High-Dimensional Manifold Geostatistics

Chintan Dalal

Department of Computer Science, Rutgers University

18th December, 2017

Thesis Committee:

Prof. Dimitris Metaxas (Chair)
Prof. Vladimir Pavlovic
Prof. Kostas Bekris
Prof. Jonathan Stroud (Georgetown University)
Prof. Douglas Nychka (National Center for Atmospheric Research, University of Colorado Boulder)

Sea-Level Changes

Observed Sea-Level Trends

153 million lives affected by 2100

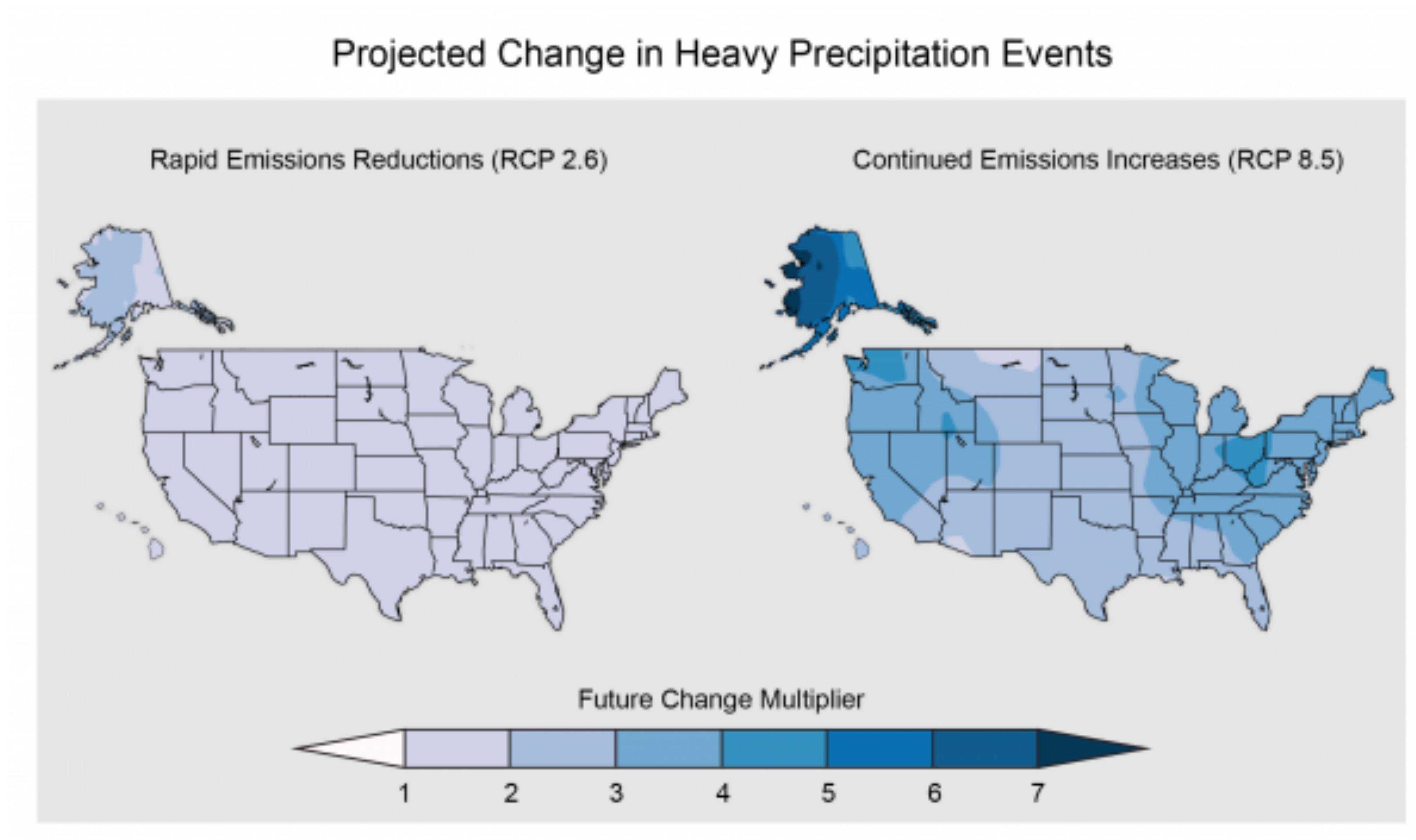
Scientific Goal:

Improve the estimates of sea-level trends using:

1. Spatial information
2. Multiple sources of datasets

Kopp et al 2017;
National Climate Assessment (2014)

Precipitation Changes



Scientific Goal:

1. Compare Existing Earth System Models
2. Emulate Future Climate Scenarios

National Climate Assessment (2014)

Climate Data Science

Observations
1. Gauges
2. Satellite
3. …

Process/ Clean Data Sets

Explore the Problem & Data
1. Climate Networks
2. Causality
3. …
4. Our Approach: High-Dimensional Manifold

Models & Algorithms
1. For Estimation:
   1. …
   2. Our focus: Non-stationary Model
2. For Simulation
   1. …
   2. Our focus: Multi-variate Normal

Data - products
1. Earth System Models
2. …

Communicate Visualize, Reports
1. IPCC reports
2. …

Make Decision

Real

Our Approach

Develop geostatistical models

by exploring

high-dimensional geometric structures

on a manifold

Talk outline

- **Scientific Goal 1**: Improve estimates of sea-level trends using spatial information
  - Regression Models
  - Simulation Study
  - Sea-Level Datasets - Spatial (Dim = 2), Spatial-temporal (Dim = 3)
- **Scientific Goal 2**: Inference from multiple sources of datasets
  - Data-Fusion Model
  - Multiple Sea-Level Datasets (Spatial, Spatial-temporal)
- **Scientific Goal 3**: Inter-comparison of Earth System Models (Dim > 3)
- **Scientific Goal 4**: Emulate future climate scenarios (Dim > 3)

Talk outline

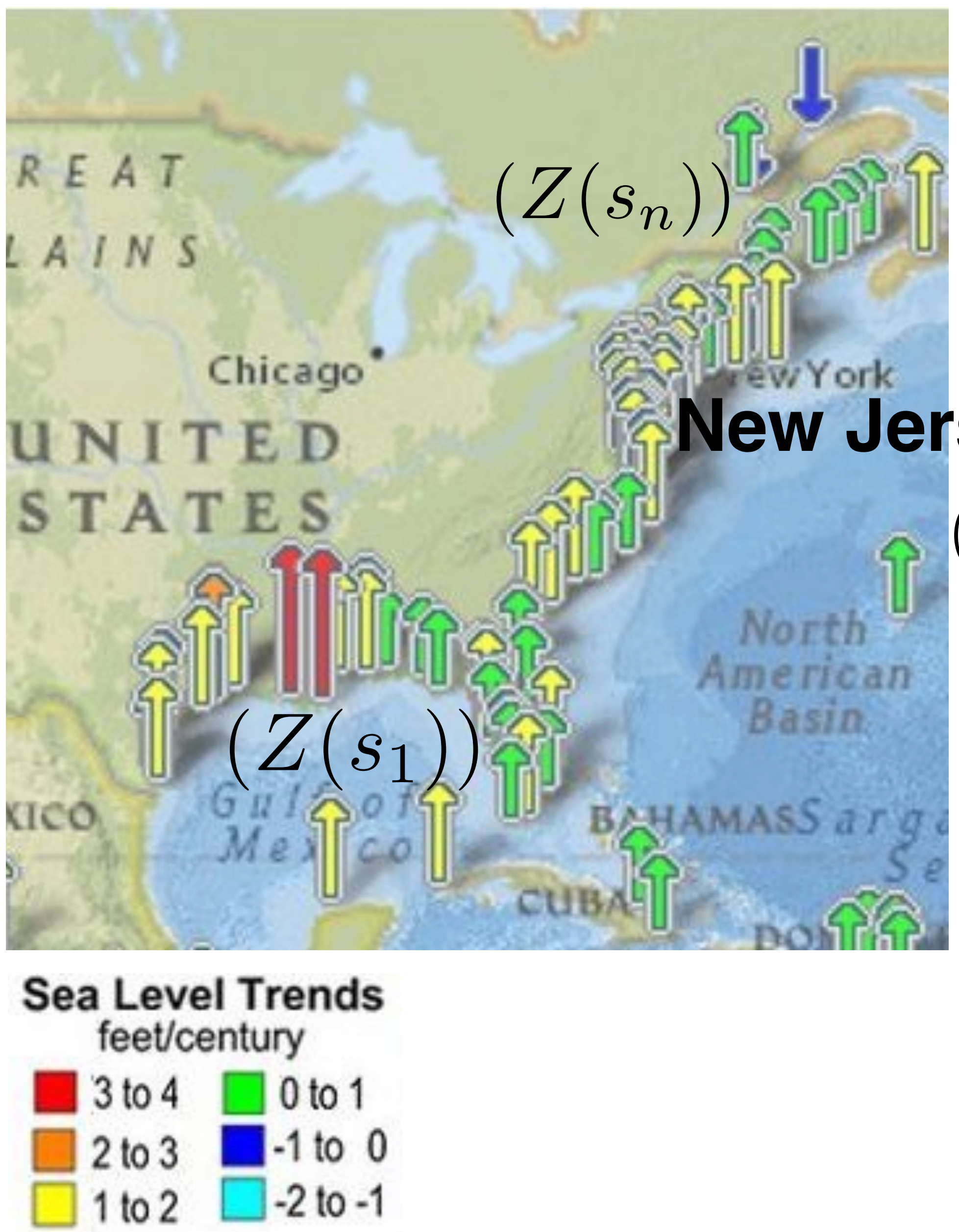
- Scientific Goal 1: Improve estimates of sea-level trends using spatial information
  - **Regression Models**
  - Simulation Study
  - Sea-Level Datasets - Spatial & Spatial-temporal
- Scientific Goal 2: Inference from multiple sources of datasets

- Scientific Goal 3: Inter-comparison of Earth System Models

- Scientific Goal 4: Emulate future climate scenarios

Geostatistics - Estimation

To model the stochastic process $\{Z(s) : s \in G \subset R^d\}$



$$(Z(s^*)) = k_1 Z(s_1) + \cdots + k_n Z(s_n)$$

Covariance:

$$K(s, s') = Cov\{Z(s), Z(s')\}$$

Regression Model

$$Z(s) = \mu(s) + Y(s) + \epsilon(s)$$

$$\mu(s) = E[Z(s)] \qquad \epsilon(s) \sim \mathcal{N}(0, \tau^2)$$

Gaussian process model: $(Y(s_1), \ldots, Y(s_n))' = \mathbf{Y} \sim G(0, K)$

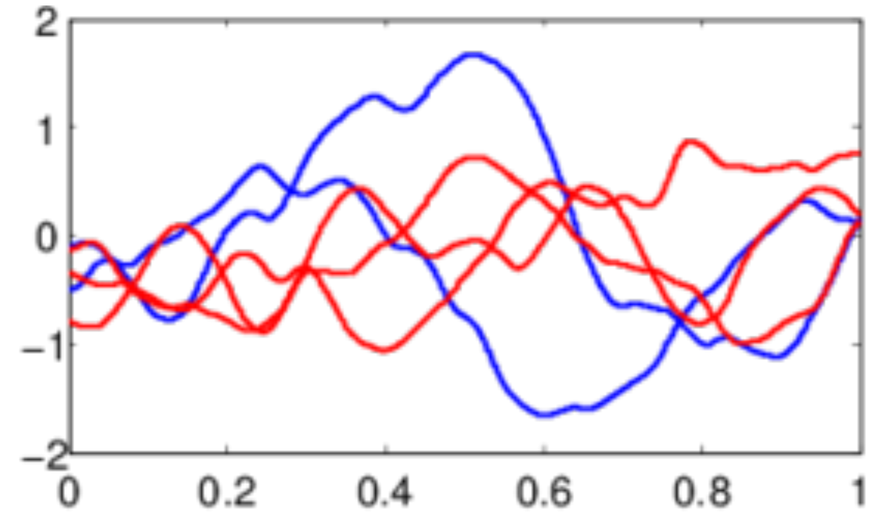
Key ingredient $K(s, s') = Cov\{Z(s), Z(s')\}$

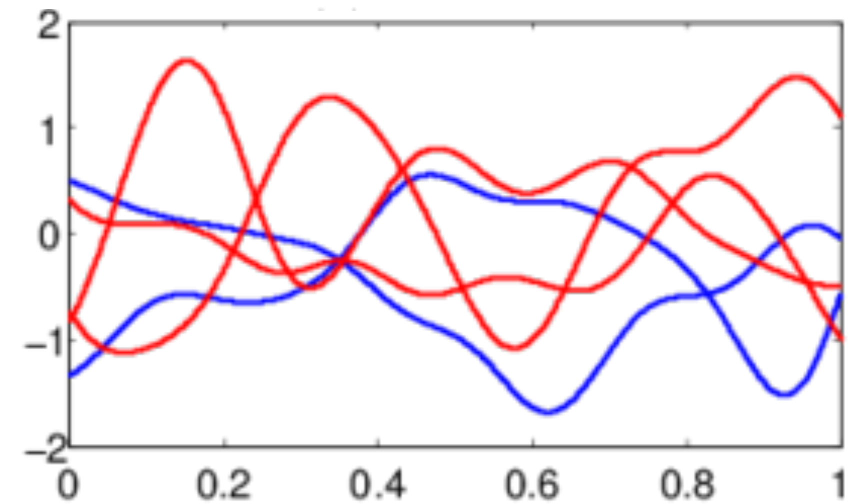
Covariance Function

$$k_{ij} \propto \mathcal{H}(Q_{ij})$$

Correlation Function: $\mathcal{H}$

E.g.
Matern function ($\mathcal{M}_\nu$) ,
Squared Exponential ($\nu \to \infty$),
etc

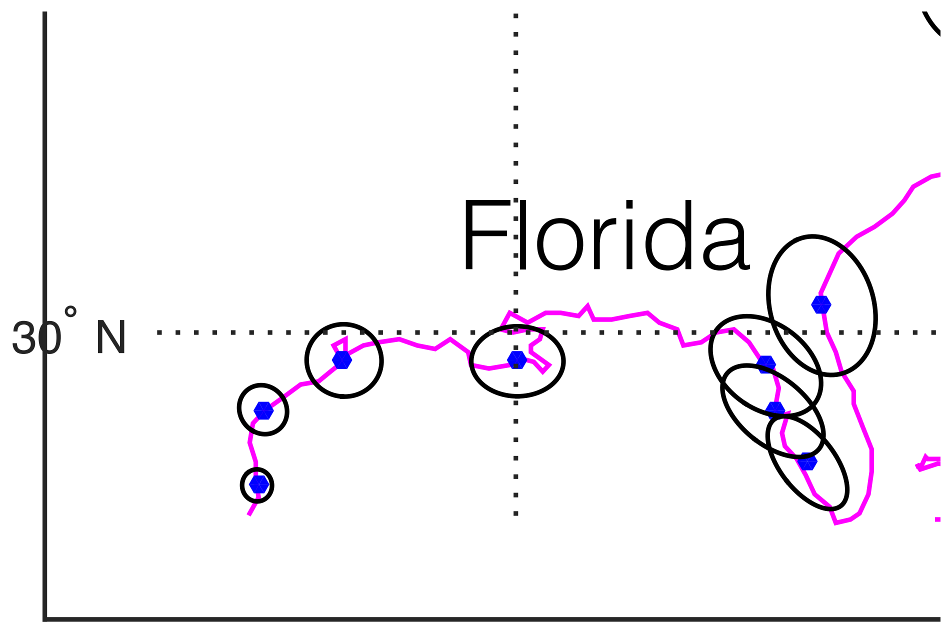


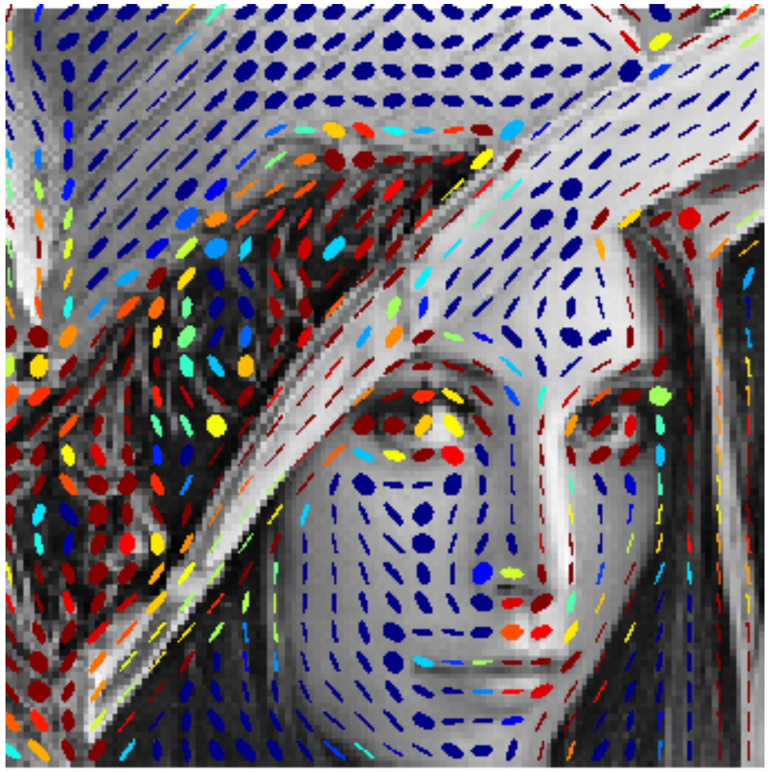


Scaled Distance Length: $Q_{ij} = \Delta(s_i, s_j)\Sigma^{-1}\Delta(s_i, s_j)$

Geometric Anisotropy (Scale): Σ

Geometric Anisotropy (Scale)

