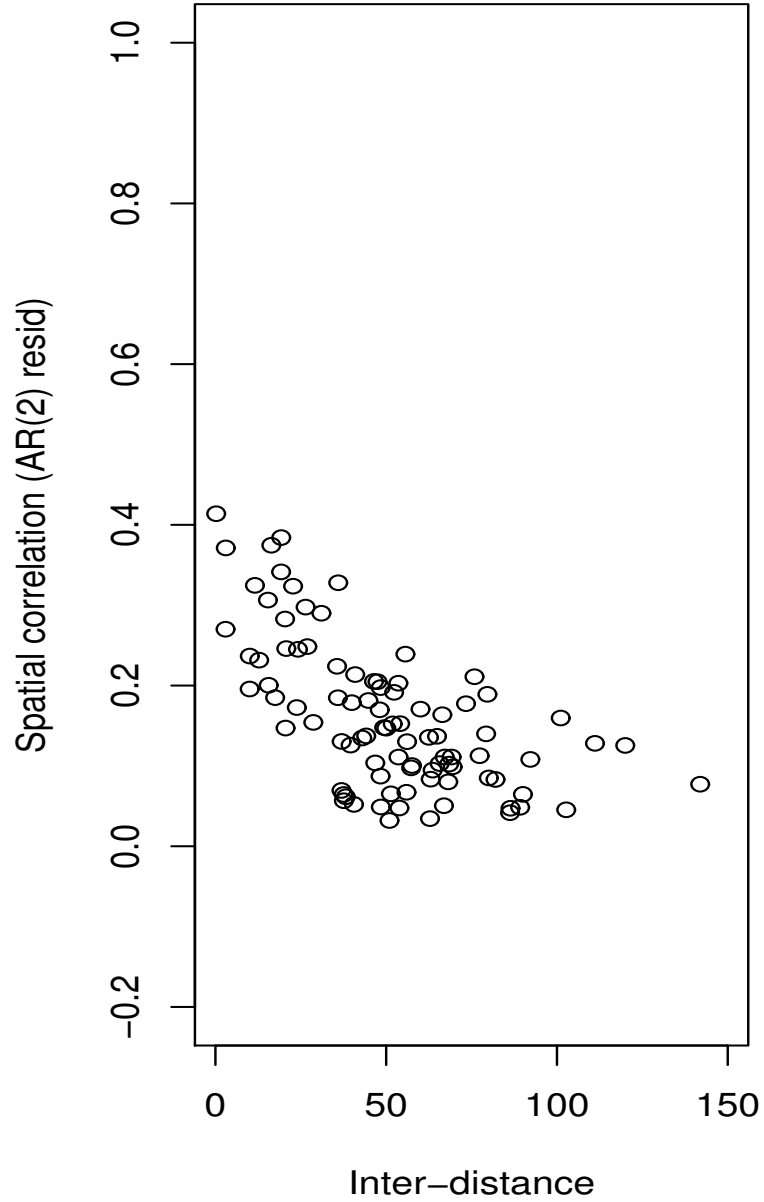
## BSP - Adaptation

Need to deal with  $A$  - assumed known in the theory

$$Y \mid \beta, \Sigma \sim N(Z\beta, A \otimes \Sigma)$$

- Estimated using data - Prefiltering
- Use only 5 hours to predict the 5<sup>th</sup> hour,  $A = I_n$ 
  - Ex. Obs 4-8pm to predict 8pm at ungauged locations
  - Partial use of data
  - Opportunity to capture hourly changes in covariance
  - Repeat 24 times (about 30' each)

# BSP: Improvement in spatial correlation



## BSP: Benefits from between hour correlation

$$\Psi = \Lambda \otimes \Omega \text{ (between stations and hours)}$$

$$\Omega = \begin{array}{c} \begin{array}{ccccc} & 12\text{pm} & 1\text{pm} & 2\text{pm} & 3\text{pm} & 4\text{pm} \end{array} \\ \left[ \begin{array}{ccccc} 1.00 & 0.76 & 0.54 & 0.44 & 0.37 \\ 0.76 & 1.00 & 0.76 & 0.58 & 0.49 \\ 0.54 & 0.76 & 1.00 & 0.80 & 0.66 \\ 0.44 & 0.58 & 0.80 & 1.00 & 0.81 \\ 0.37 & 0.49 & 0.66 & 0.81 & 1.00 \end{array} \right] \end{array}$$

# DLM Approach

Stroud, Muller, Sanso (2001), Huerta, Sanso, Stroud (2004)

Model:

$$Y_t(s) = Z_t(s)' \beta_t + S_{1t} \alpha_{1t}(s) + S_{2t} \alpha_{2t}(s) + \epsilon_t(s)$$

- $S_{jt}$  sine's + cosine's for 12hr and 24hr cycles
- $\alpha_{jt}$  capture amplitudes
- $\text{Cov}(\epsilon_t) = \sigma_y^2 \exp(-D/\lambda)$  induces regional smoothness
- Parameters change dynamically - random walk

$$\begin{aligned}\beta_t &= \beta_{t-1} + \omega_t \\ \alpha_{jt}(s) &= \alpha_{j,t-1}(s) + \omega_{jt}(s)\end{aligned}$$

