

# **Design-based Empirical Orthogonal Functions**

**Breda Munoz-Hernandez**

**Virginia M. Lesser**

**Department of Statistics  
Oregon State University**

**This presentation was partially supported under STAR Research Assistance Agreement No. CR82-9096-01 awarded by the U.S. Environmental Protection Agency to Oregon State University. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and EPA does not endorse any products or commercial services mentioned in this publication.**

# Outline

- **Overview**
  - EOF model
  - **Design-based EOF model**
    - Main results
- **Combining information in the design-based EOF model**
- **Future research**

# **Empirical Orthogonal Functions Model (EOF)**

- **Alternative technique for modeling space-time stochastic processes**
  - **No spatial stationary conditions**
  - **Common analysis tool in atmospheric sciences:**
    - **Fit orthogonal functions**
    - **Reduction /essential features of the process**

# EOF Background

- Under regularity conditions a random process  $X(\mathbf{s}, t)$

$$X(\mathbf{s}, t) = \sum_{k=1}^{\infty} Z_k(t) \varphi_k(\mathbf{s})$$

**converge absolutely in mean square in a compact set**

**(Mercer's Theorem, (Adler, 1981))**

# EOF Background

- $\varphi_k(\mathbf{s})$  satisfies

$$\int_R C_X(\mathbf{s}, \mathbf{s}') \varphi_k(\mathbf{s}') d\mathbf{s}' = \lambda_k \varphi_k(\mathbf{s})$$

and

$$\int_R \varphi_k(\mathbf{s}) \varphi_j(\mathbf{s}) d\mathbf{s} = \delta_{kj}$$

# EOF Background

- $Z_k(t)$  is defined as

$$Z_k(t) = \int_R X(\mathbf{s}, t) \varphi_k(\mathbf{s}) d\mathbf{s}$$

$$\min \left( \int_D \left( X(\mathbf{s}, t) - \sum_{k=1}^N Z_k(t) \varphi_k(\mathbf{s}) \right)^2 \right)$$

# Comments in EOF

- **Finite set of network sites results in poor space/time coverage**
  - **Numerical instability when calculating components**
  - **Reduction in statistical confidence of estimates and possible misleading interpretation of the results**

# Design-based EOF

- Design-unbiased estimate of  $Z_k(t)$

$$\hat{Z}_k(t) = \sum_{s=1}^n \frac{X(\mathbf{s}, t)}{\pi(\mathbf{s})} \varphi_k(\mathbf{s})$$

- Estimates are generated as solutions:

$$\sum_{i=1}^n \frac{C_X(\mathbf{s}_i, \mathbf{s}_j)}{\pi(\mathbf{s}_i)} \varphi_k(\mathbf{s}_i) = \lambda_k \varphi_k(\mathbf{s}_j)$$

# Design-based EOF

- Final estimate:

$$\hat{X}(\mathbf{s}, t) = \sum_{k=1}^K \hat{Z}_k(t) \hat{\phi}_k(\mathbf{s})$$

- Interpolation non-observed site  $\mathbf{s}_0$ :

$$\hat{X}(\mathbf{s}_0, t) = \sum_{k=1}^K \hat{X}(\mathbf{s}_k, t) w(\mathbf{s}_0, \mathbf{s}_k)$$

# Similarity Coefficients

$$w(\mathbf{s}_0, \mathbf{s}_k) = \frac{r(\mathbf{s}_0, \mathbf{s}_k)}{\sum_{i=1}^n r(\mathbf{s}_0, \mathbf{s}_i)}$$

**Example of similarity coefficient constructed based on auxiliary variables  $y$ :**

$$r(\mathbf{s}_0, \mathbf{s}_i) = \frac{|y_j(\mathbf{s}_0) - y_j(\mathbf{s}_i)|}{\text{range}(y_j)}$$