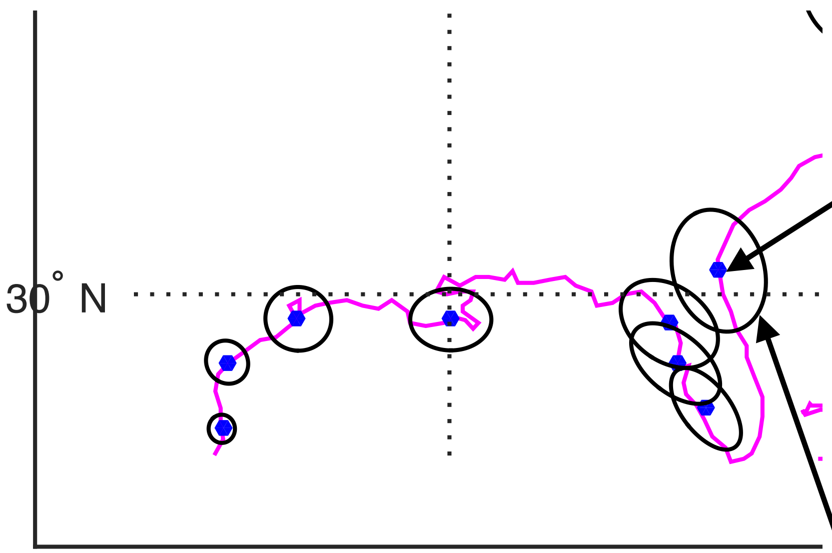




Geostatistics:
Rate of decay
at various geo-locations

Vision:
Texture
at various locations

High-Dimensional Geometric Structure



Points in high-dimension: $s_i \in \mathcal{R}^d$

$$x(s_i) = (x_1(s_i), \ldots, x_m(s_i)) \in \mathcal{R}^m$$

E.g., {latitude, longitude, elevation, time}

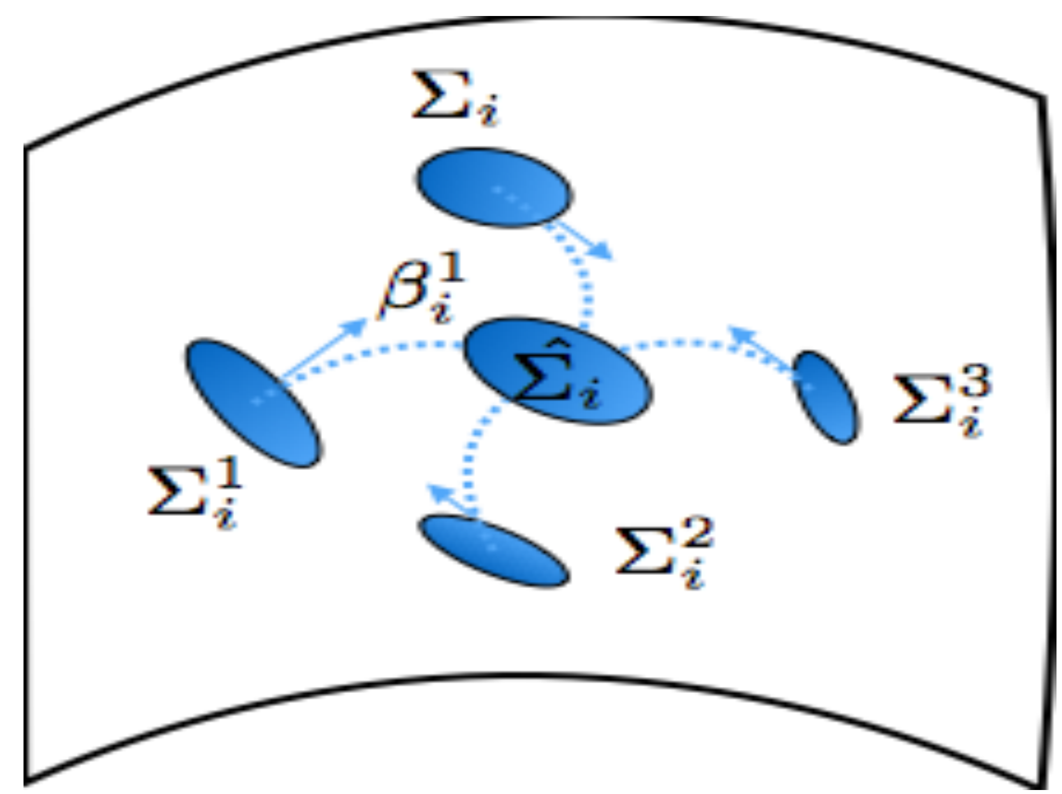
$\boxed{\text{High-Dimensional Geometric Structure}}$

High-Dimensional Manifold

Symmetric Positive Definite (m)

$$\Sigma(\cdot) = \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \\ l_{m1} & & l_{mm} \end{bmatrix}$$

Manifold of $\Sigma_i \in SPD(m)$



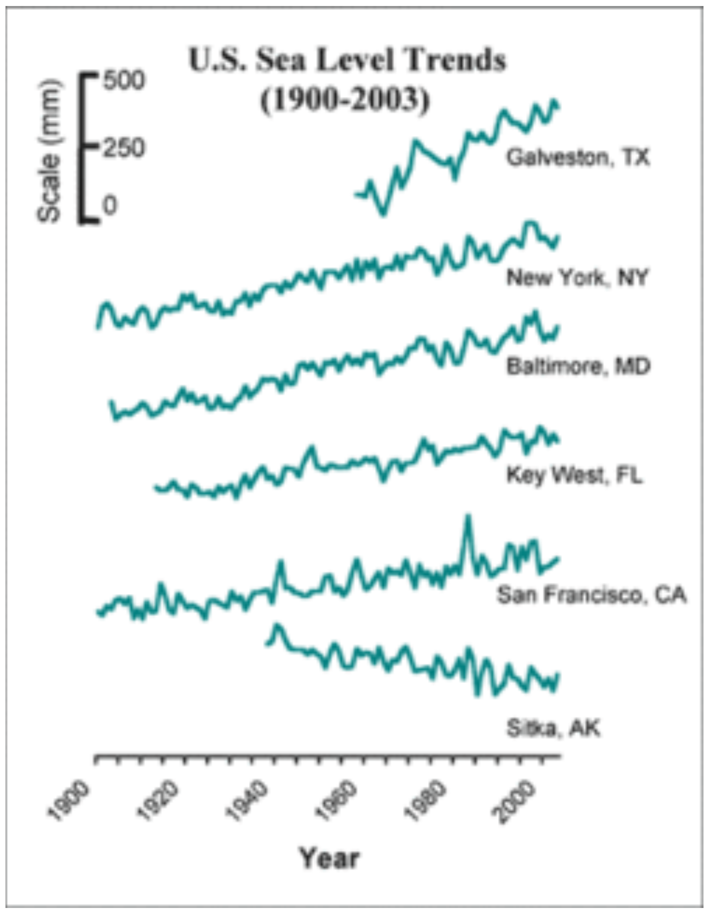
$\hat{\Sigma}_i \leftarrow$ Mean estimate

$\beta_i^1 \leftarrow$ Direction to $\hat{\Sigma}_i$

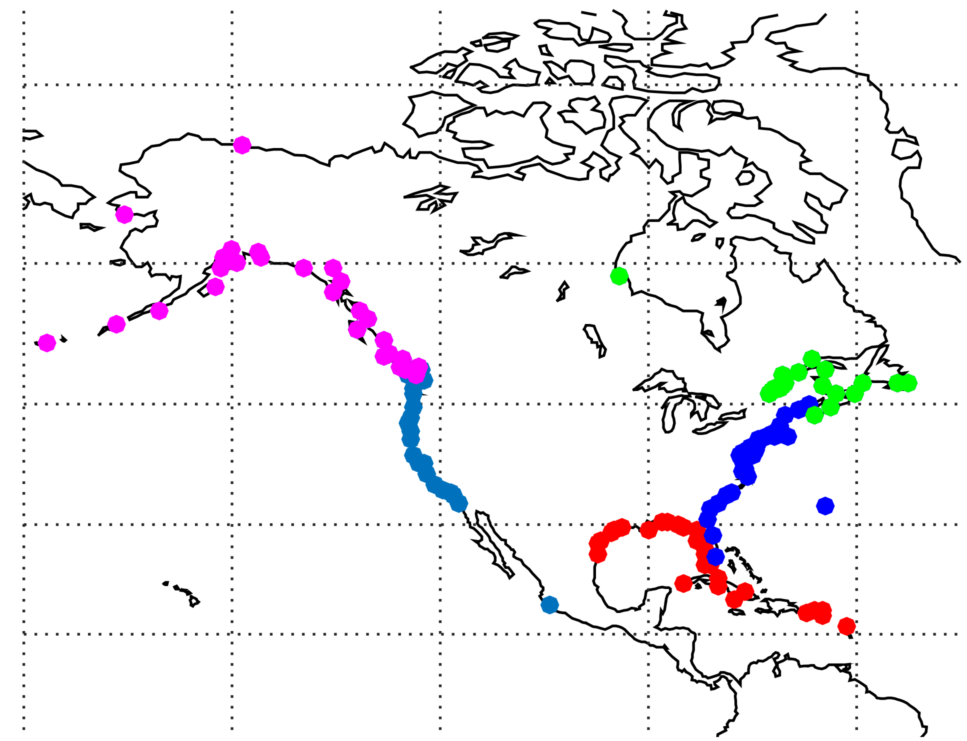
Sea Level Changes

Climate Scientist Approach:

Step 1:



Step 2:



Geophysical Clusters of Stations
[Kopp et al., Nature (2013); Hay et al., Nature (2015)]

Physical understanding does not mean we know how likely it is

Goal: Provide a systematic approach to incorporate spatial structures for estimating local sea level changes.

Spatial Models

Deformation Models:
Sampson & Guttorp (1992), Anderes & Stein (2008)

Processes on a Sphere:
Jun & Stein (2008)

Spatially Varying Model:
Hidgon (1999), Paciorek (2005), Riser (2014)

Weighted Average Models: Fuentes (2002)

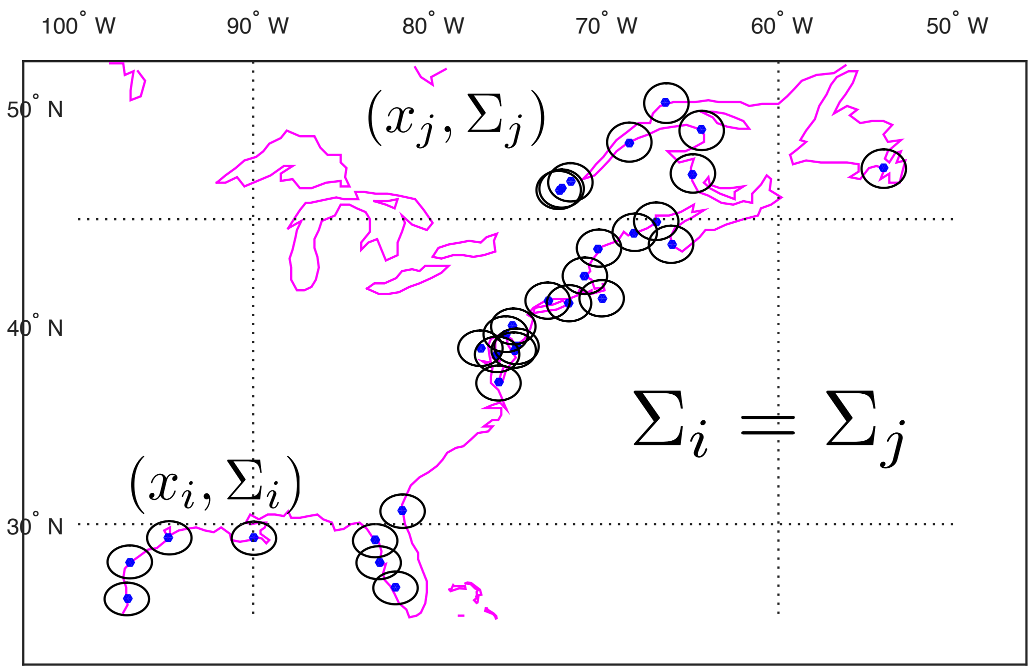
Basis Function Expansion: Holland et al. (1998), Nychka et al. (2002), Matsuo et al. (2011), Katzfuss (2013)

Models for Large Geo-statistical Datasets:
Lattice Kriging (Nychka et al. (2014)), Bayesian Nearest Neighbors Geostatistics (Banerjee (2016)), Stochastic PDE - INLA (Lindgren et al. (2011))

Spatially Varying Covariance Function

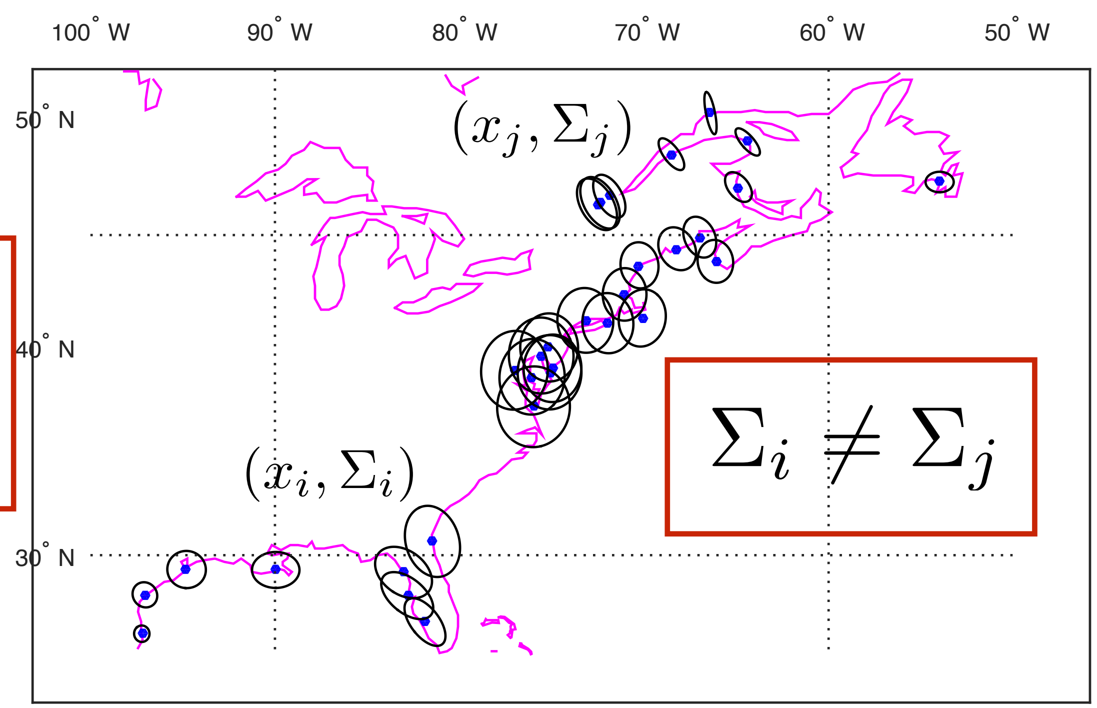
Stationary Model

$$Q_{ij} = (x_i - x_j)^T \Sigma^{-1} (x_i - x_j)$$



Non-stationary Model

$$\boxed{Q_{ij} = (x_i - x_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (x_i - x_j)}$$



Geo-regional Sea-Level Changes

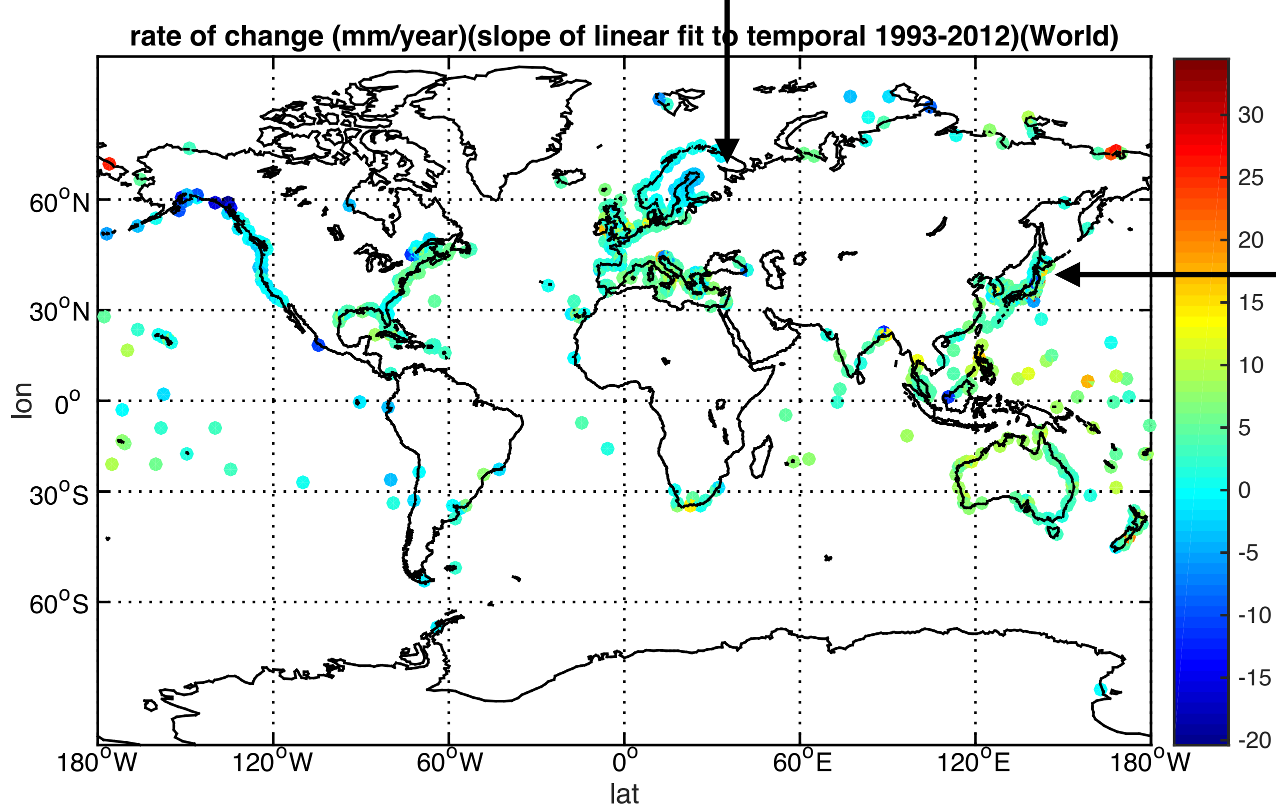
Complex spatial patterns results from ocean dynamical processes, movements of the sea floor, and changes in gravity due to water mass redistribution in the climate system.