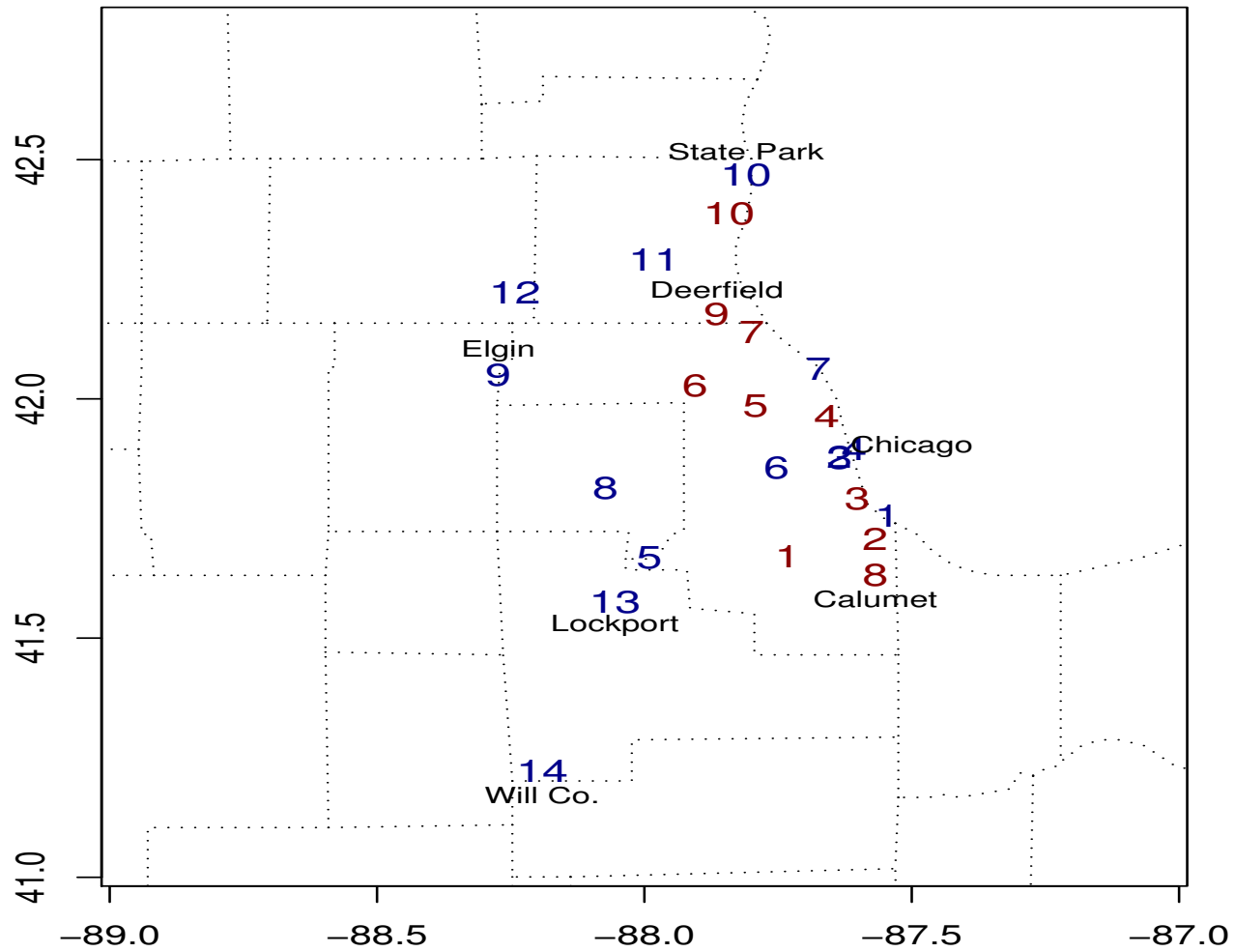
# DLM Approach

Model: 
$$Y_t(s) = Z_t(s)' \beta_t + S_{1t} \alpha_{1t}(s) + S_{2t} \alpha_{2t}(s) + \epsilon_t(s)$$