# Fuzzy Similarity Coefficients

- Fuzzy membership function example

$$\mu: D \rightarrow [0,1]$$

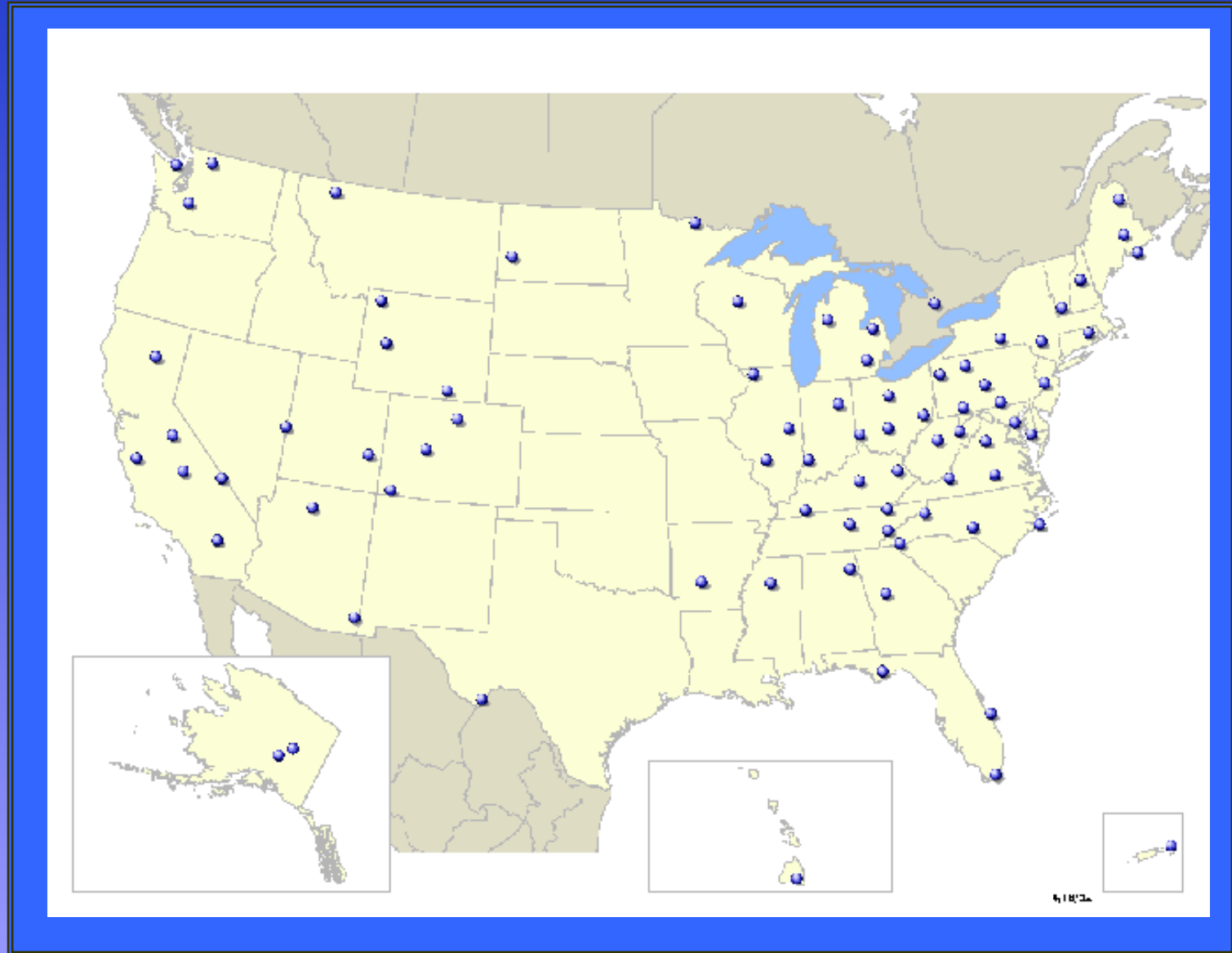
$$s \mapsto \frac{1}{1+a^b (s-c)^b}$$

- $\mu$  is selected based on the auxiliary variables (discrete, continuous)

# **Illustration Background**

- **Total acid dry deposition estimations from the Clean Air Status and Trends Network (CASTnet) (1987 - 1998).**
- **Dry deposition estimations are based on measured air pollutant concentrations and modeled dry deposition velocities estimated from meteorology, land use, and site characteristic data.**

# CASTnet site locations



# **MSE validation statistics**

## **Castnet dry deposition data**

### **(1987-1998)**

	<b>EOF classic 5</b>	<b>EOF classic 4</b>	<b>EOF Fuzzy 4</b>	<b>Kriging</b>
<b>Whole data</b>	<b>0.3043</b>	<b>0.3042</b>	<b>0.405</b>	<b>1.272</b>
<b>Eastern data</b>	<b>0.1594</b>	<b>0.1596</b>	<b>0.208</b>	<b>0.942</b>

# Current Research

- **Further model evaluation/comparison using simulations and different datasets**
- **Combining information from two different sources**
  - **Adjusting model to consider data from different sources**
  - **Evaluation using simulation and real data**

# Combining data

- Assume information from two sources:

- Site level:  $\{S_1, S_2, \dots, S_n\}$

- Region level  $\{D_1, D_2, \dots, D_m\}$

block average:

$$\{X(D_1, t), X(D_2, t), \dots, X(D_m, t)\}$$

# Combining data

$$\begin{aligned} \text{Cov}(X(\mathbf{s}_i, t), X(D_j, t)) &= \frac{1}{|D_j|} \int_{D_j} C_X(\mathbf{s}_i, \mathbf{s}) d\mathbf{s} \\ &\approx \frac{1}{|\hat{D}_j|} \sum_{k=1}^m \frac{C_X(\mathbf{s}_i, \mathbf{s}_k)}{\pi(\mathbf{s}_k)} \end{aligned}$$

where  $|\hat{D}_j| = \sum_{k=1}^m \frac{I(\mathbf{s}_k \in D_j)}{\pi(\mathbf{s}_k)}$

# Future Research

- **Evaluation of monitoring networks**
  - Are additional sites providing new information?
  - Evaluate other sampling designs (e.g. number and location of sites)
- **Incorporation of seasonal variability**