$Z(s_{\text{Northern Europe}})$

Driven by melting of glaciers

$Z(s_{\text{Japan}})$

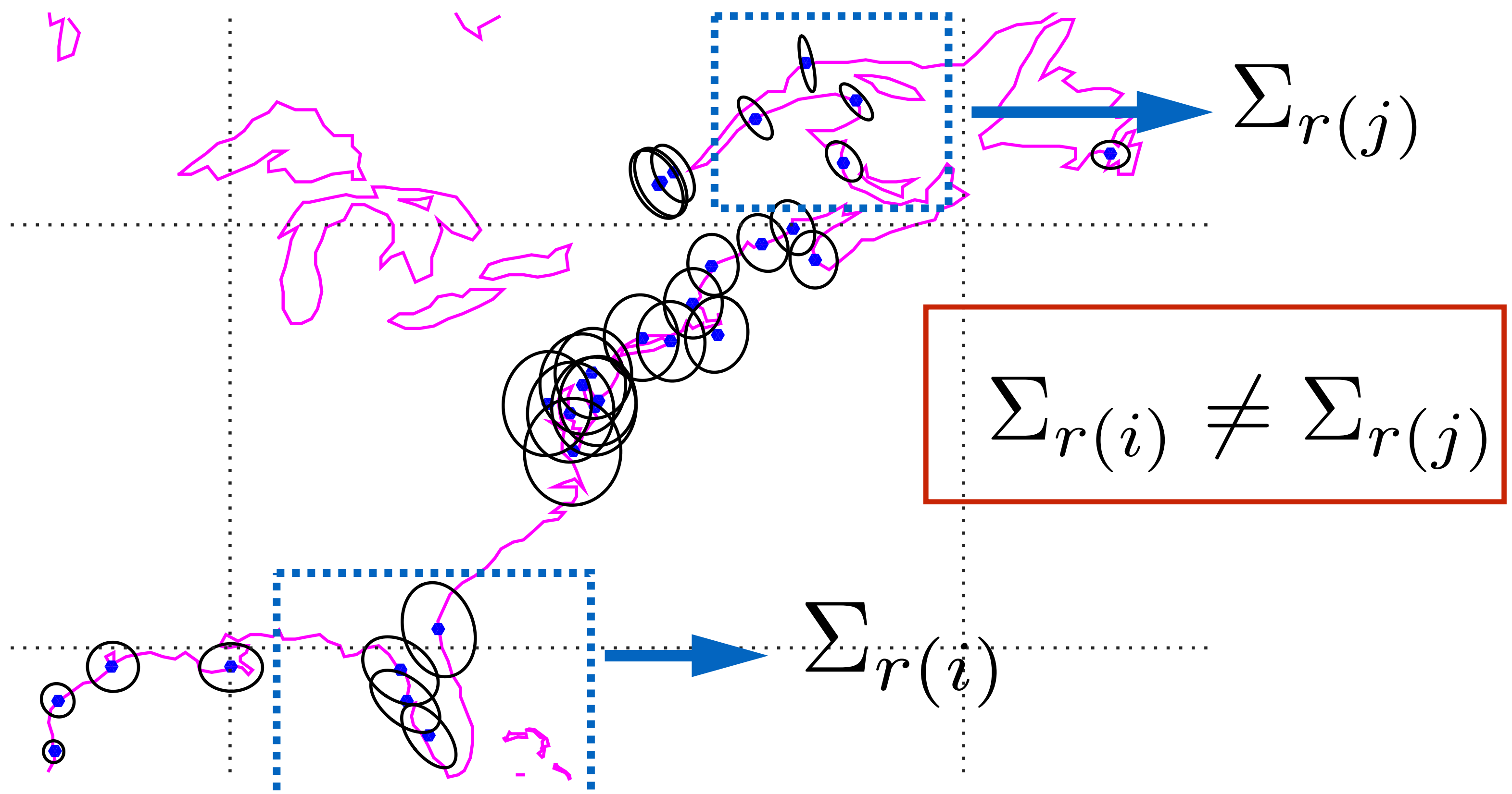
Driven by tectonics



IPCC (2014)

Proposed Covariance Function

Geometric anisotropy now incorporates regional structure of the geolocations

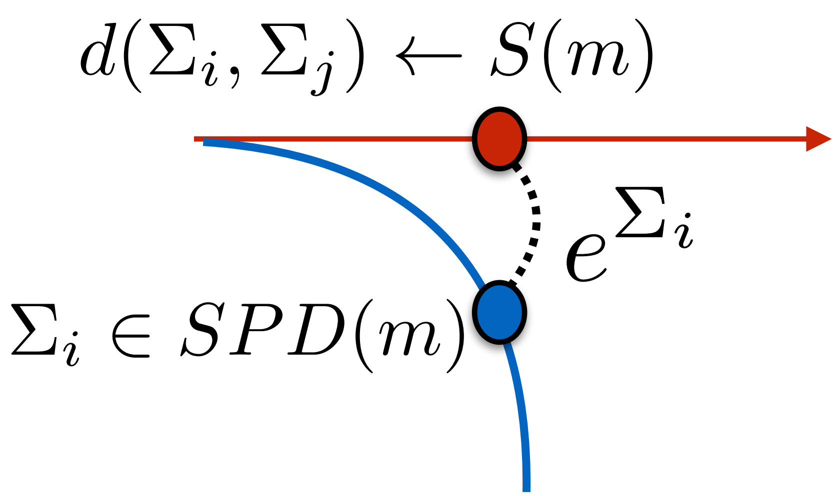


$$Q_{ij} = (x_i - x_j)^T \left(\frac{\Sigma_{r(i)} + \Sigma_{r(j)}}{2} \right)^{-1} (x_i - x_j)$$

Dalal et al. (2014); Dalal et al. (2015) (Climate Informatics, **Best paper award)**

Proposed Covariance Function Construction

Choose a metric and a distance function for $\{\Sigma(s) \in SPD(m)\}$



Affine-invariant Metric

$d(\Sigma_i, \Sigma_j) \leftarrow$ Rao's Riemannian distance

Sample or Estimate: $\Sigma_{r(i)}$

Theorem 1: The proposed covariance function is a valid non-stationary covariance function.

Dalal et al. (2014); Dalal et al. (2015)
(Climate Informatics, **Best paper award**)

Parameter Estimation

Markov Chain Monte Carlo scheme

1. For geometric anisotropy:
   1. Sample from $\Sigma_{r(i)} \sim \mathcal{N}(\hat{\Sigma}, \hat{\Lambda})$

      $\hat{\Sigma} \leftarrow \mathrm{argmin}_{\hat{\Sigma}} \sum d^2(\hat{\Sigma}, \Sigma_i^k)$

      $\hat{\Lambda} = \sum_k \beta_i^k (\beta_i^k)^T$
   2. Estimate for k-nearest neighbors:
      1. Profile likelihood
      2. Proposal range is a measure of dispersion $|\hat{\Lambda}| < 1$ in $\{\Sigma\}_{i \in \mathrm{neighbors}}$
2. For smoothness ν : Use profile likelihood and discrete priors {0.5,1,…,5.5}
3. Other parameters (σ_f, σ_n): Jointly propose from their conjugate priors

$$k_{ij} = \sigma_f^2 \mathcal{H}(Q_{ij}) + \sigma_n^2$$

Dalal et al. (2014);
Dalal et al. (2015);
Dalal et al. (2017) [In prep]

Model Improvements

1. Reduced the parameter space & Scalable to higher dimension:

    1. Previous work (NSGP): $\{s_i\}_n \to \{\Sigma_i\}_n$

parameterization of scale

GP priors for the scale parameters

$$\boxed{(2m-1+\frac{m(m-1)}{2})\cdot n} + \boxed{(2m-1+\frac{m(m-1)}{2})\cdot(m+\frac{m(m-1)}{2})\cdot 3}$$

For m=3: 8n + 144

After model simplification: 8n + 12

2. Our Approach (proposedNSGP): $\{s_i\}_n \to \{\Sigma_i\}_k$, $k << n$

Sample Scale Matrix

k-neighbors

$$\boxed{(2m+\frac{m(m-1)}{2})\cdot k} \boxed{+\ 1}$$

For m=3: 9k + 1

Dalal et al. (2014); Dalal et al. (2015)
(Climate Informatics, **Best paper award**)

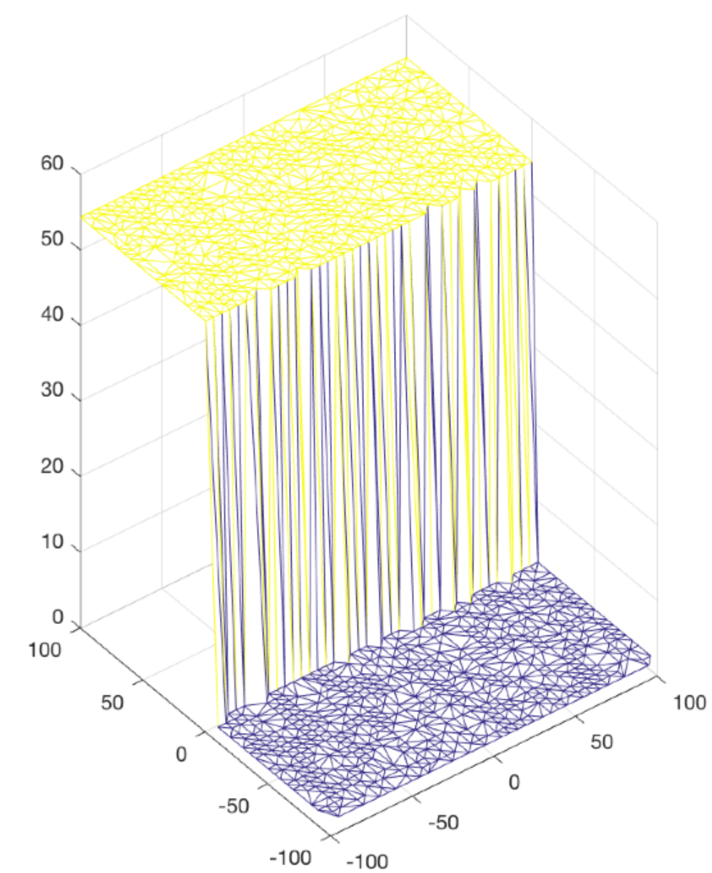
Model Improvements

2. Faster mixing (convergence) in MCMC scheme

3. Flexibility in the smoothness of geometric anisotropy:

    1. NSGP: Not suitable for jump datasets (discontinues process parameters)

    2. Proposed NSGP: Flexible for jump datasets



Dalal et al. (2014); Dalal et al. (2015)

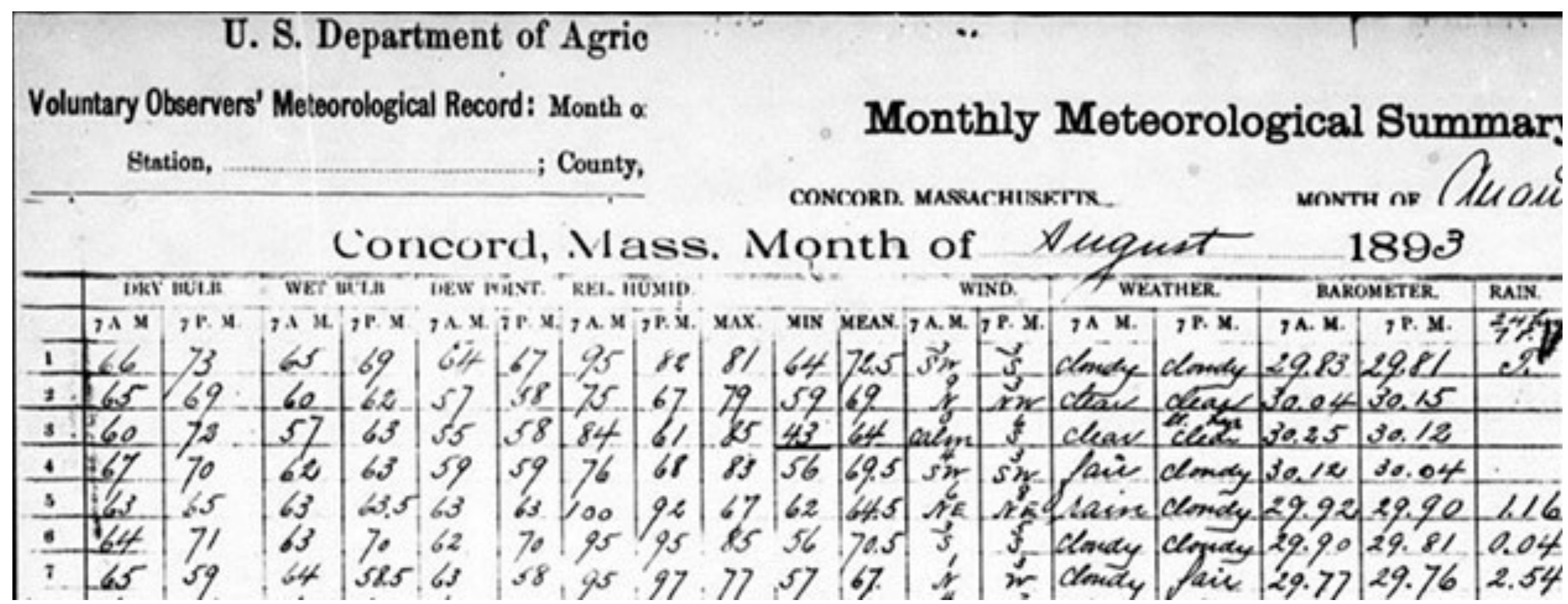
Talk outline

- Scientific Goal 1: Improve estimates of sea-level trends using spatial information
  - Regression Models
  - **Simulation Study**
  - Sea-Level Datasets - Spatial & Spatial-temporal
- Scientific Goal 2: Inference from multiple sources of datasets
- Scientific Goal 3: Inter-comparison of Earth System Models
- Scientific Goal 4: Emulate future climate scenarios

Simulation Studies

Importance of a good simulation study in Climate Data Science

Data — Cat in Images
[Google's Deep Dream Project (2016)]

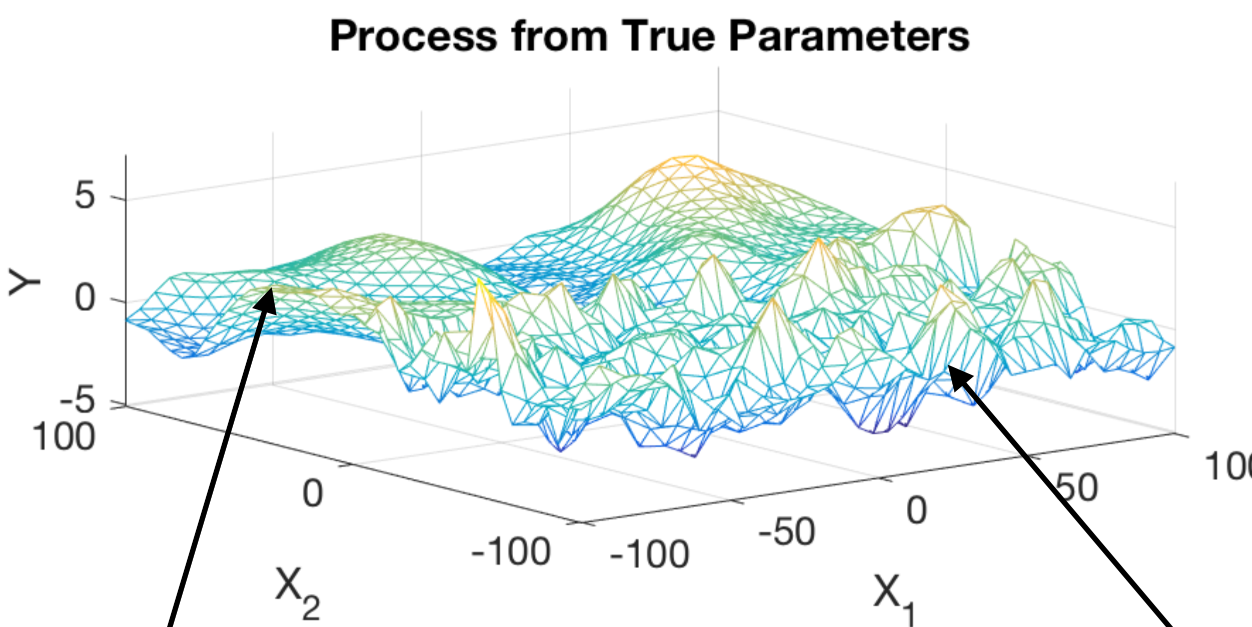


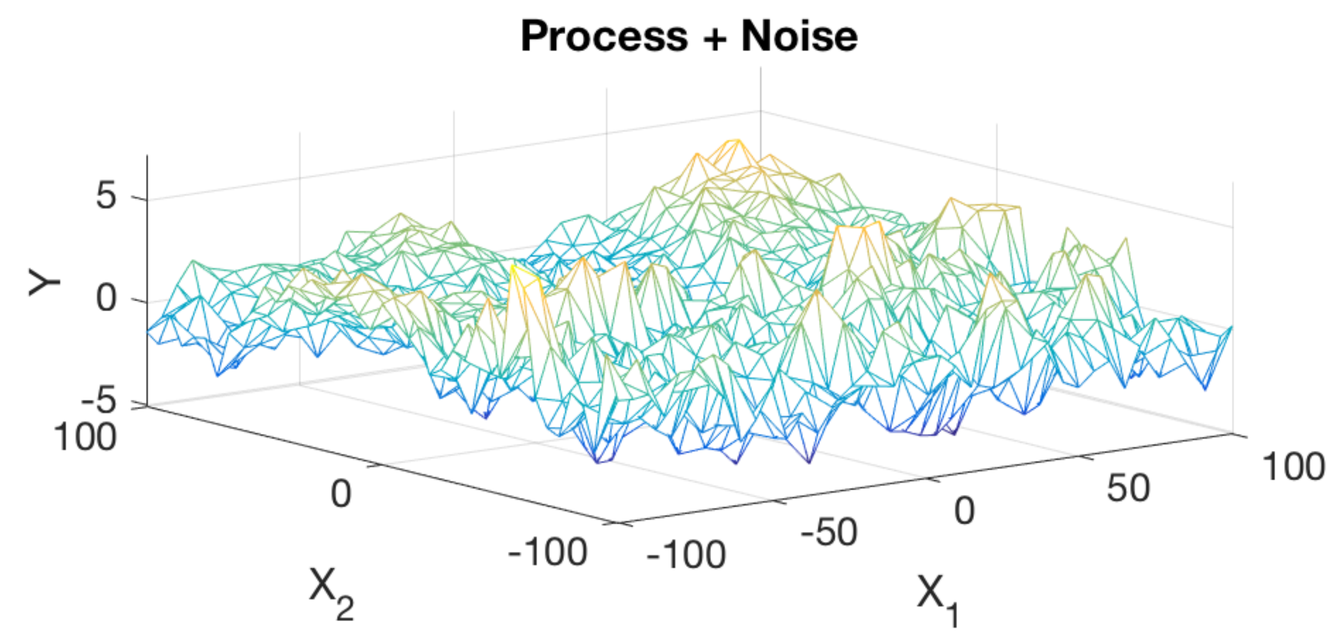
U. S. Department of Agric
Voluntary Observers' Meteorological Record: Month o
Station, ; County,
Monthly Meteorological Summary
CONCORD, MASSACHUSETTS. MONTH OF
Concord, Mass. Month of August 1893

	DRY BULB		WET BULB		DEW POINT.		REL. HUMID.					WIND.		WEATHER.		BAROMETER.		RAIN.
	7 A. M.	7 P. M.	7 A. M.	7 P. M.	7 A. M.	7 P. M.	7 A. M.	7 P. M.	MAX.	MIN	MEAN.	7 A. M.	7 P. M.	7 A. M.	7 P. M.	7 A. M.	7 P. M.	
1	66	73	65	69	64	67	95	88	81	64	72.5	SW	S	cloudy	cloudy	29.83	29.81	[illegible]
2	65	69	60	62	57	58	75	67	79	59	69	N	NW	clear	clear	30.04	30.15	
3	60	73	57	63	55	58	84	61	85	43	64	calm	S	clear	clear	30.25	30.12	
4	67	70	62	63	59	59	76	68	83	56	69.5	SW	SW	fair	cloudy	30.12	30.04	
5	63	65	63	63.5	63	63	100	92	67	62	64.5	NE	NE	rain	cloudy	29.92	29.90	1.16
6	64	71	63	70	62	70	95	95	85	56	70.5	S	S	cloudy	cloudy	29.90	29.81	0.04
7	65	59	64	58.5	63	58	95	97	77	57	67	N	W	cloudy	fair	29.77	29.76	2.54

Climate Data — Records
[NASA]