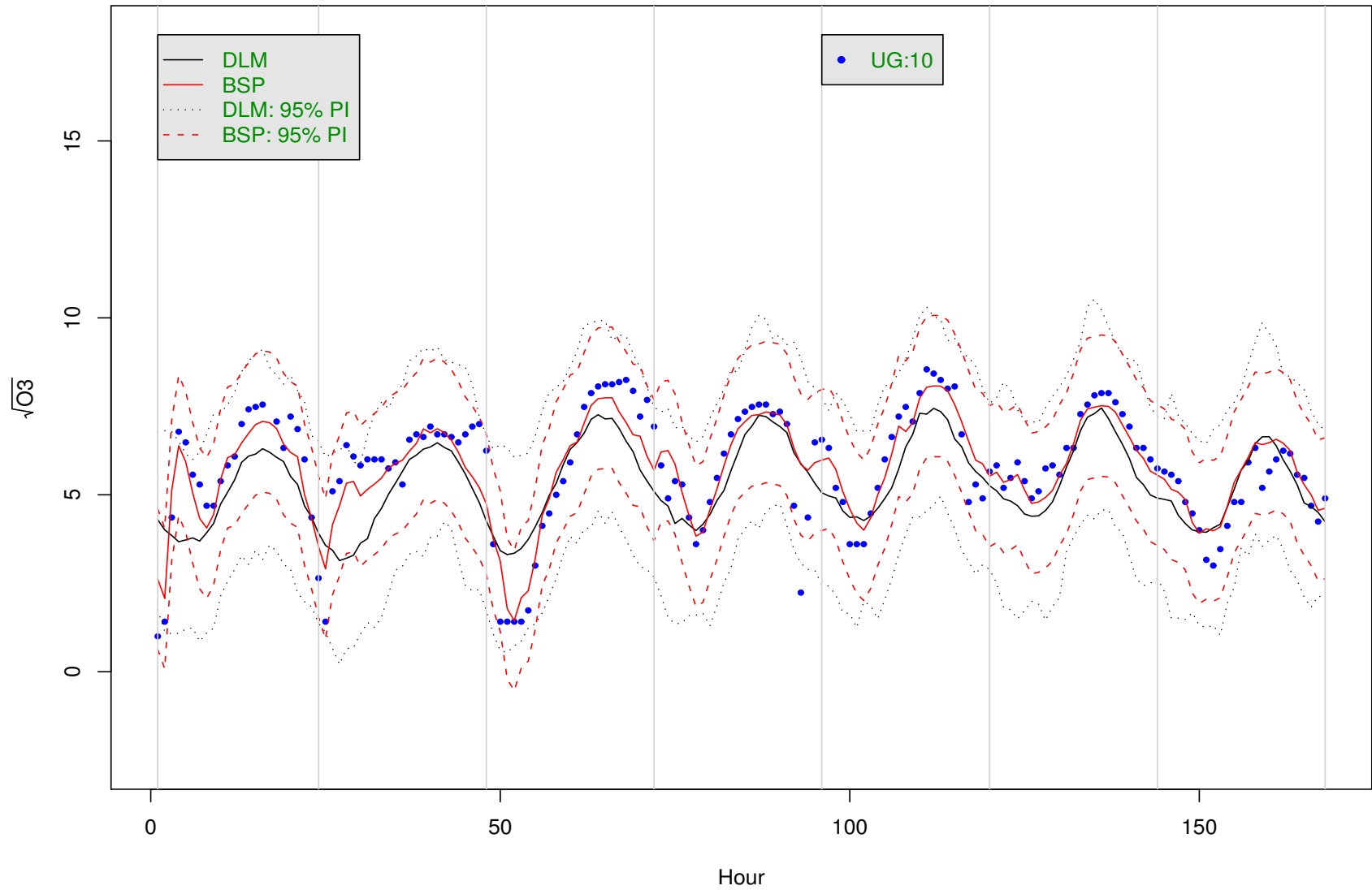
$$\beta_t = \beta_{t-1} + \omega_t, \quad \alpha_{jt}(s) = \alpha_{j,t-1}(s) + \omega_{jt}(s)$$

- $\text{Cov}(\omega_{jt}) = \sigma_y^2 \tau_j^2 \exp(-D/\lambda_j)$  smoothness across sites
- $\omega_t \sim N(0, \sigma_y^2 \tau_y^2)$
- **Very flexible: incorporate trend, space and time correlation (non-separable), etc directly via model parameters**
- **Build sub-models for meteorology, etc**
- Implement via MCMC - computationally expensive
- $\tau_j^2, \tau_y^2$  fixed in advance - trial & error (not easy!)  
results sensitive to these.