Simulation Studies

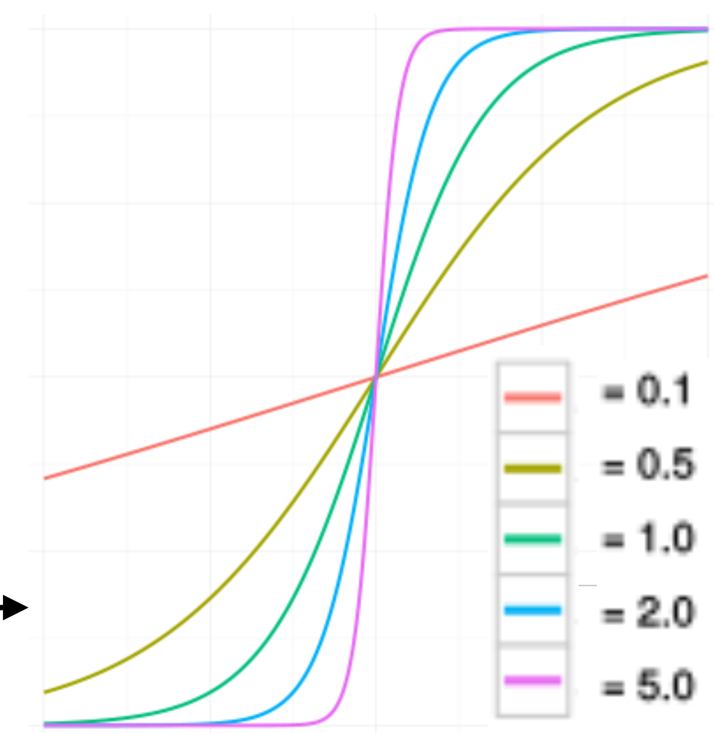




$$f_1(x) \sim GP(0, K_1(x)) \qquad f_2(x) \sim GP(0, K_2(x))$$

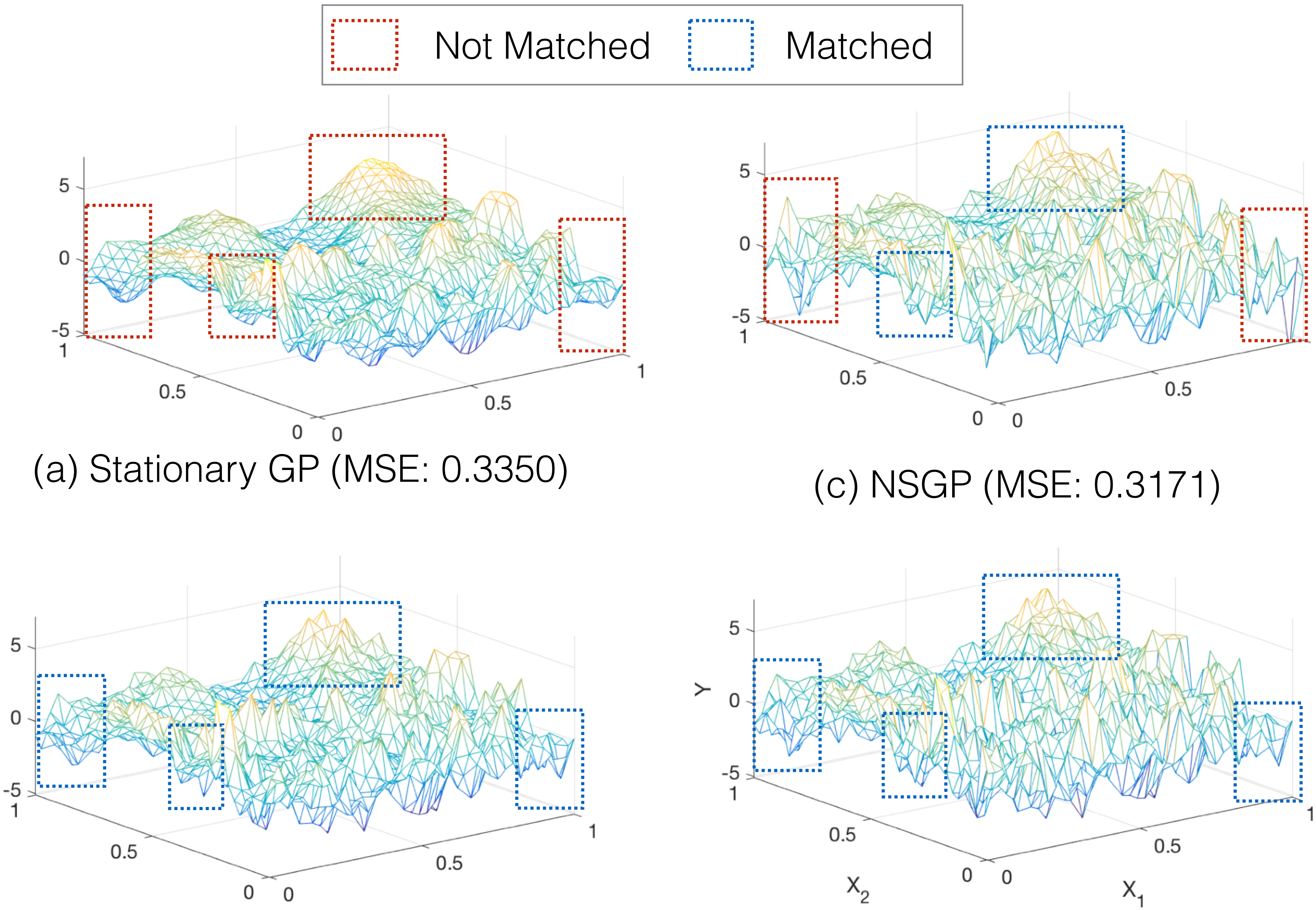
Simulation Function: $y = f_1(x)w(x) + f_2(x)(1 - w(x))$

Generalized Logistic Function ⟶ $w(x)$ ⟶



Dalal et al. (2017) [In Prep]

Simulation Studies (Surfaces)



(a) Stationary GP (MSE: 0.3350)

(c) NSGP (MSE: 0.3171)

(b) Proposed NSGP (MSE: 0.2910)

(d) Target Surface

Evaluation Metrics

1. Mean Square Error (smaller is better): mean estimates

2. Negative Log Predictive Density (smaller is better): mean + variance estimates

3. Continuous Rank Probabilistic Score (larger is better): Probabilistic forecasting

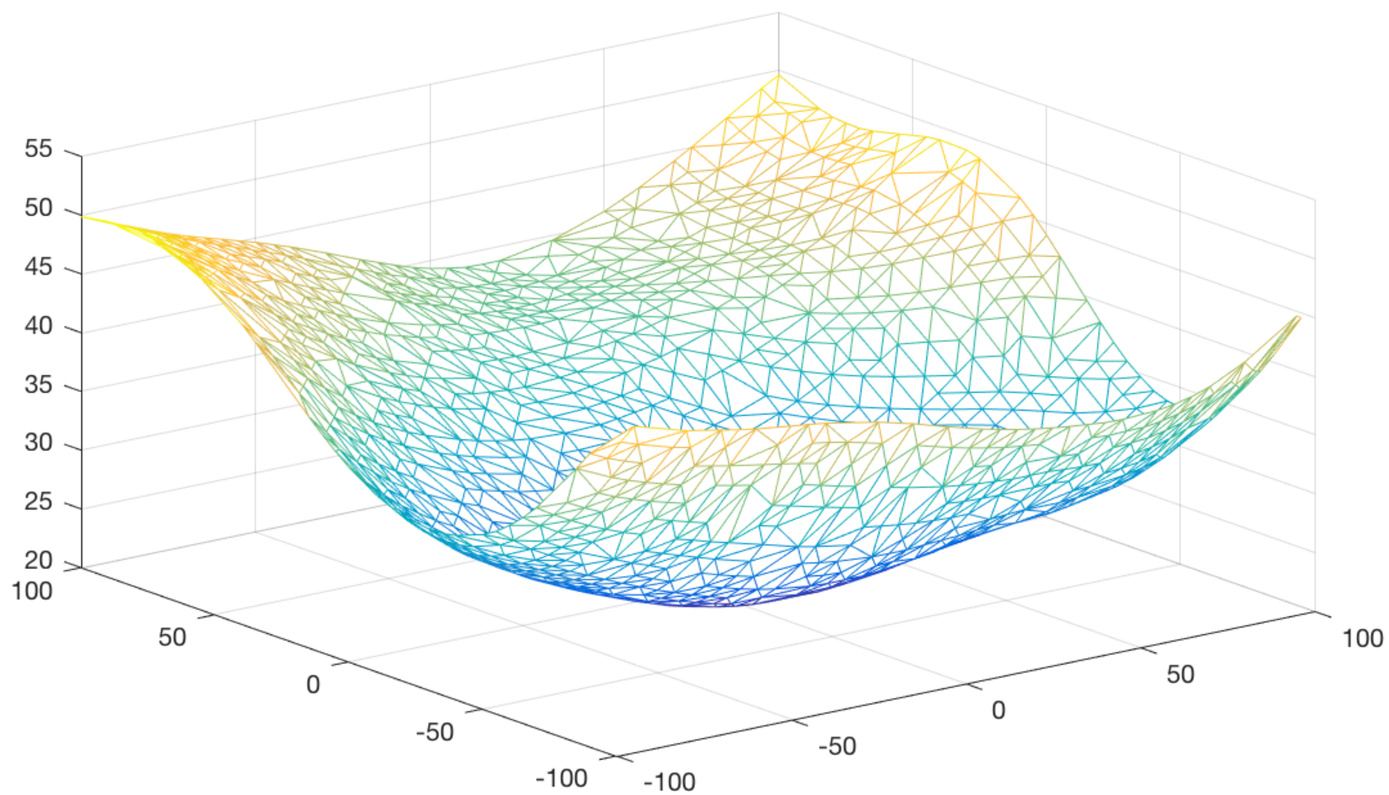
GP models	statGP	NSGP	proposed NSGP	True-par-GP
MSE	0.33(0.02)	0.31(0.01)	**0.30(0.01)**	**0.30(0.01)**
NLPD	0.87(0.01)	0.80(0.01)	**0.78(0.01)**	**0.77(0.01)**
CRPS	0.47(0.2)	0.49(0.1)	**0.50(0.1)**	**0.50(0.1)**

Proposed NSGP recovers the underlying process.

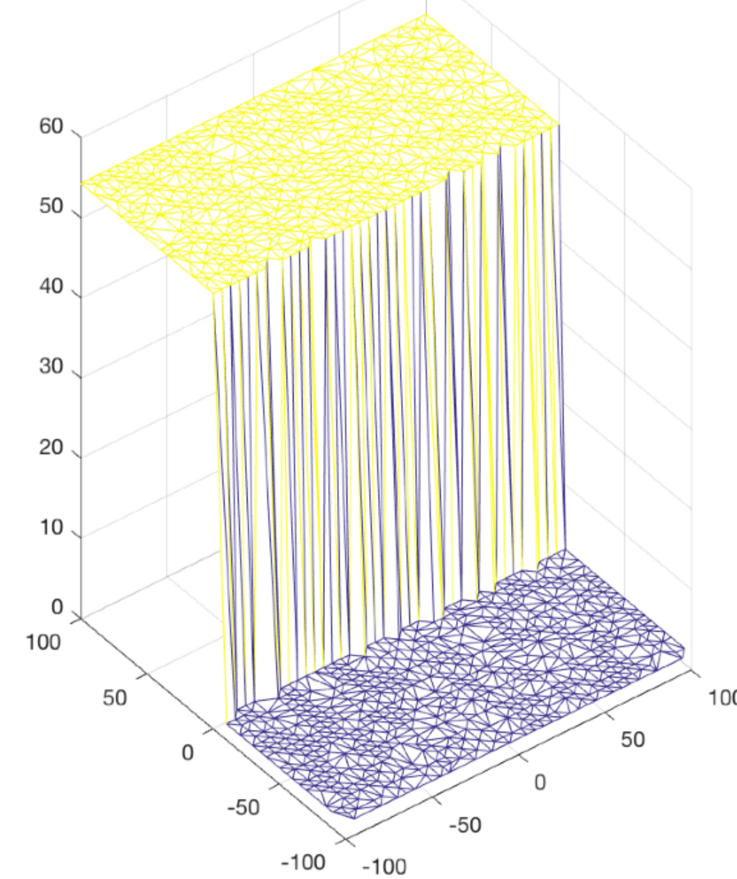
MCMC

Reconstructed surface of the scale matrix range parameter from the MCMC draws

MCMC-NSGP



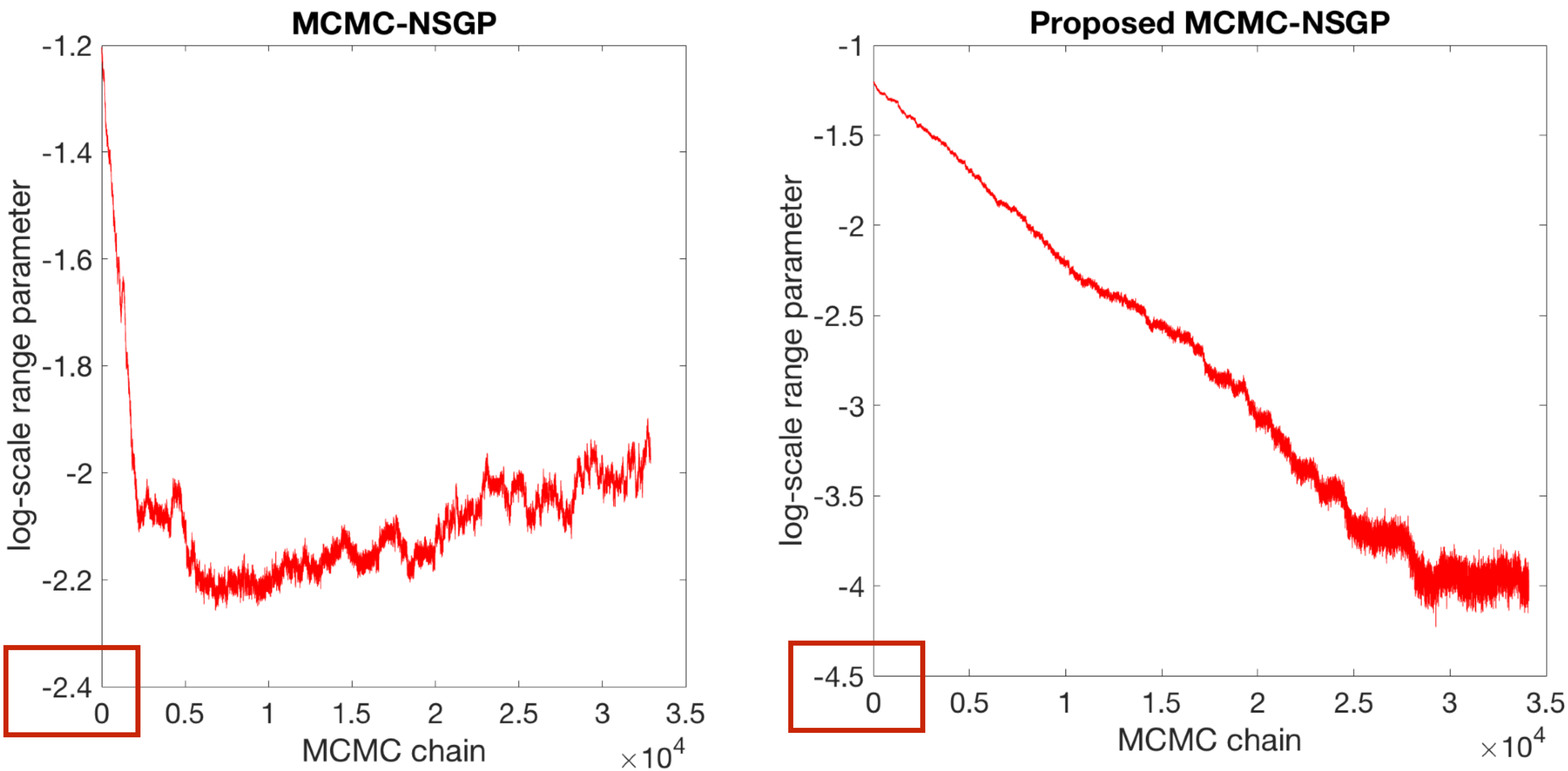
Proposed MCMC-NSGP



The proposed MCMC-NSGP recovers the underlying geometric anisotropy (Scale)

MCMC

MCMC draws of the scale matrix range parameter at s_i

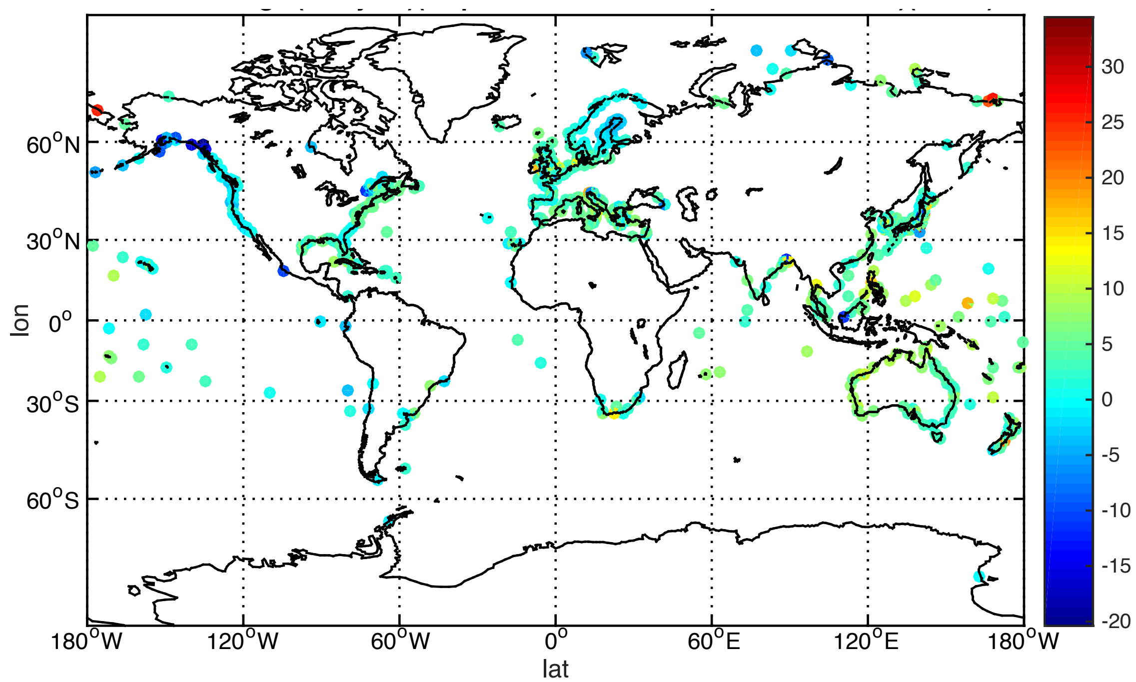


MCMC-NSGP does not converge (slow mixing issue)

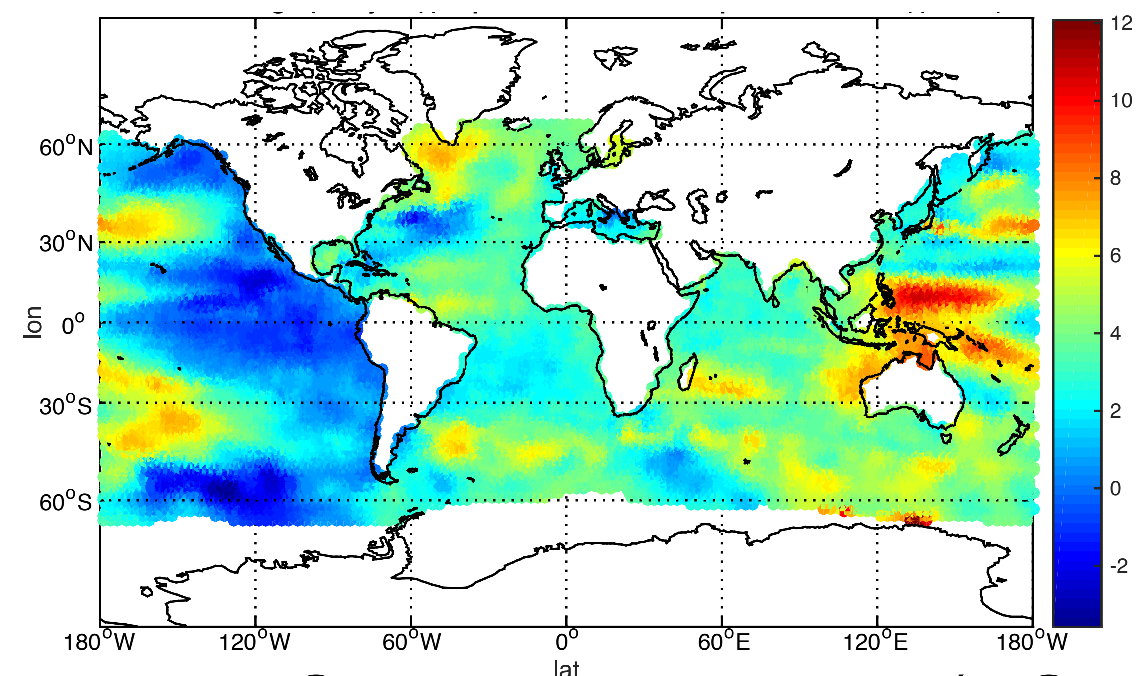
Talk outline

- Scientific Goal 1: Improve estimates of sea-level trends using spatial information
  - Regression Models
  - Simulation Study
  - **Sea-Level Datasets: Spatial (Dim=2) & Spatial-temporal (Dim=3)**
- Scientific Goal 2: Inference from multiple sources of datasets
- Scientific Goal 3: Inter-comparison of Earth System Models
- Scientific Goal 4: Emulate future climate scenarios