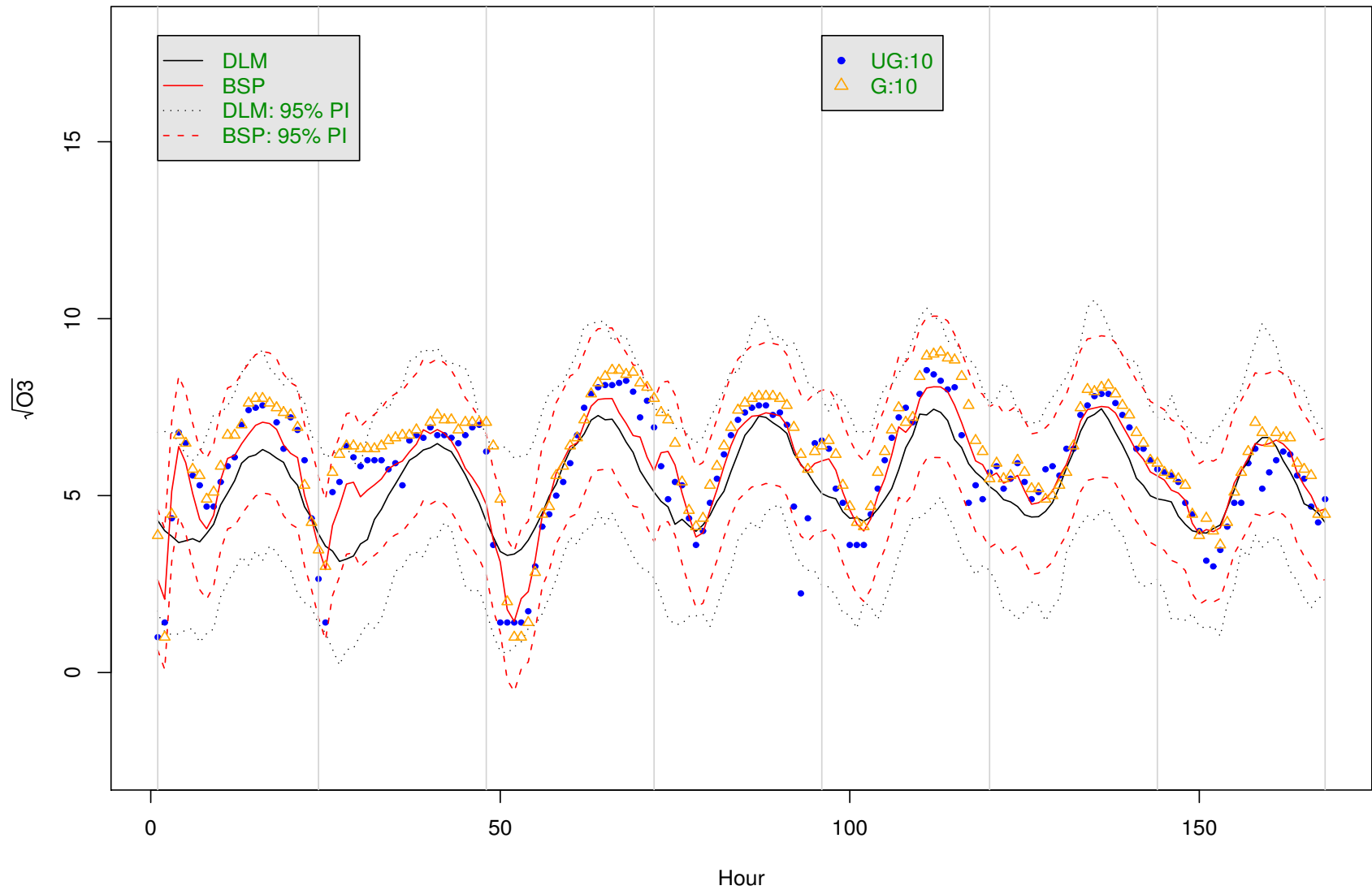
# Results



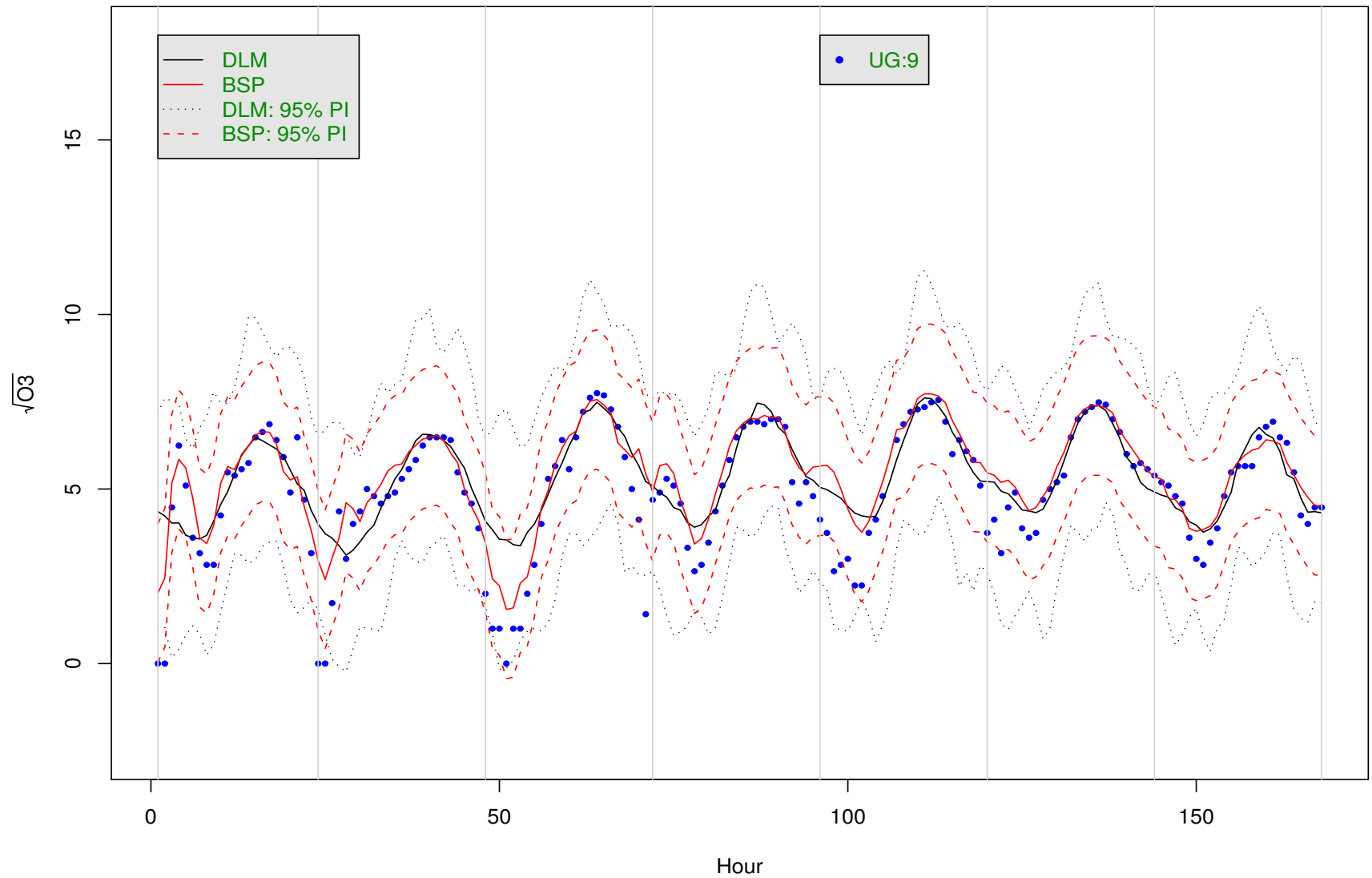
# Pred mean + 95% CI for ungauged site 10 - week 1



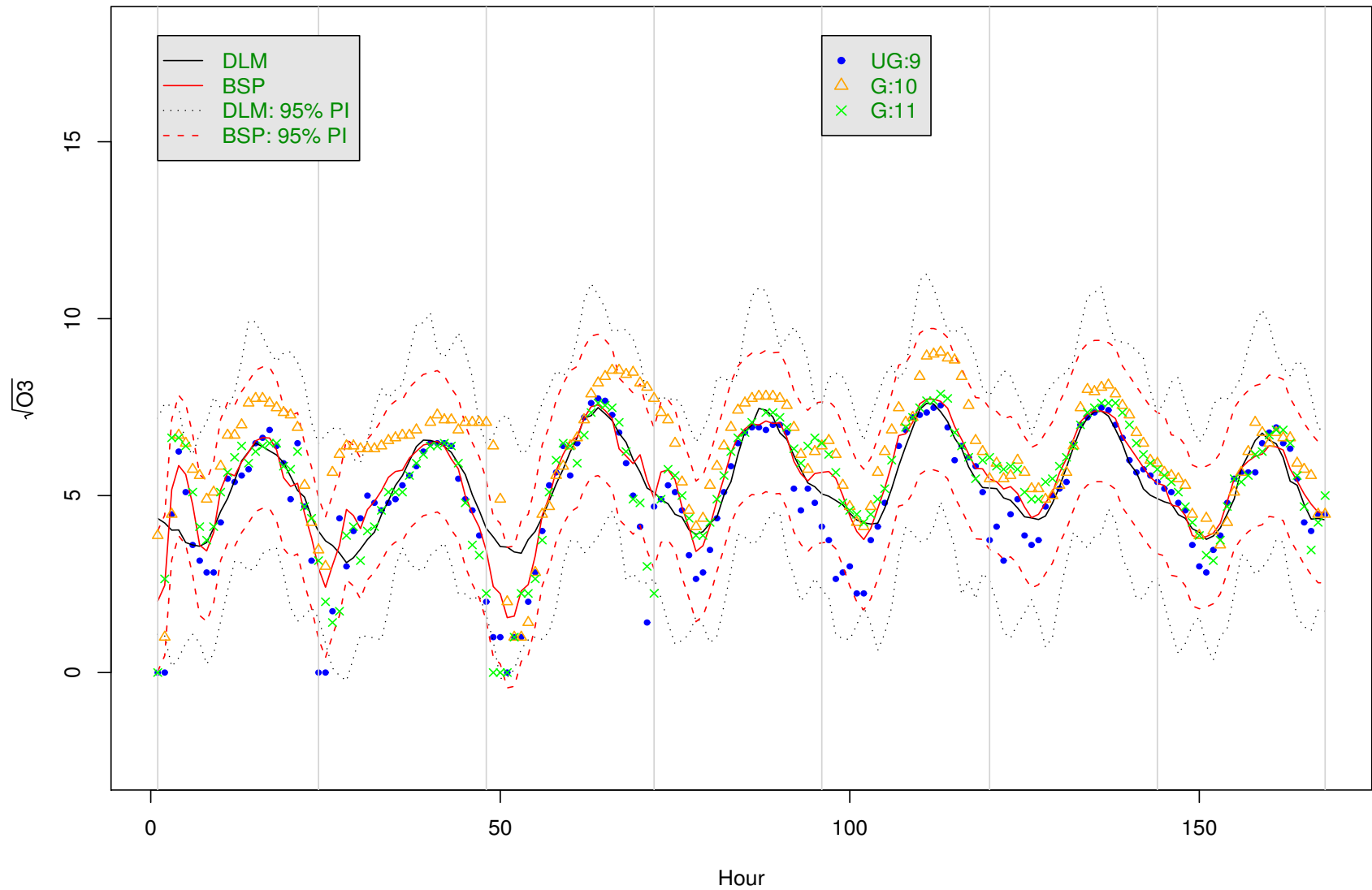
# Pred mean + 95% CI for ungauged site 10 - week 1