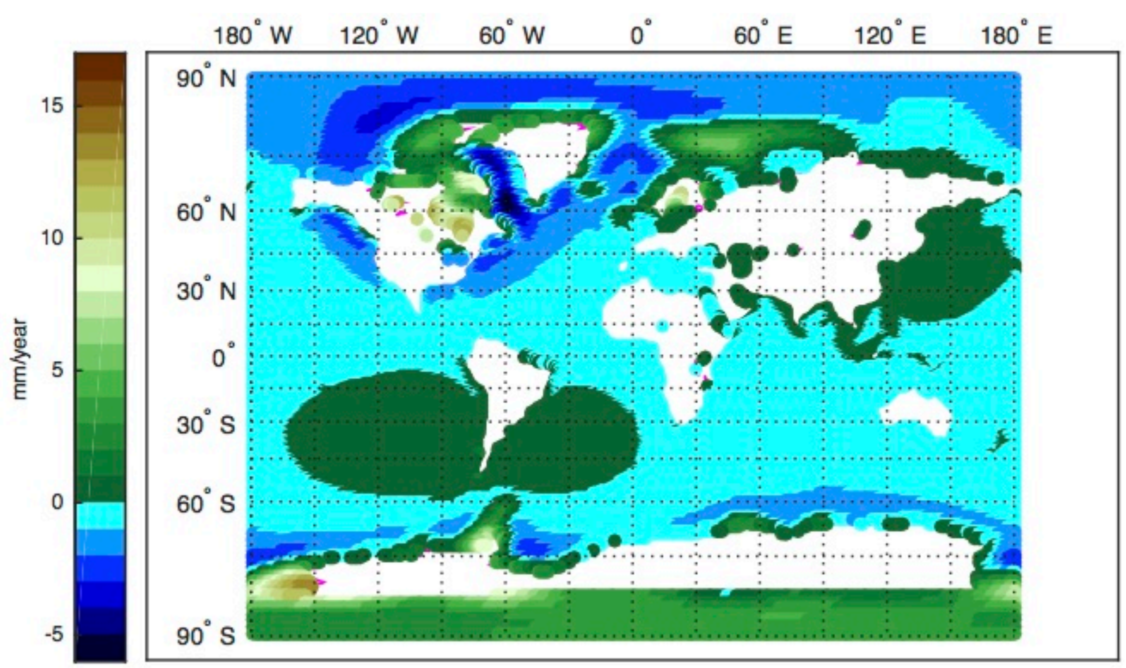
Climate (Sea Level) Dataset



Station Dataset (PSML)



Remote Sensing Dataset (TOPEX)



Geophysical Simulation (GIA-VLM)

Experimental Set-up:

1. Time Frame: 1993 to 2012
2. Region-wise hold-out
3. Trends are found using time-series regression
4. All units in mm/year

Results

Results for Spatial Datasets (m = 2)

GP Models	statGP		NSGP		propNSGP	
Eval Metric	MSE	NLPD	MSE	NLPD	MSE	NLPD
Tide Gauge	0.85	2.81	0.75	2.82	**0.71**	**2.78**
GIA-LVM	0.58	3.08	0.56	2.00	**0.54**	**1.94**

Dalal et al, 2015 (Climate Informatics, **Best Paper Award**)

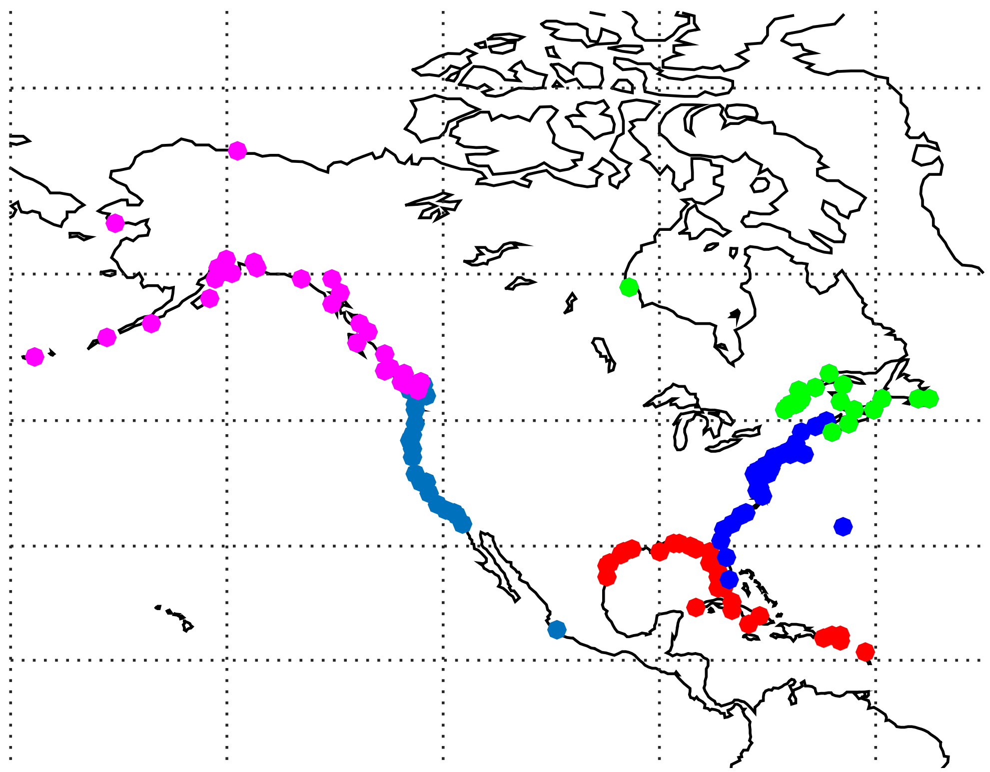
Estimates with regional structure is better!

Results for Spatio-Temporal Datasets (m = 3)

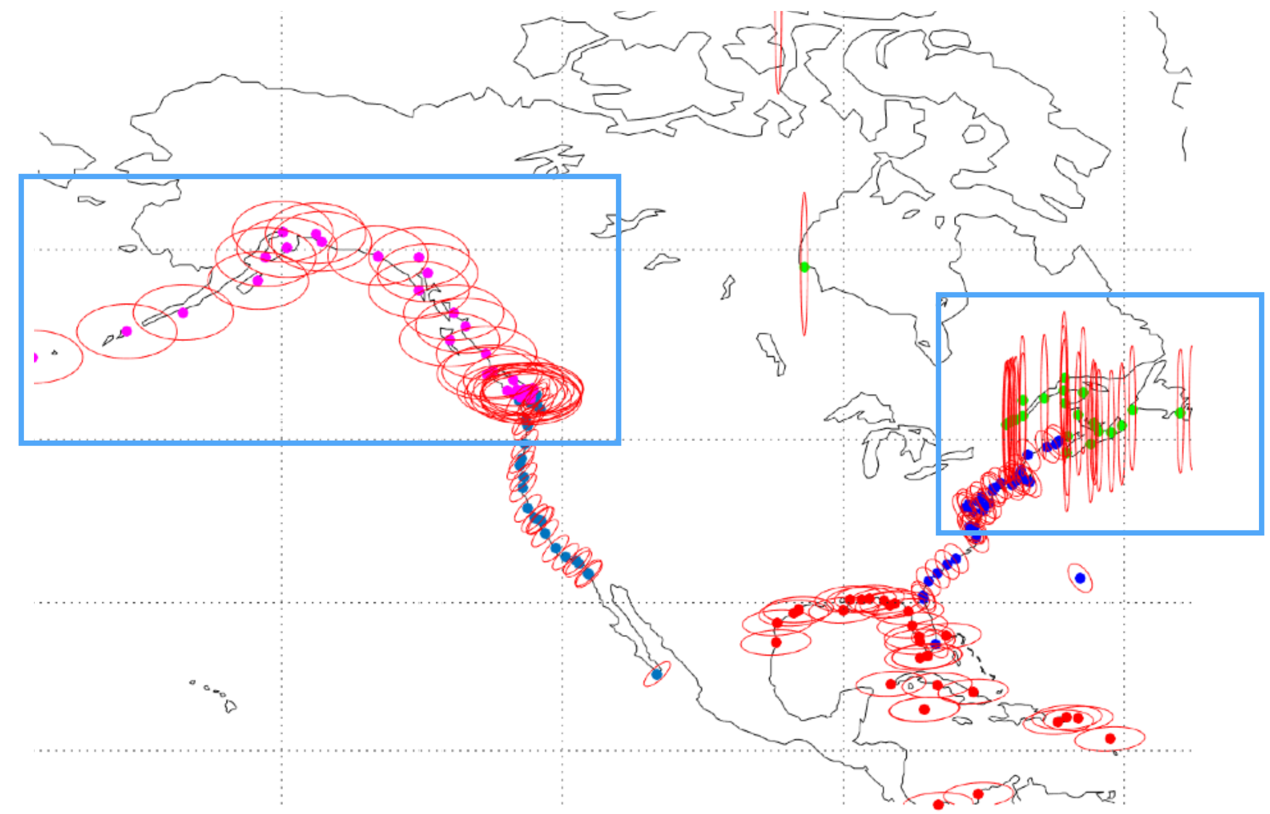
GP models	statGP		NSGP		propNSGP	
Eval Metric	MSE	NLPD	MSE	NLPD	MSE	NLPD
Remote Sense	1.38	4.29	1.18	4.25	**1.10**	**4.15**

Dalal et al. (2015), American Geophysical Union

Climate (Sea Level) Dataset



Geophysical Clusters of Stations
[Kopp et al., Nature (2013);
Hay et al., Nature (2015)]



Proposed Model
[Dalal et al. (2015);
Dalal et al. (2017) [In Prep]]

Geometric Anisotropy (ellipses) aligns with the orientation of the near by points at the coastline

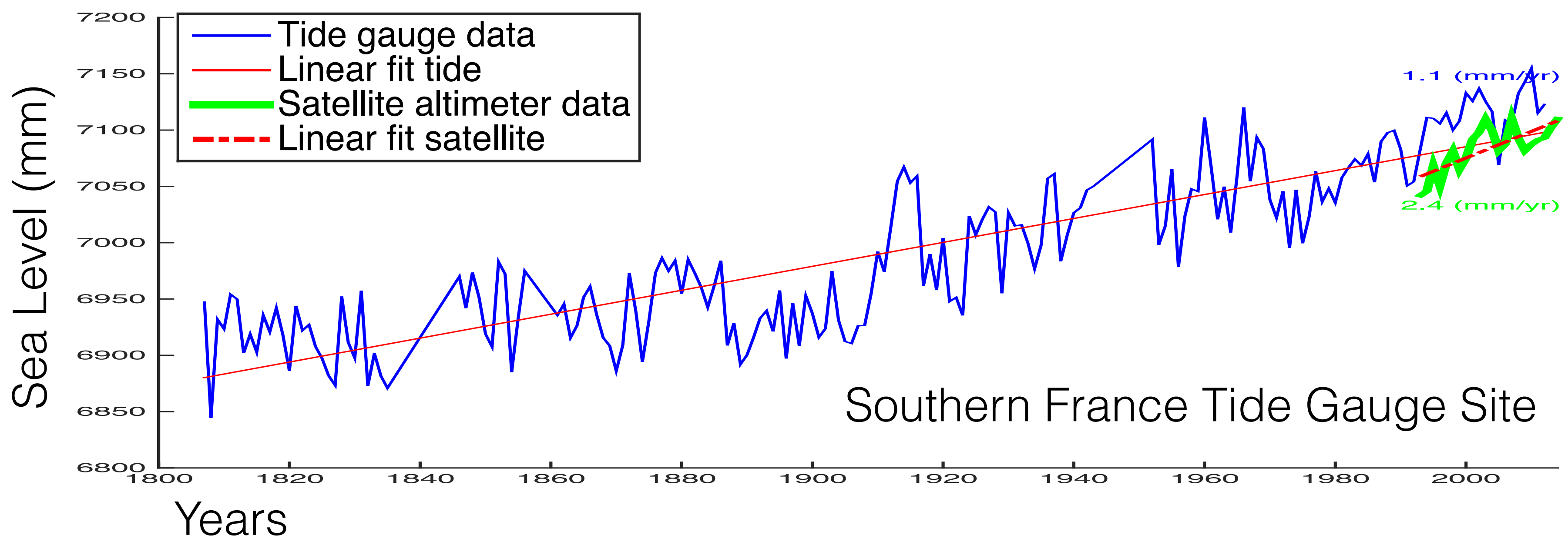
Data clusters conforms with the geophysical clusters

Talk outline

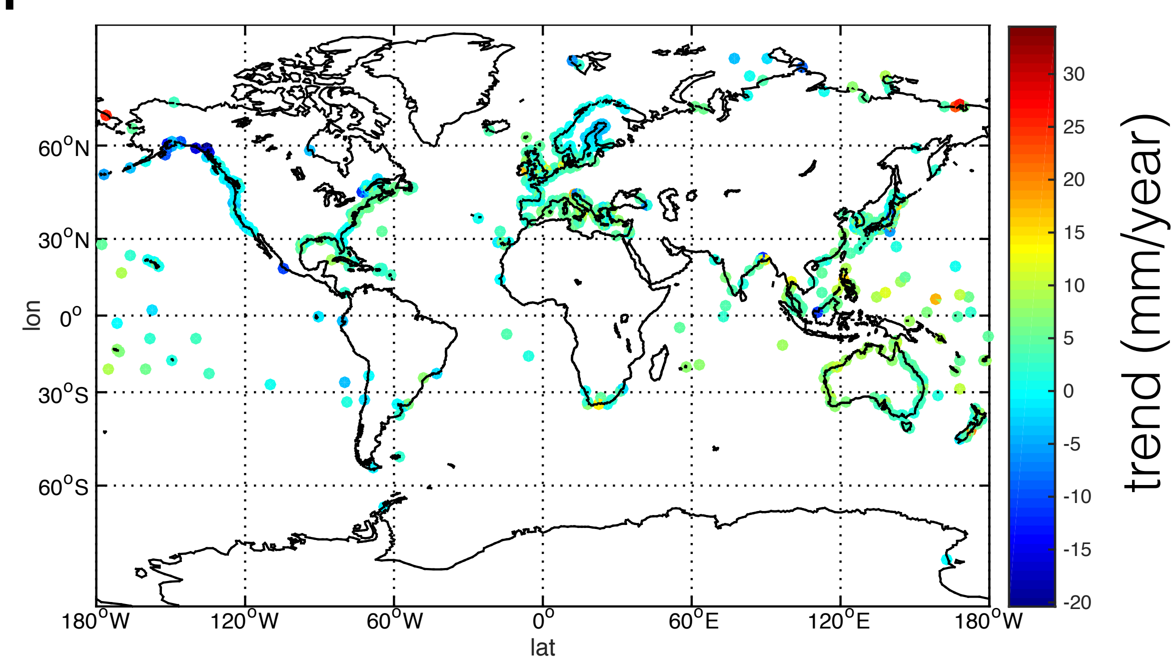
- Scientific Goal 1: Improve estimates of sea-level trends using spatial information
- **Scientific Goal 2: Inference from multiple sources of datasets**
- Scientific Goal 3: Inter-comparison of Earth System Models (Dim > 3)
- Scientific Goal 4: Emulate future climate scenarios (Dim > 3)

Multiple Data-products

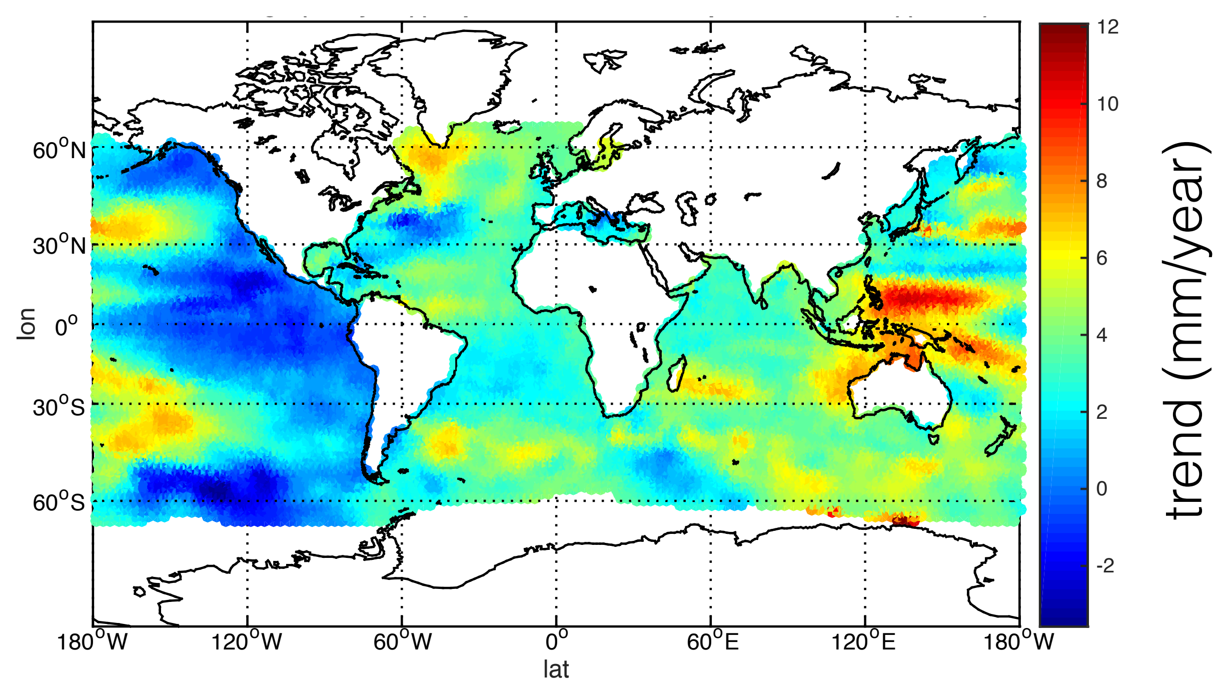
Temporal Data



Spatial Data



Spatially Sparse (Station Data)



Spatially Dense(Remote Sensing)

Prior Approaches in Climate

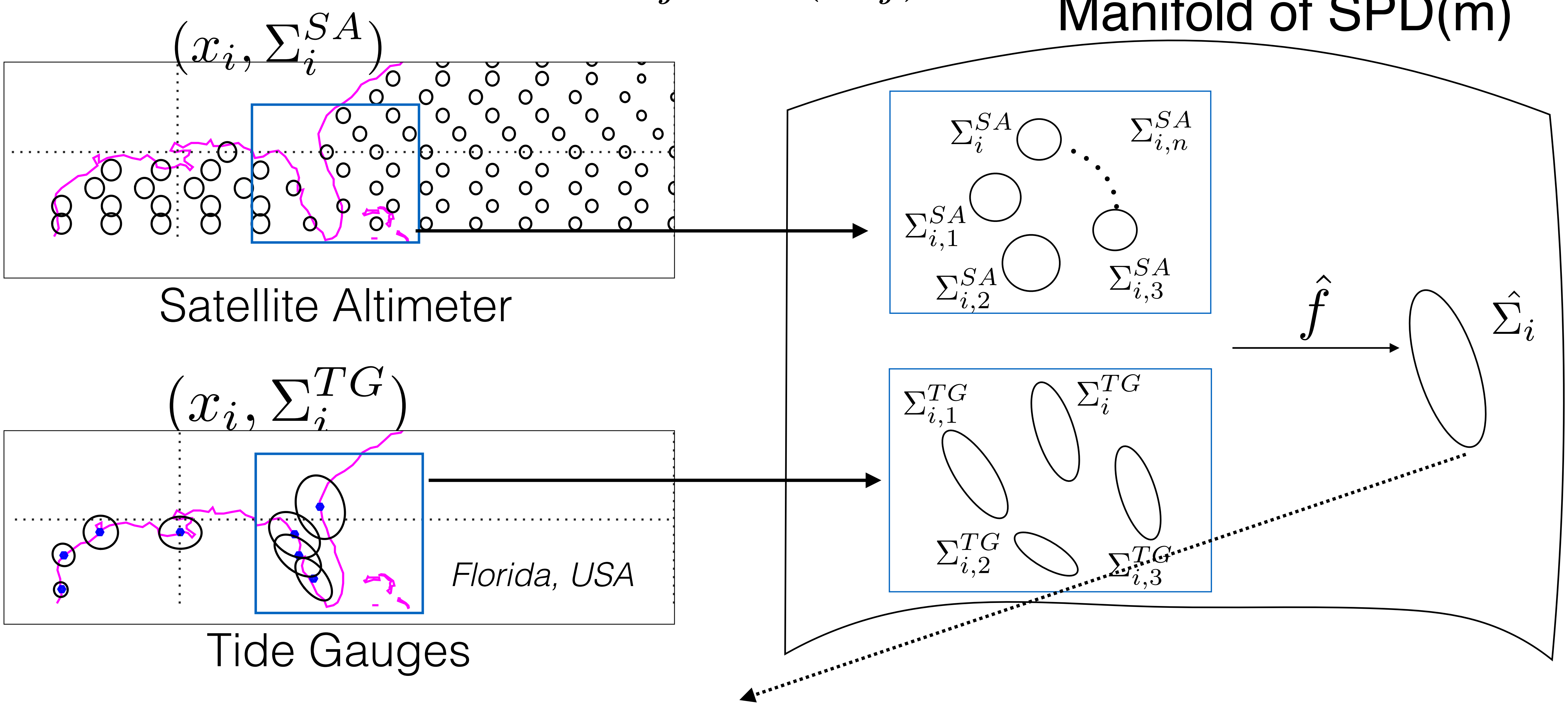
For time series with one variable: Church & White (2011)

Region specific: Davis et al (2010), Ouillon (2003)

Climate variable specific: IPCC (2013)

Data - Fusion Model

$$k_{ij} \propto \mathcal{H}(Q_{ij})$$



$$Q_{ij} = (x_i - x_j)^T \left(\frac{\hat{\Sigma}_i + \hat{\Sigma}_j}{2} \right)^{-1} (x_i - x_j)$$

Parameter Estimation

Regression for $\Sigma \in SPD(m)$: $\Sigma_{SA}(s) = \phi_1(s)\Sigma_{TG}(s) + \phi_2(s)$

But, $\Sigma \in SPD(m) \to \phi_1 \Sigma \notin SPD(m)$

Thanks to affine-invariant metric!

$$\log : SPD(m) \to S(m) \quad , \quad \exp : S(m) \to SPD(m)$$

$$\boxed{\hat{\Sigma}_i \longrightarrow \Sigma_{SA}(s) = \exp(\phi_1(s)\log(\Sigma_{TG}(s))) + \phi_2(s)}$$

Theorem 2: The proposed covariance function is a valid covariance function.

Dalal et al. (2015);
Dalal et al. (2017) [In prep]

Results

Results for Tide Gauge Spatial Datasets (m = 2)

GP models	statGP	NSGP	propNSGP	data-fusion-NSGP
MSE	0.85	0.75	0.71	**0.66**
NLPD	2.81	2.82	2.78	**2.75**

Dalal et al. (2015)
(Climate Informatics, **Best paper award**)

Estimates from multi-sources of information is better!

Results for Remote-Sensing Spatio-Temporal Datasets (m = 3)

GP models	statGP	NSGP	propNSGP	data-fusion-NSGP
MSE	1.38	1.18	1.10	**1.09**
NLPD	4.29	4.25	4.15	**4.11**

Dalal et al. (2015),
American Geophysical Union

Talk outline

- Scientific Goal 1: Improve estimates of sea-level trends using spatial information

- Scientific Goal 2: Inference from multiple sources of datasets

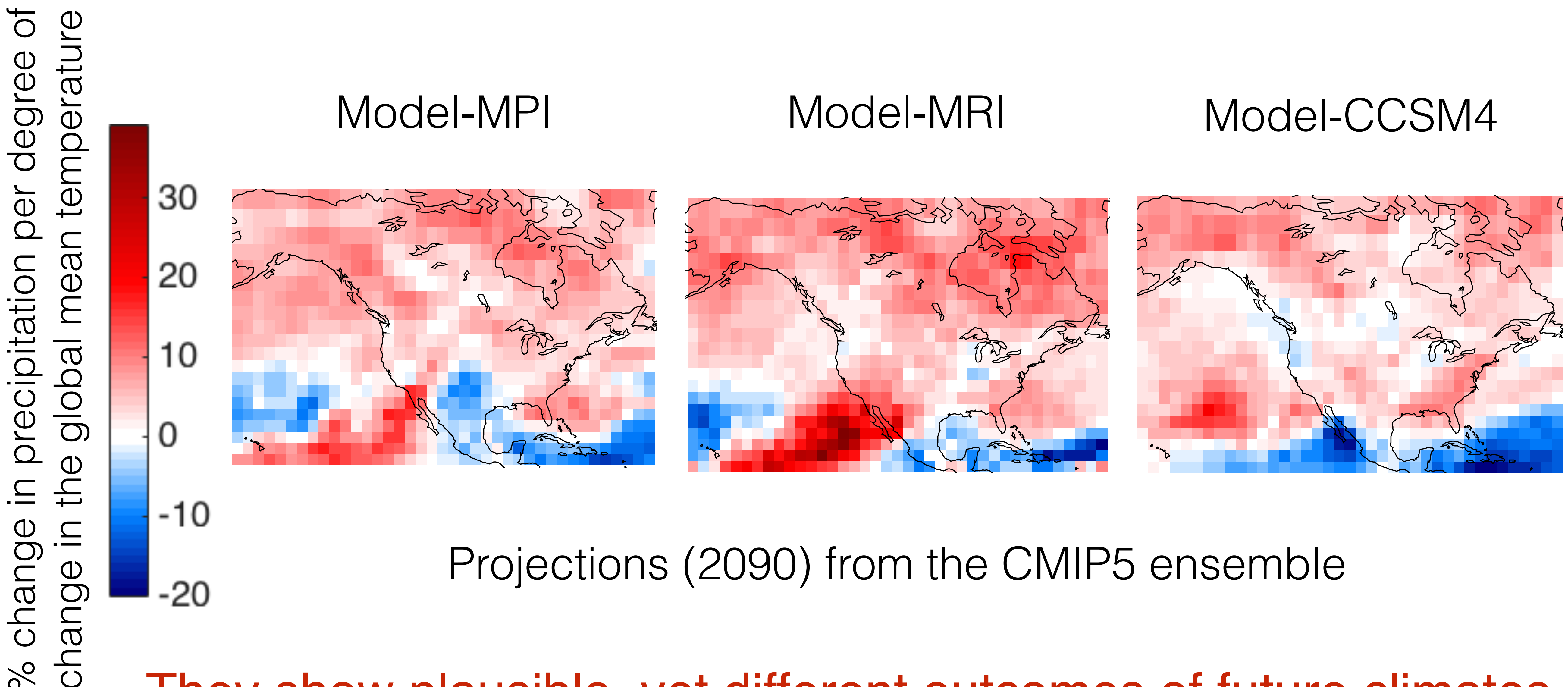
- **Scientific Goal 3: Inter-comparison of Earth System Models (Dim > 3)**

- Scientific Goal 4: Emulate future climate scenarios (Dim > 3)

"One's intuition in higher dimensional space is not worth a damn!" - Dantzig

Dimensions > 3

Climate model outputs from various Earth System Modeling groups



Projections (2090) from the CMIP5 ensemble

They show plausible, yet different outcomes of future climates.

Prior Approaches in Climate

Genealogy: Knutti et al. (2013)

Bayesian Approaches: Tebaldi et al. (2005, 2007), Furrer et al. (2007), Leith (2010)

Machine Learning Approach: Sanderson et al. (2015)

Distance Function

Step 1: Learn data parameters {sill, range, nugget} $\theta = \{\sigma, \phi, \psi\}$

$$\theta_i = \mathrm{argmax}_\theta \hat{l}(\theta, y_i)$$

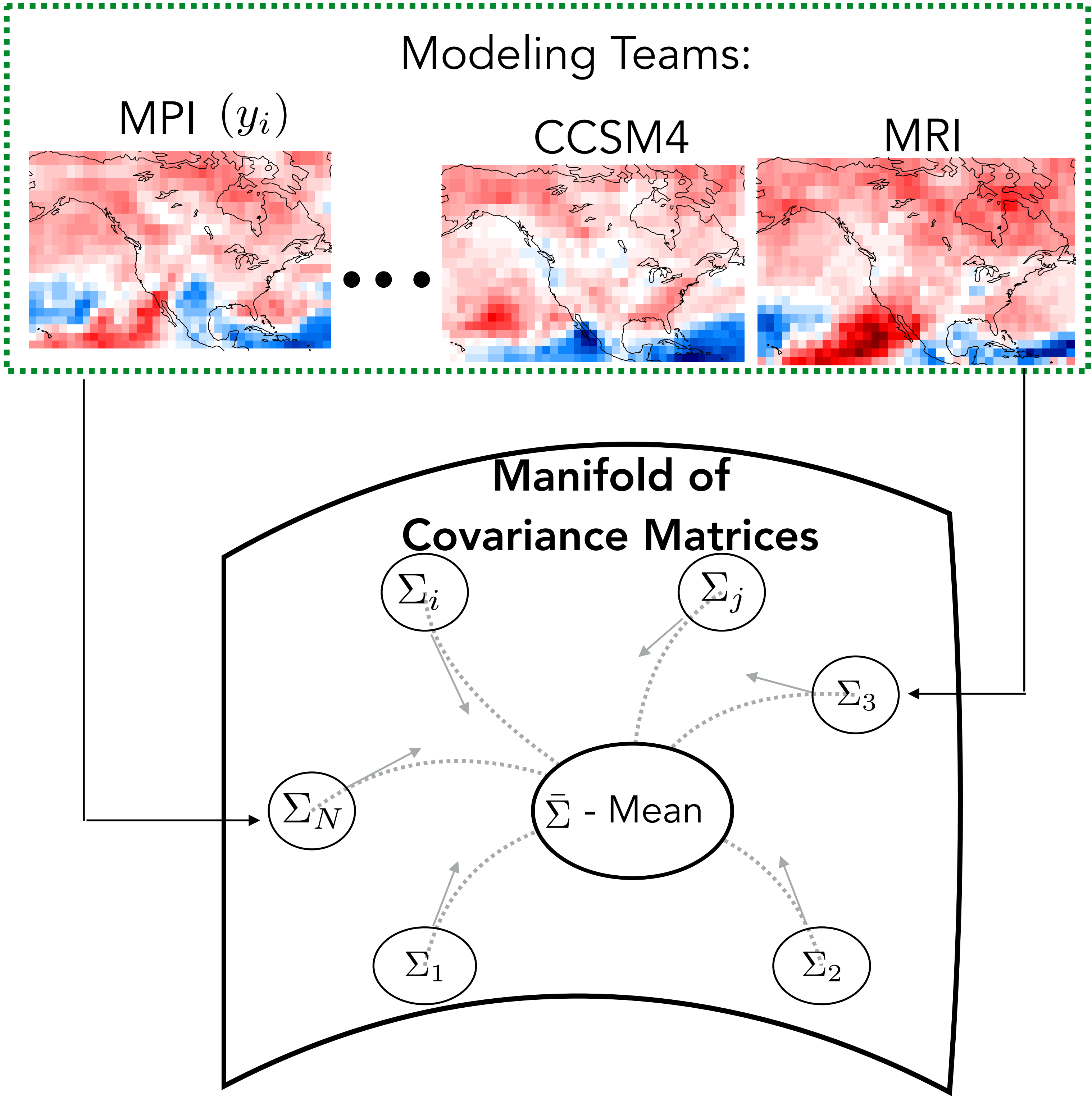
Step 2: Fitting covariance for each climate model output

$$\Sigma_i(\theta_i) = \psi_i I + \sigma_i^2 H(\phi_i)$$

Step 3: Distance on a manifold

$$\Sigma_i(\theta_i) \in SPD(n)$$

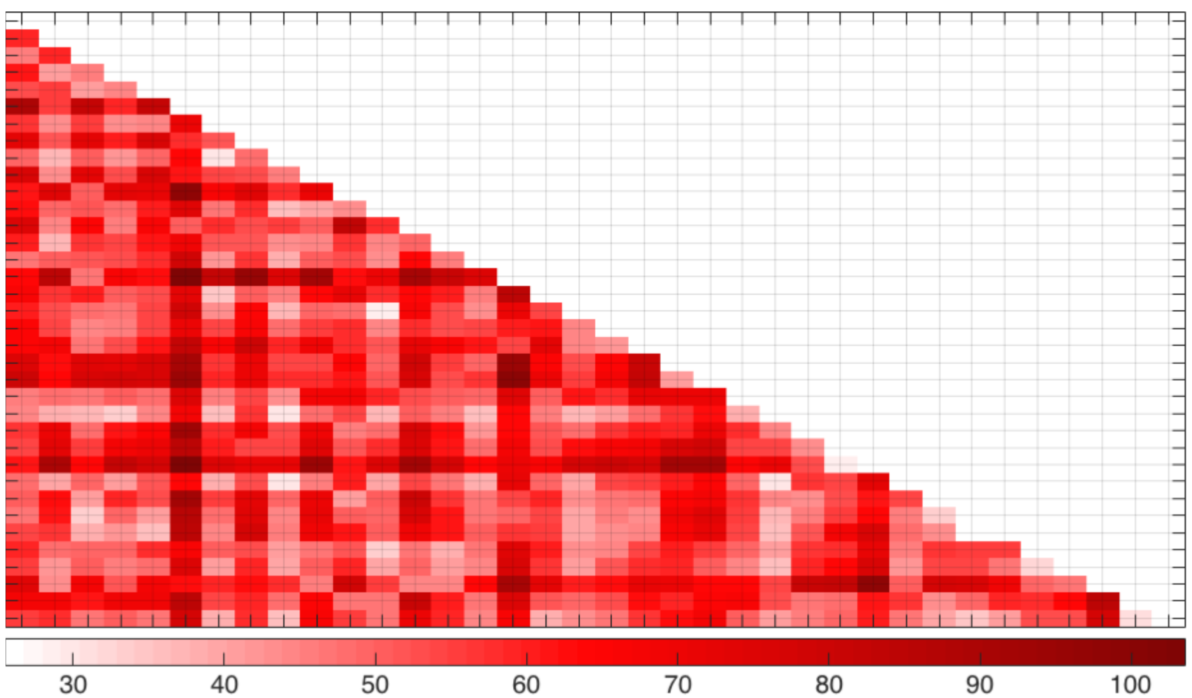
$$D^2(\Sigma_1, \Sigma_2) = \frac{1}{2}\mathrm{tr}(\log^2(\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}))$$