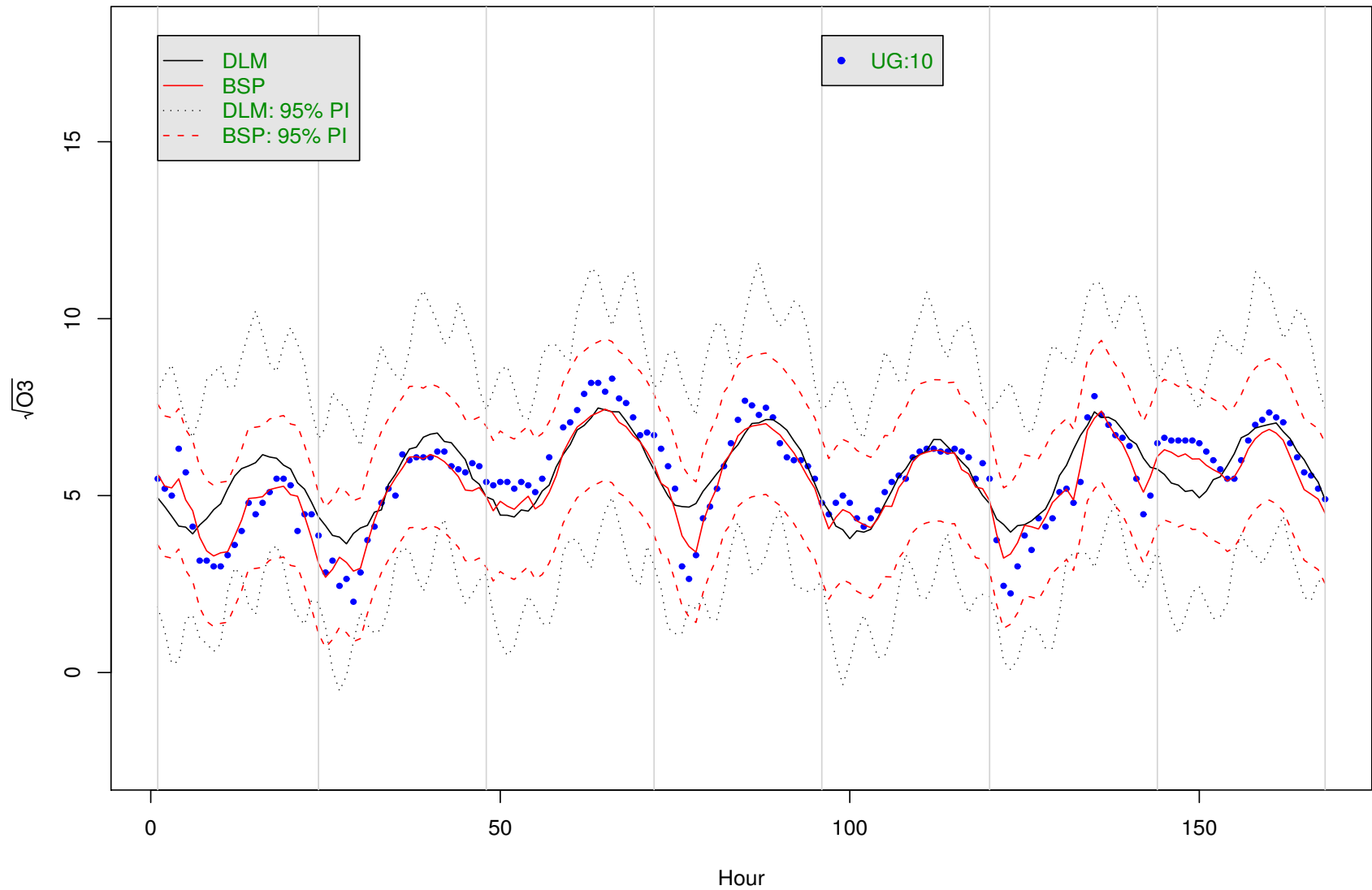
# Pred mean + 95% CI for ungauged site 9 - week 1



# Pred mean + 95% CI for ungauged site 9 - week 1

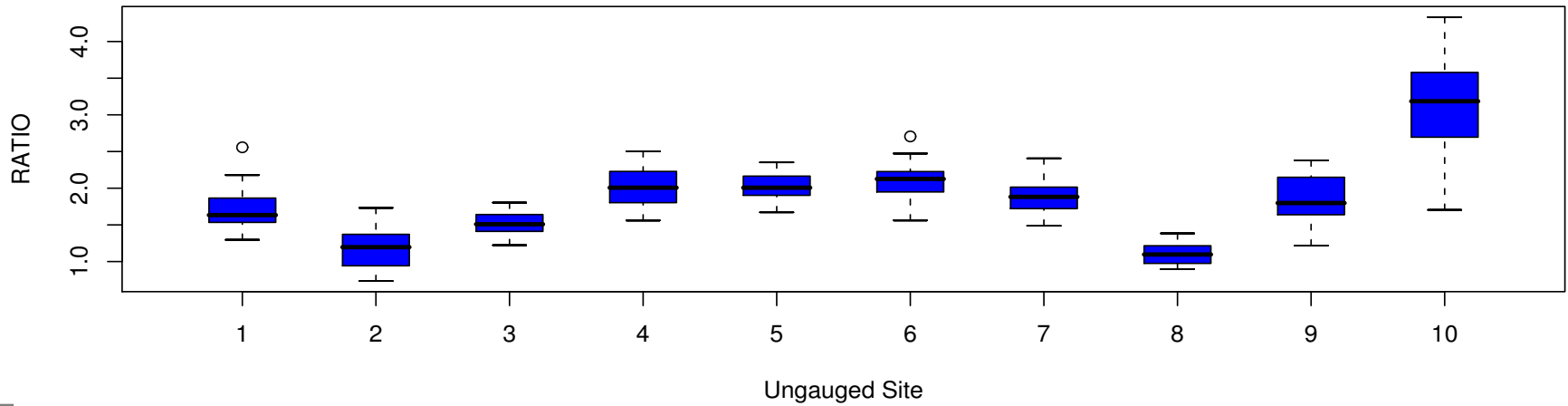
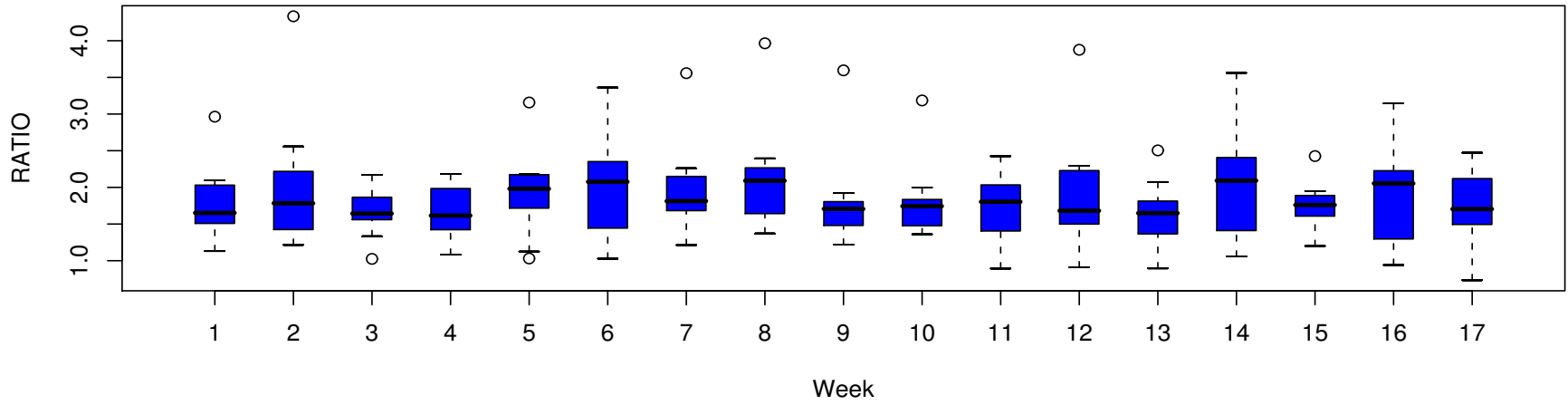