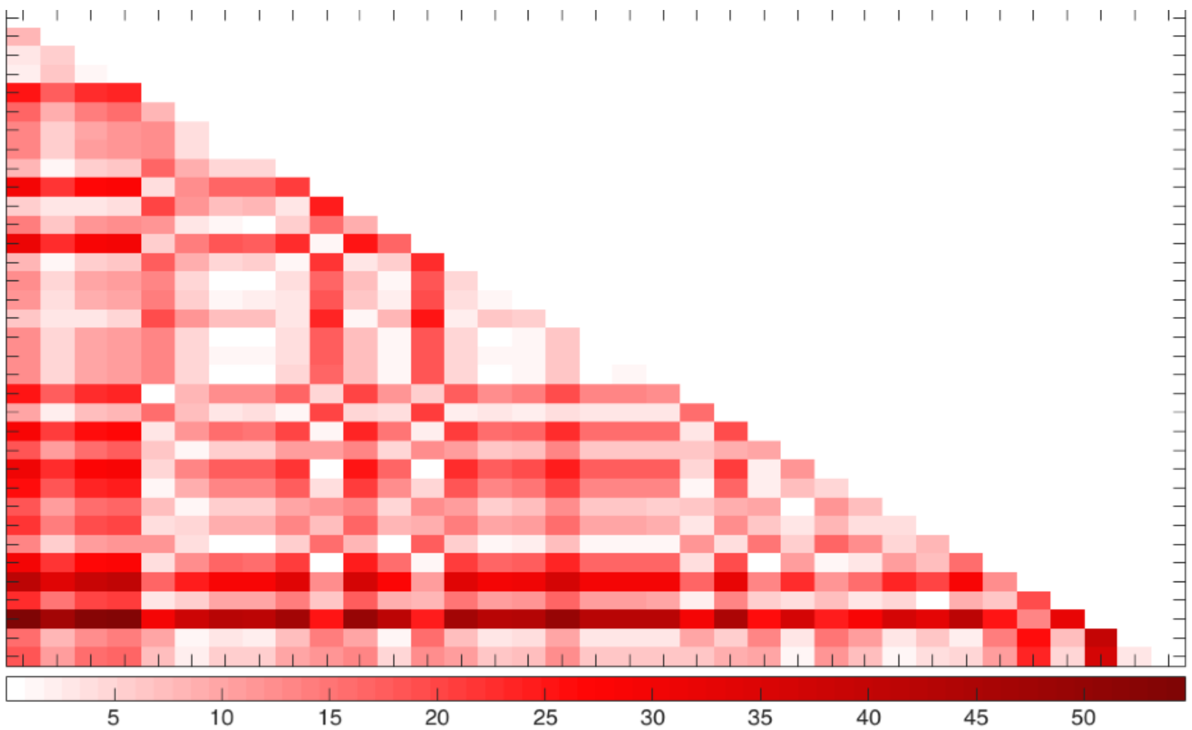
Evaluation: Distance Matrix

Little Contrast

- More Contrast
- More Dependencies Identified



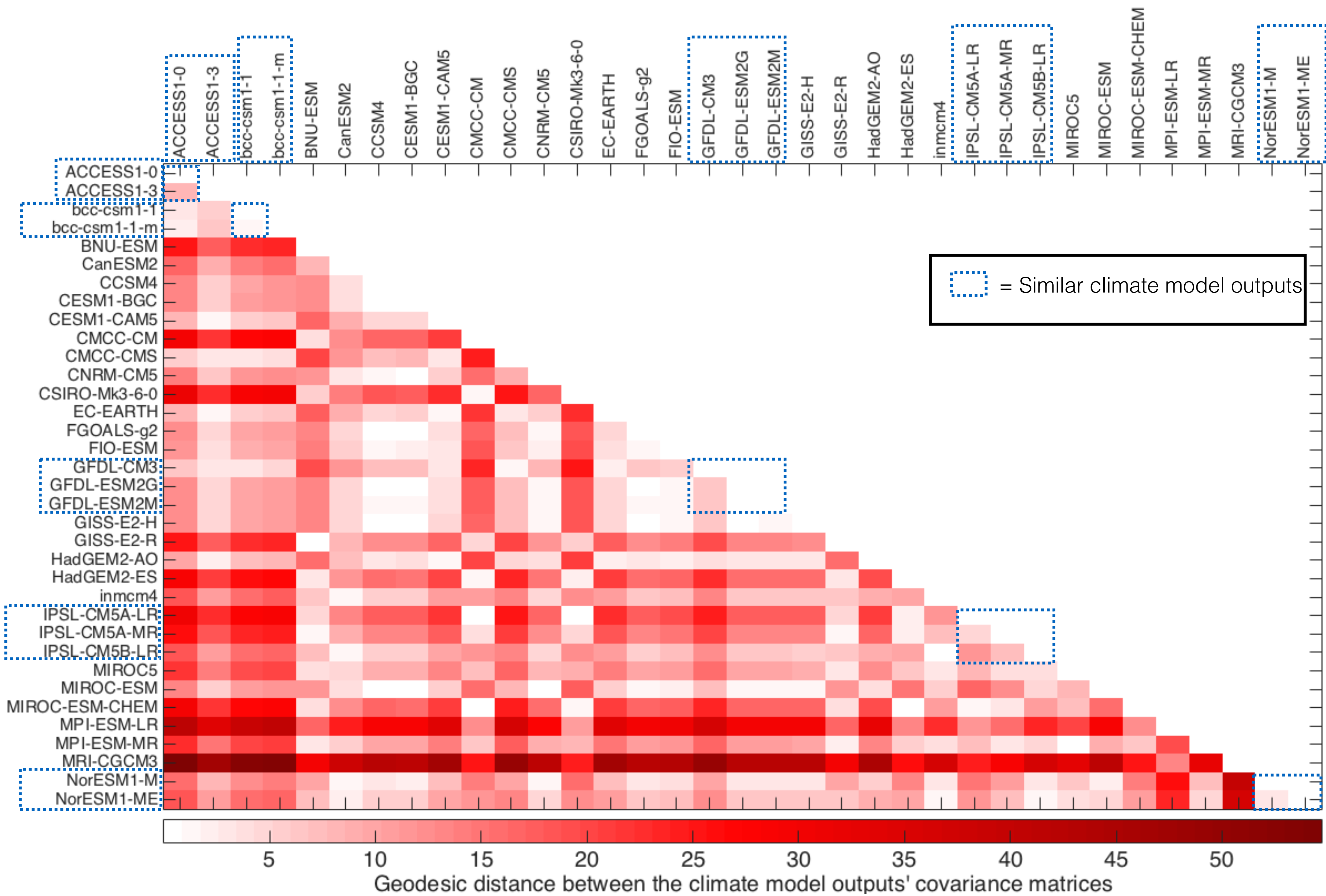
Euclidean distance between climate model outputs



Geodesic distance between covariance matrices of climate model outputs

Dalal et al. (2016)
(Climate Informatics, **Best paper award**)

Evaluation: Model Dependencies

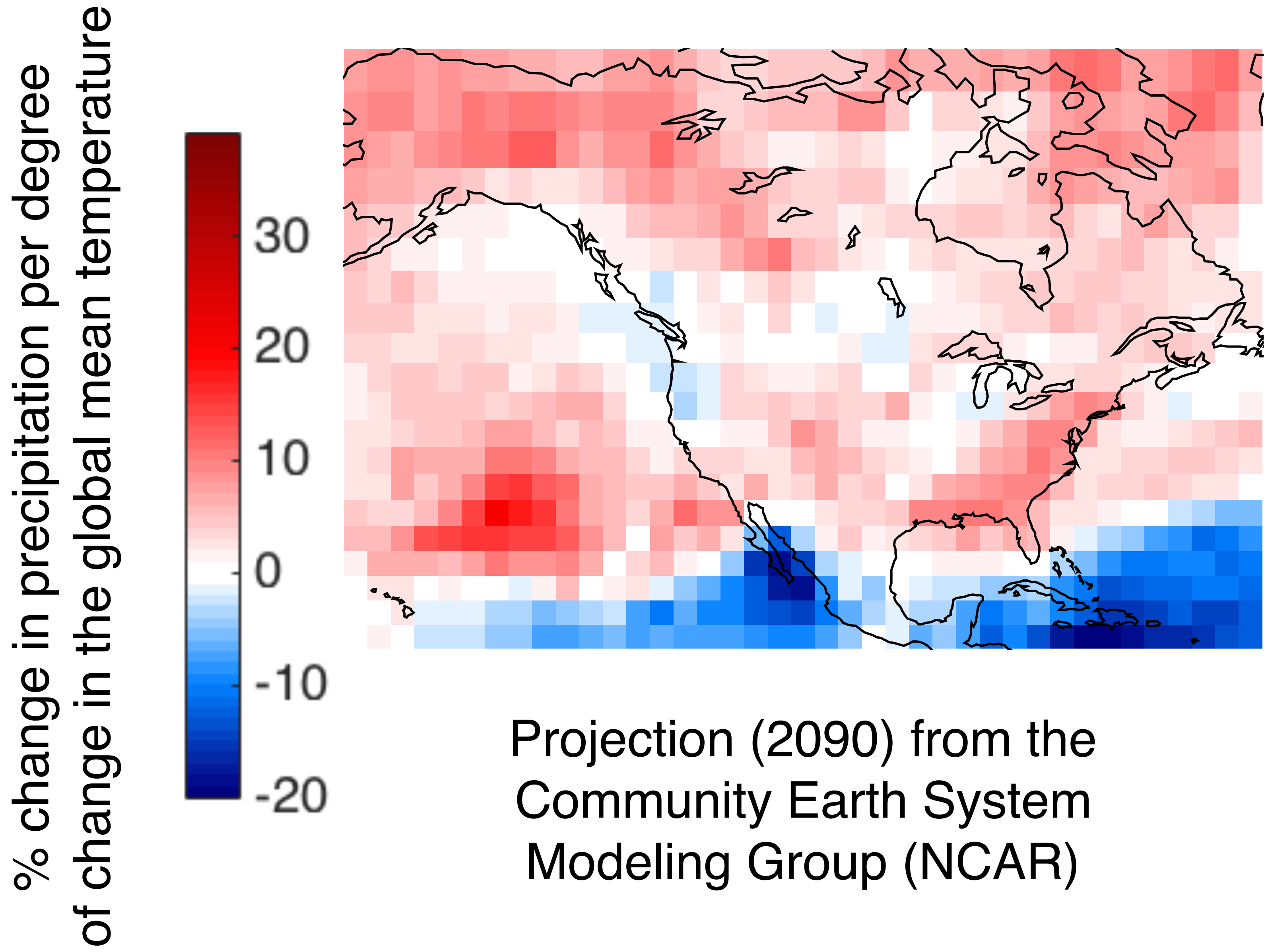


Dalal et al. (2016)
(Climate Informatics, **Best paper award**)

Talk outline

- **Scientific Goal 1**: Improve estimates of sea-level trends using spatial information
- **Scientific Goal 2**: Inference from multiple sources of datasets
- **Scientific Goal 3**: Inter-comparison of Earth System Models (Dim > 3)
- **Scientific Goal 4: Emulate future climate scenarios (Dim > 3)**

Dimensions > 3



Projection (2090) from the Community Earth System Modeling Group (NCAR)

Earth system models running on a supercomputer take several months to output projections of future climates.

Prior Approaches in Climate

Bayesian Approaches: Tebaldi et al (2005, 2007, 2009)

A list of very specific approaches: IPCC (2013)

Statistical Emulator

Multivariate Normal Sampling (MVN) Scheme

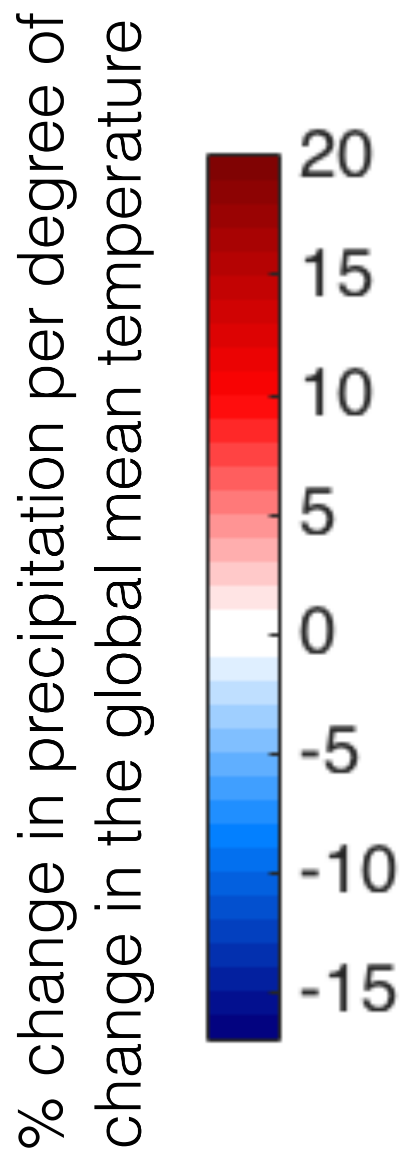
$$\tilde{\mathbf{y}} = \mu + \Sigma^{\frac{1}{2}}\epsilon$$

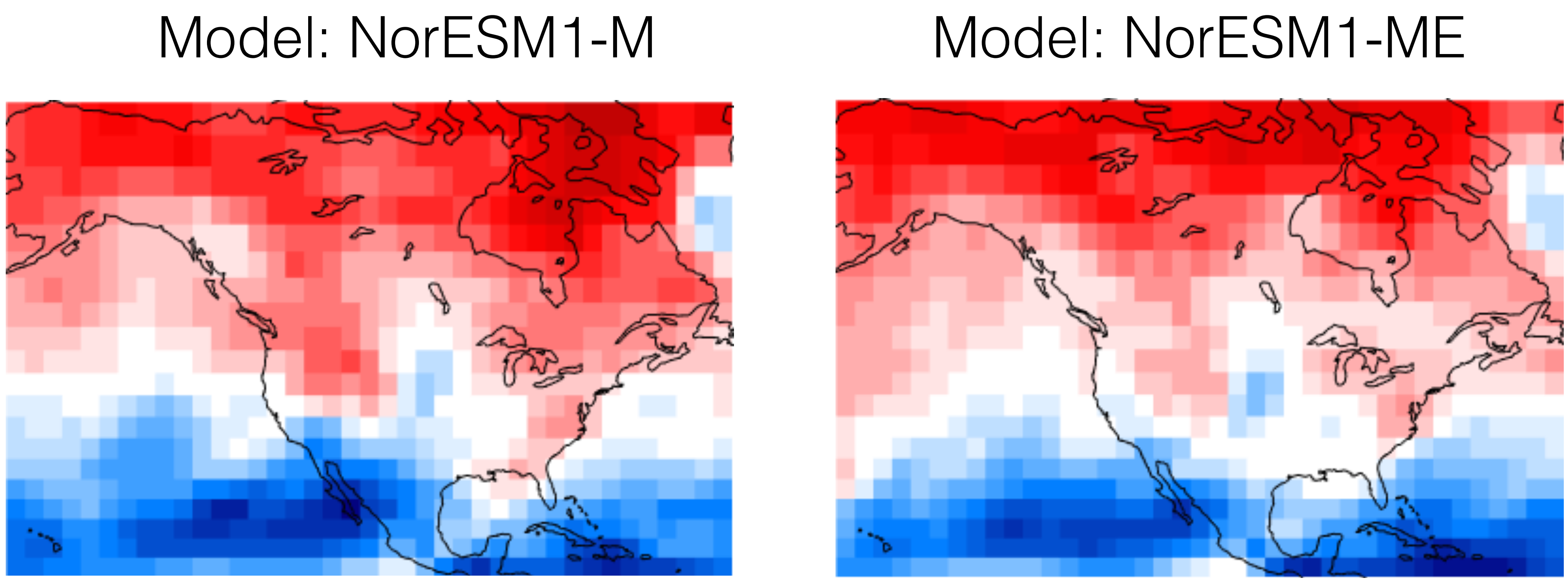
$\hat{\mu} \longleftarrow$ Multi-model ensemble mean, $\hat{\mu} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathbf{y_i}$

$\hat{\Sigma} \longleftarrow$ Covariance fitting $\Sigma(\theta)$, θ = {range, sill, nugget}

Challenges

Climate model outputs are not independent.





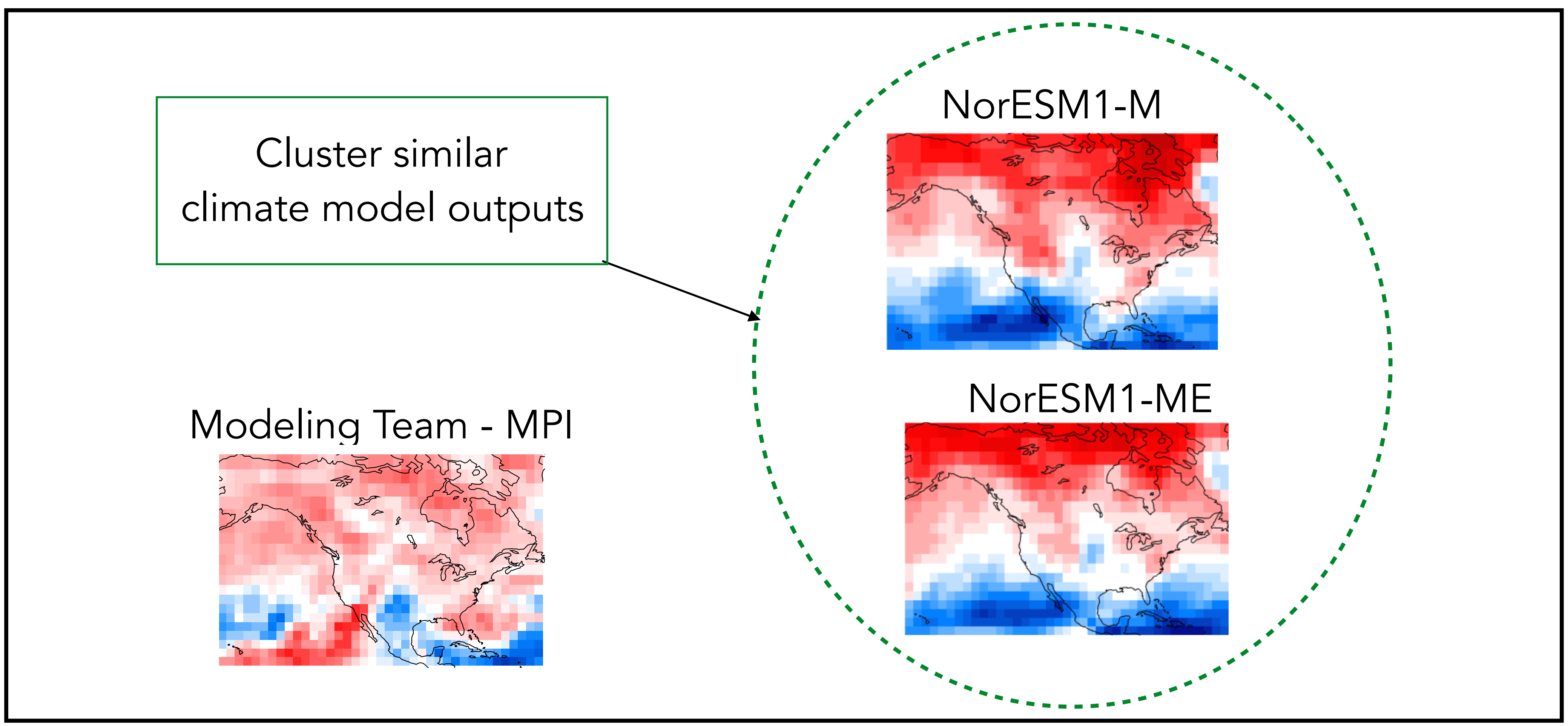
Projections (2090) from the same modeling group (NCC) in the CMIP5-RCP8.5 ensemble.

Proposed Sampling Scheme

Weighted multi-model ensemble mean:

Weights are Informed from high-dimensional geometric structure

$$\hat{\mu} = \sum_{j=1}^{n} \sum_{k=1}^{m_j} \frac{1}{nm_j} \mathbf{y}_{j,k}$$, where n is the number of clusters.



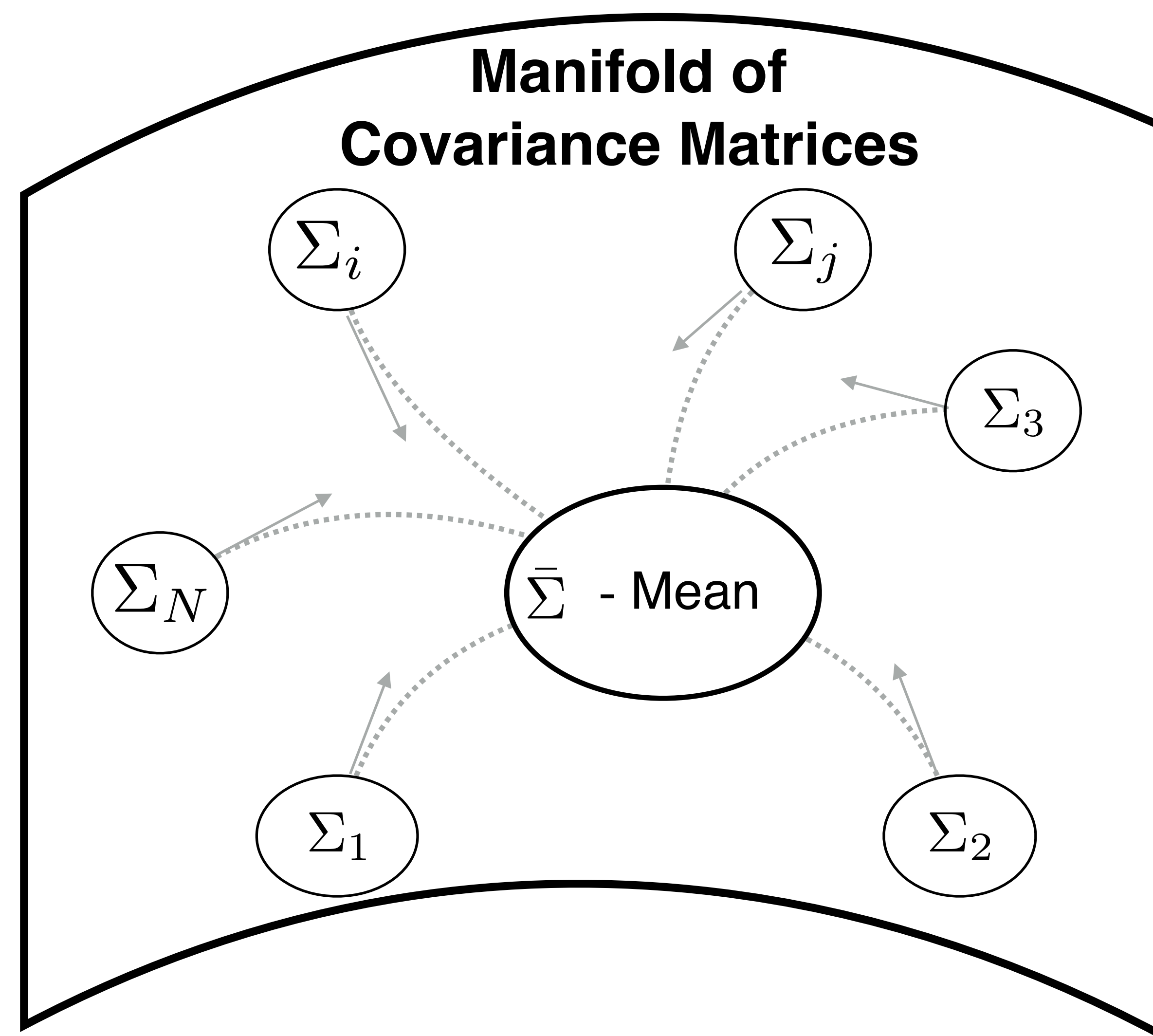
Proposed Sampling Scheme

Sample a covariance matrix from a distribution on a manifold.

$$\hat{\Sigma} \sim \mathcal{N}(\bar{\Sigma}, \Lambda | \Sigma(\theta_1), \ldots, \Sigma(\theta_{\mathrm{N}}))$$

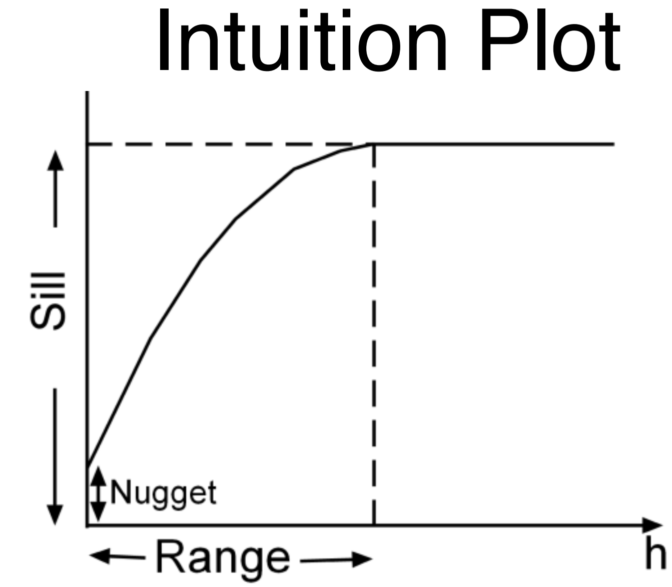
$\bar{\Sigma} \leftarrow$ Weighted mean estimates of CMO of covariance matrices

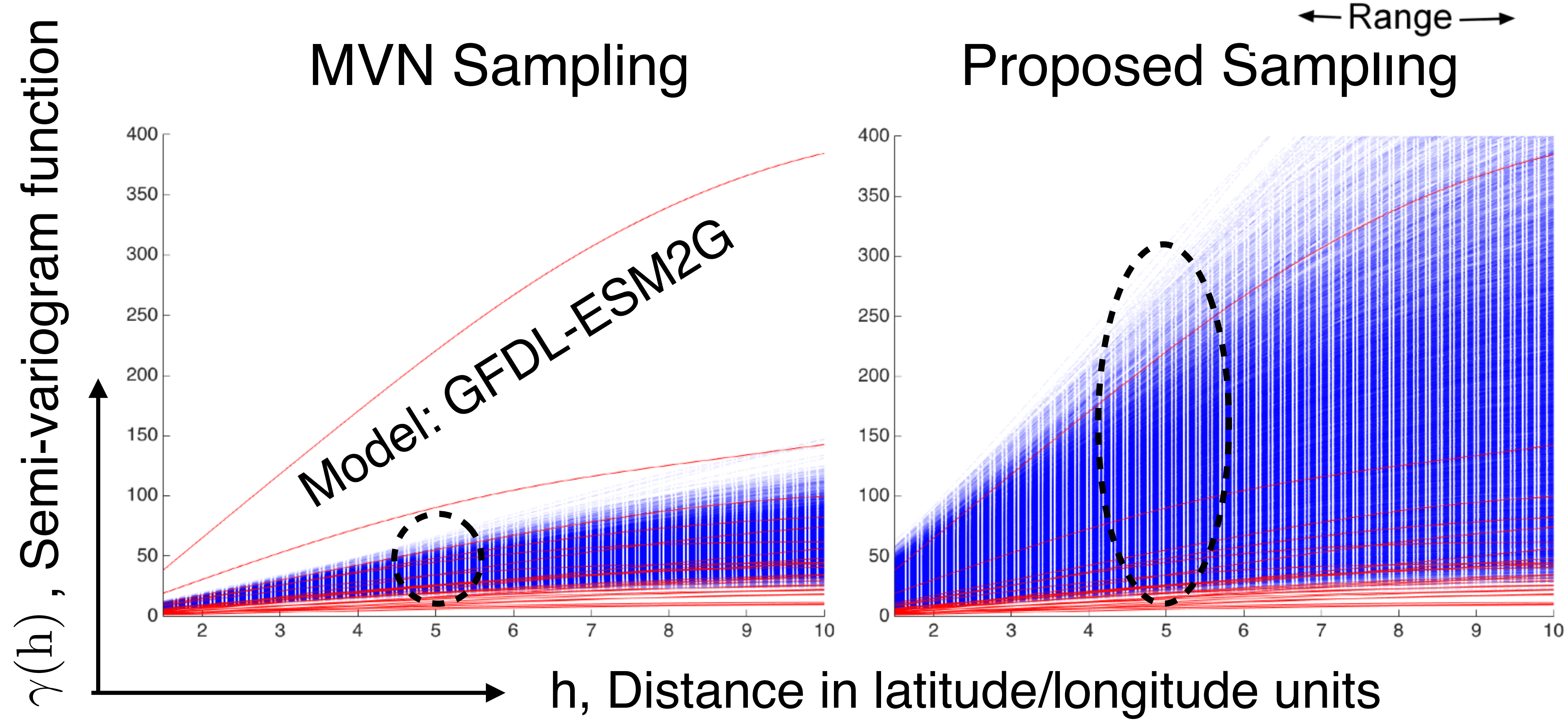
$\Lambda \leftarrow$ Covariance of CMO covariance matrices



Evaluation: Semi-variogram Plots

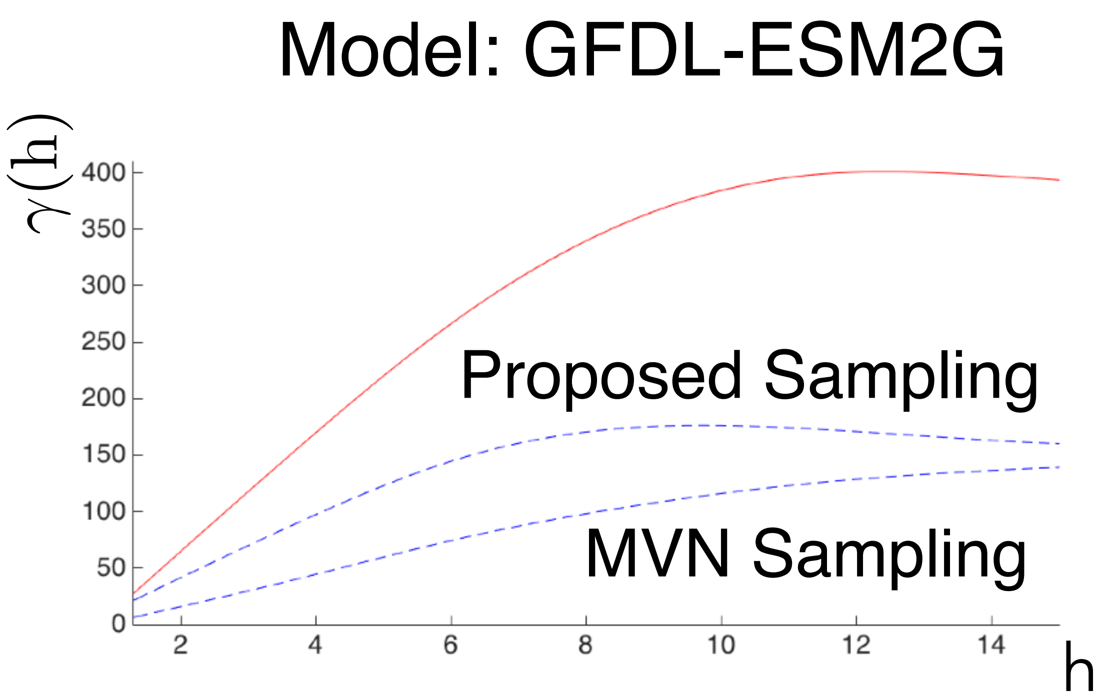
- Climate output models from the CMIP5-RCP2.6 ensemble
- Samples
- Spread

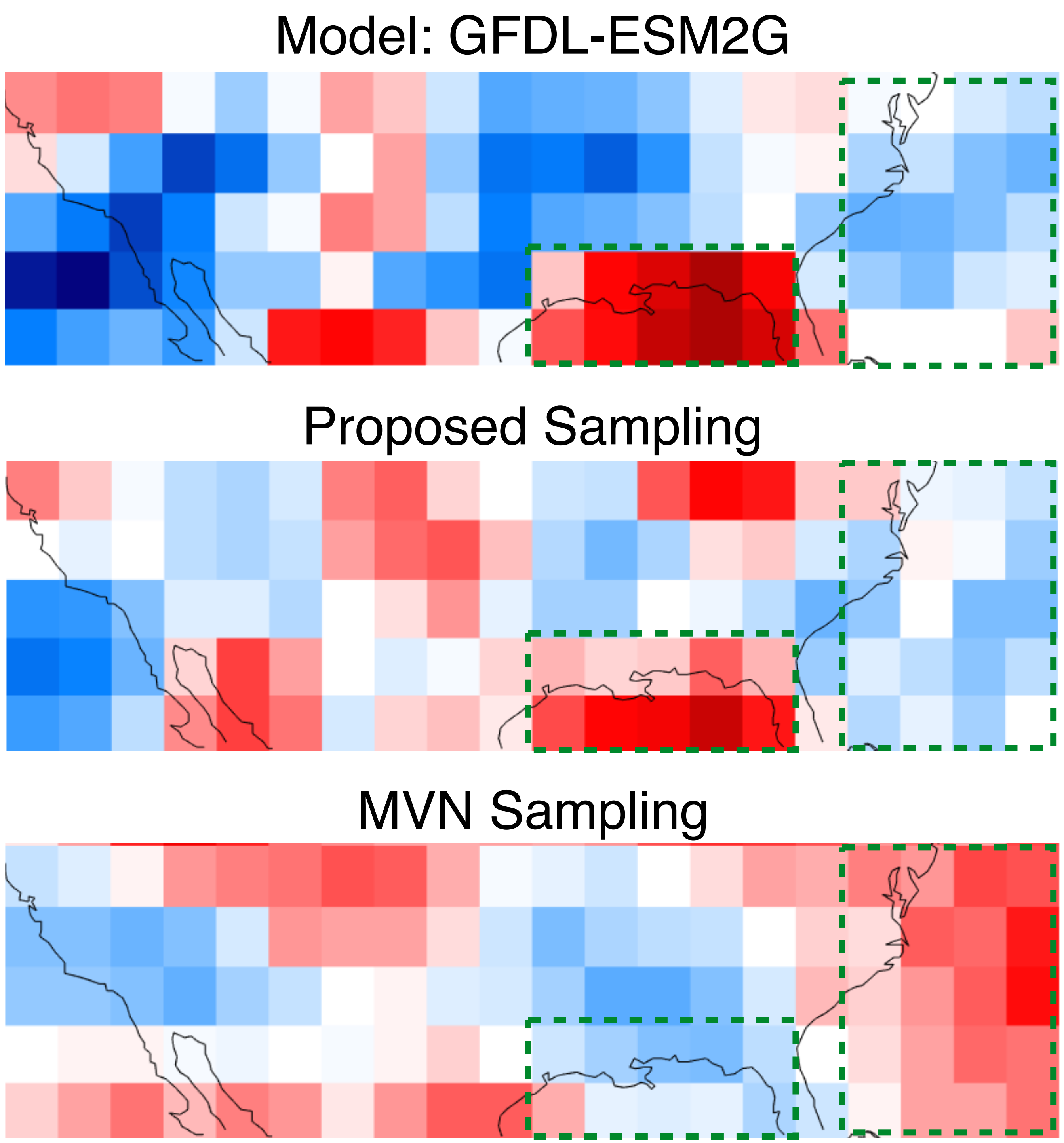




Dalal et al. (2016)
(Climate Informatics, **Best paper award**)

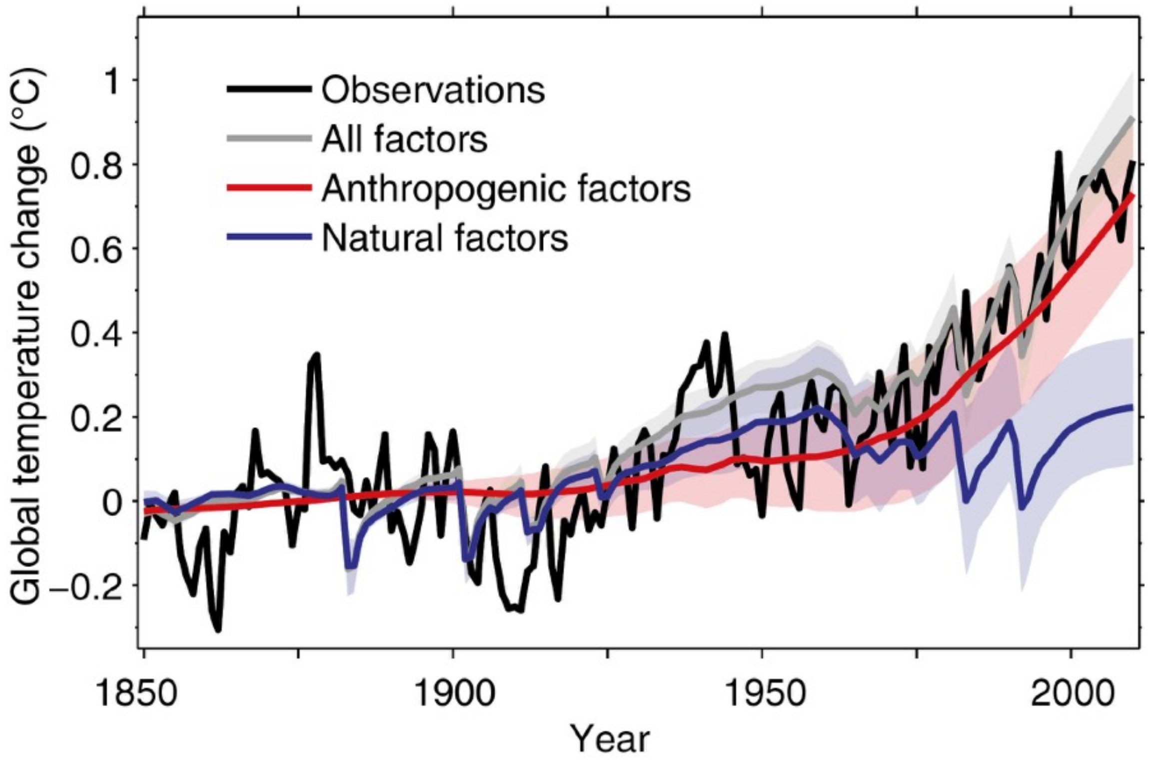
Evaluation: Spatial Field





Conclusion

- **Scientific Goal** : Reduce the climate change's projection uncertainty

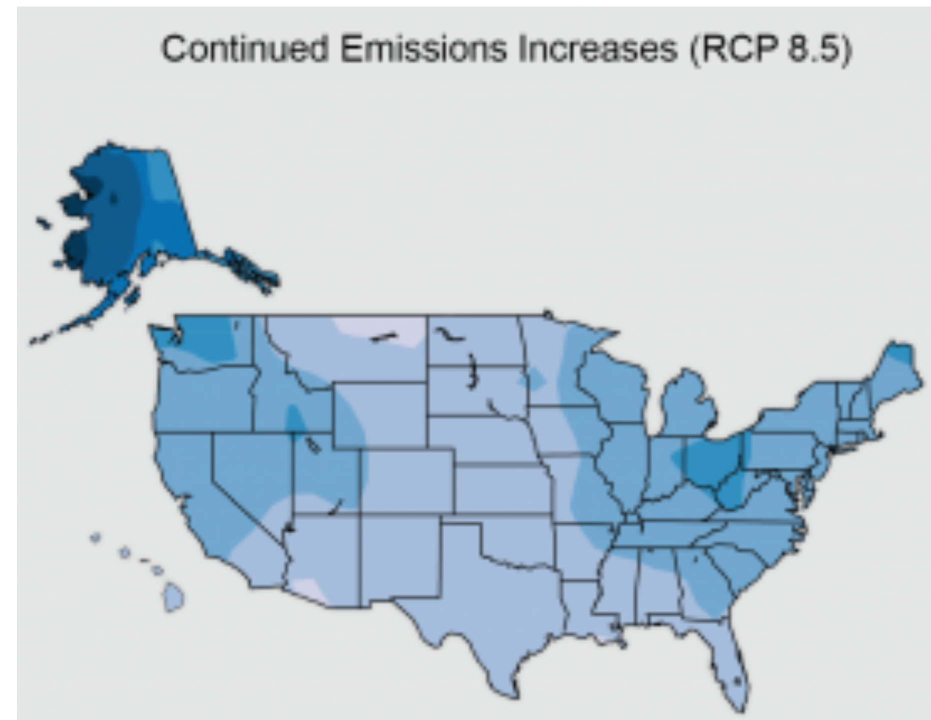


Aim: By developing geostatistical models for geometric structures

1. Improved parameter modeling
2. Improved parameter estimation
3. Leveraging multiple sources of information

Conclusion

- **Scientific Goal** : Emulate future climate change scenarios



Aim: By developing geostatistical models for geometric structures

1. Improved comparison of existing models that provide future scenarios.
2. Emulator does not need a supercomputer and few months; but just a laptop and few hours.

Acknowledgements

- U.S. Department of Homeland Security under Grant Award Number 2012-ST-104-000044 (CCICADA Fellowship)
- U.S. Department of Education (GAANN Fellowship)
- Computational Biomedicine Imaging and Modeling Center
- IMAGe at National Center for Atmospheric Research
- Thesis Committee Members
- CBIM & NCAR Colleagues

Thank you

Extra