# Pred mean + 95% CI for ungauged site 10 - week 10

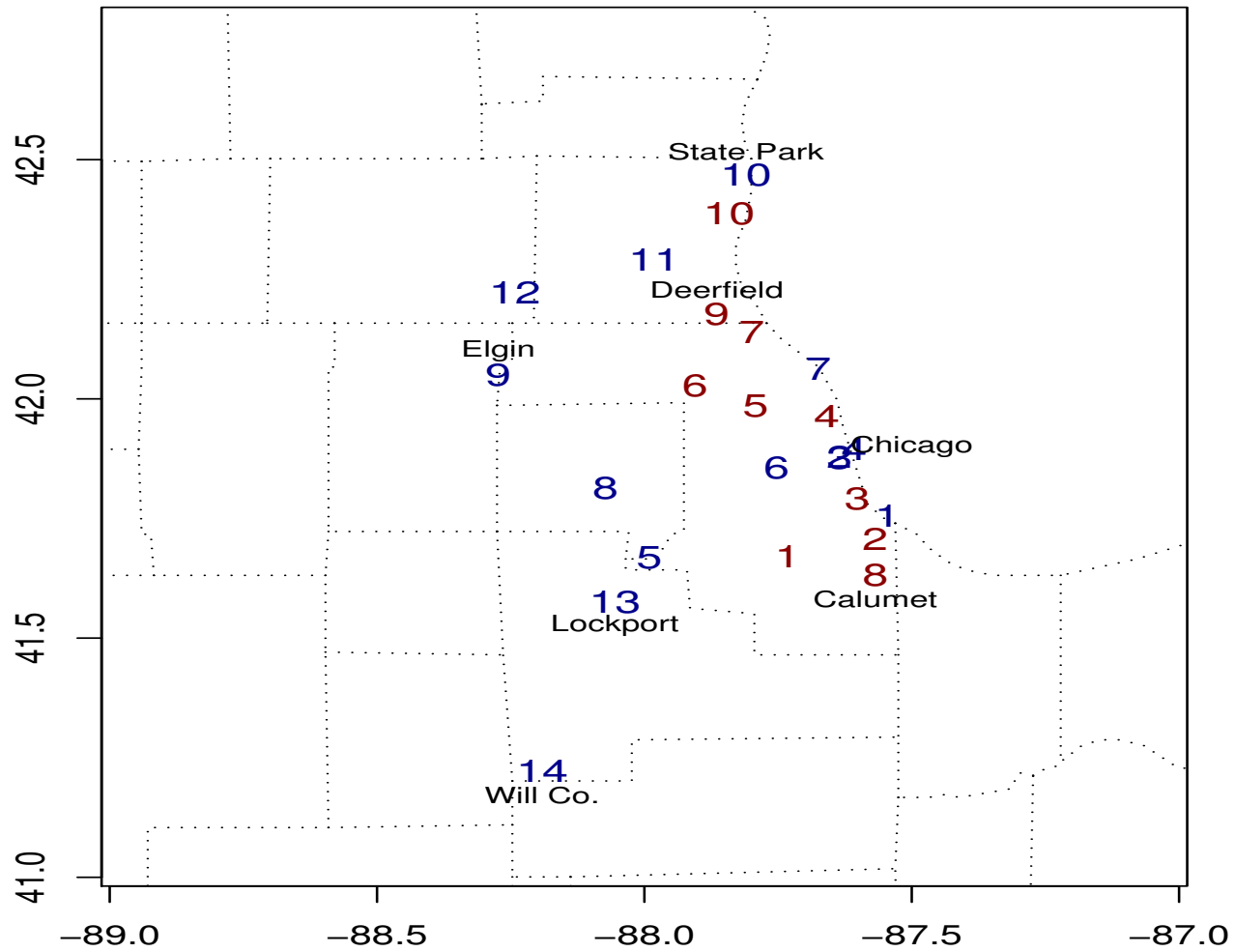


# Ratio of MSE's

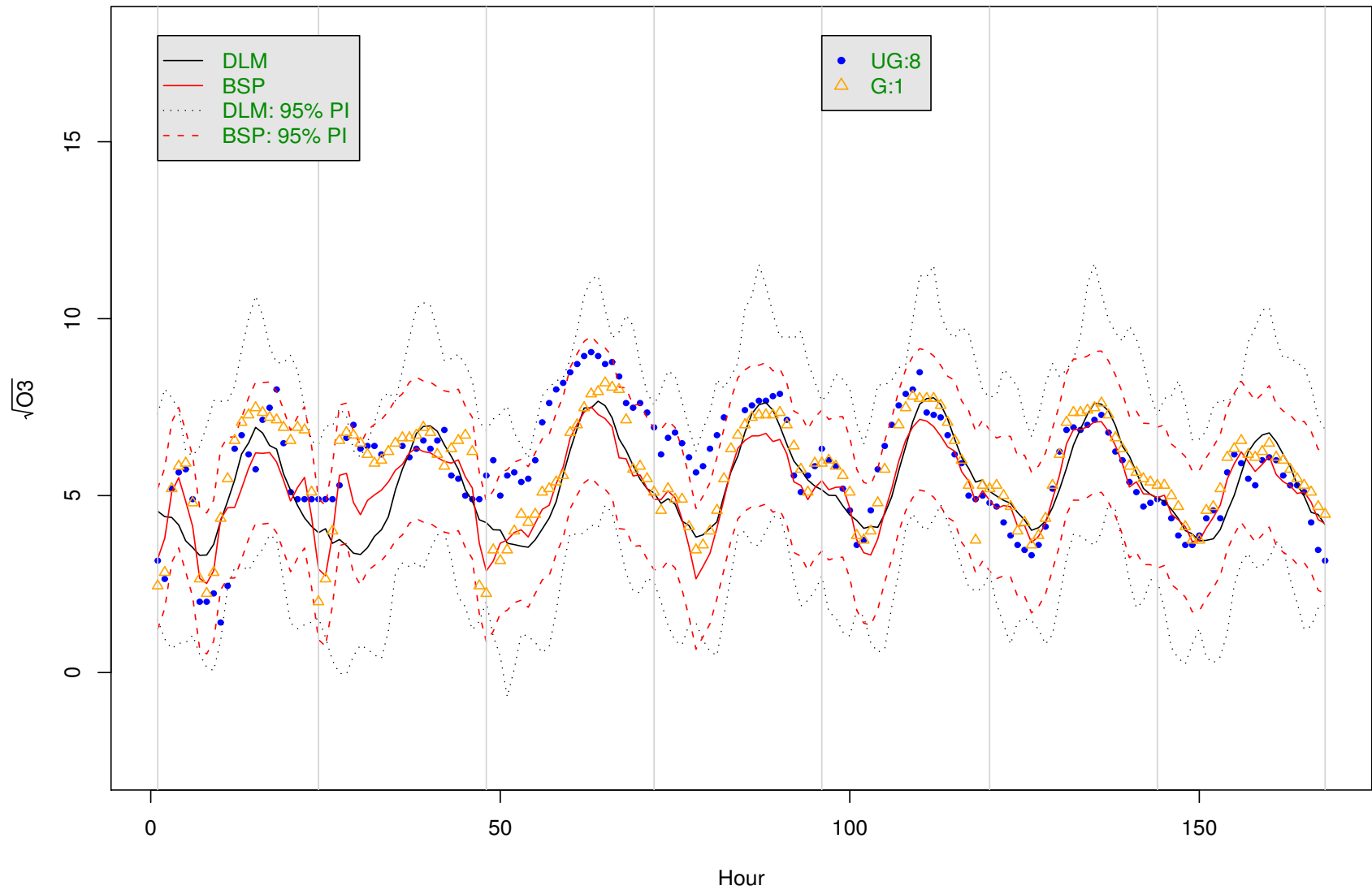
Boxplot of MSE(DLM)/MSE(BSP)



# Ungauged Location 8



# Pred mean + 95% CI for ungauged site 8 - week 1



# Confidence Interval Coverage

Nominal Level	DLM	BSP
95%	95	90
80%	88	79
60%	77	66

DLM:

•  $\tau_j^2, \tau_y^2$ : trial & error to get good coverage 95%CI

Ones - good for other levels - quite different

# Remarks

BSP:

- Perform reasonably well
- Handle multiple pollutants + hourly data
- Limitations
  - Time dependency
    - Adding a distribution on A
    - Implement via MCMC
  - Non-separability

# Remarks

DLM:

- Limited success
- Powerful and flexible but computationally expensive  
10 days to get 3000 MCMC replicates
- Limitations
  - Hyperparameter estimation  
method of moments (Wikle & Cressie, 1999)
  - Computational burden  
ensemble Kalman filter (Bengtsson et al. 2003)
  - Variance increasing with time  
discount factor (West & Harrison 1997)
  - Spatial correlation leakage?

# Remarks

Correlation leakage is an important issue in space-time modelling - not received much attention yet!

Software available at <http://enviro.stat.ubc.ca>

Tutorial in Chapter 14

Le & Zidek (2006). Statistical Analysis of environmental space-time processes (Springer)