Scalable Gaussian Processes on Physically Constrained Domains

Spatial Modelling with Intrinsic Geometry

Bora Jin,
Amy H. Herring, David Dunson
Duke University

Duke

Constrained Domains

- Groundwater

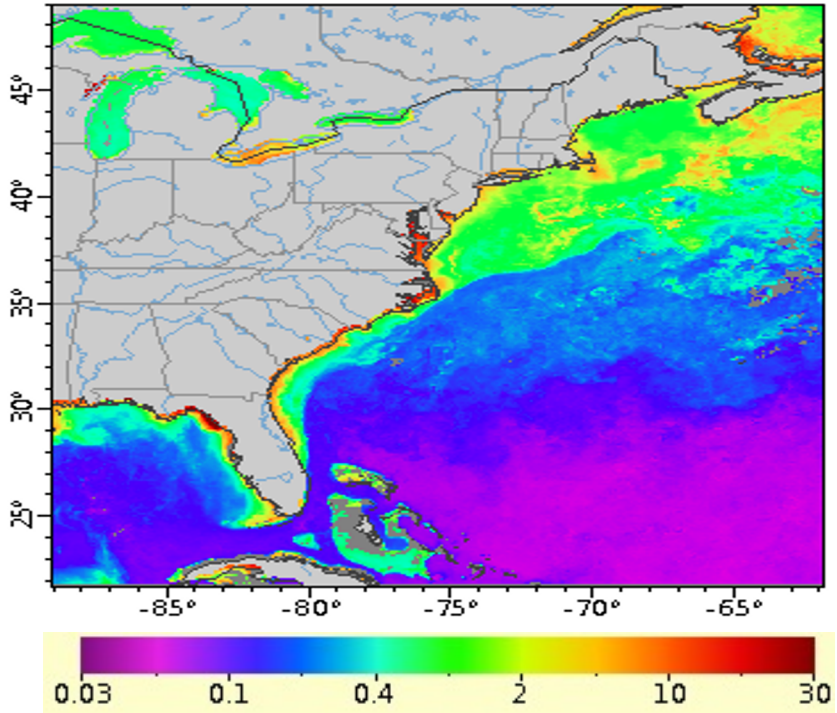
Duke

Constrained Domains

- Groundwater
- Neighborhood/wildlife disconnected by barriers

Duke

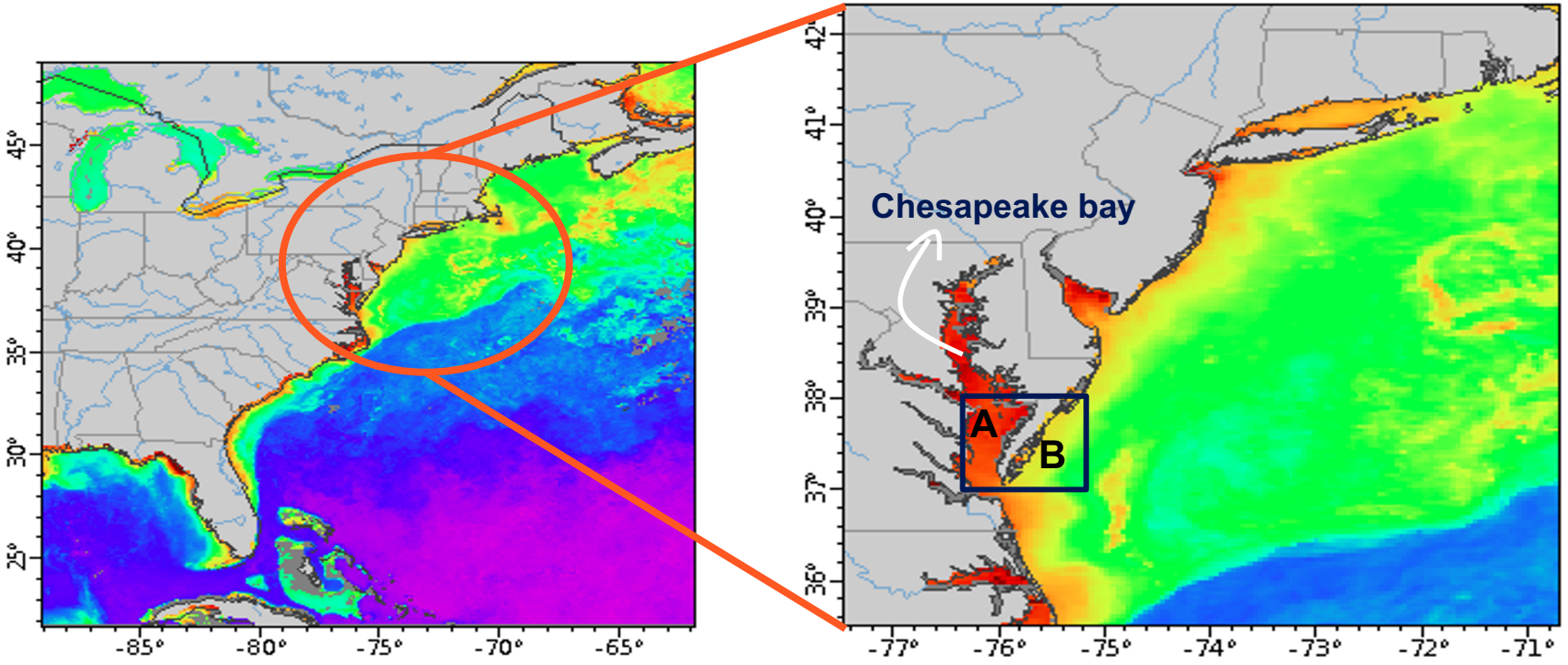
Constrained Domains



- Groundwater
- Neighborhood/wildlife disconnected by barriers
- Coastlines

Duke

Constrained Domains



Duke

https://coastwatch.pfeg.noaa.gov/erddap/griddap/erdMH1chlamday.graph

BORA-GP

Duke

Background

Goals

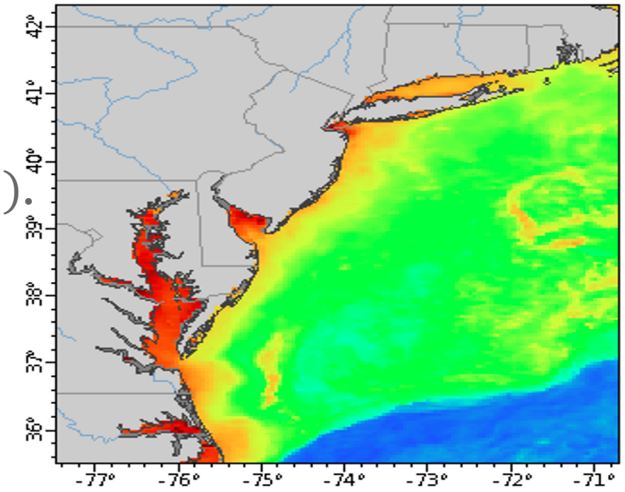
Methods

Analysis

Conclusions

Background

- Many applications with unique geometry including boundaries or barriers.
- Traditional spatial Gaussian process (GP) models ignore the unique geometry of the domain.
  - Inappropriate smoothing over physical barriers
  - Likely produce sub-optimal results
- Traditional GP has high computational cost $\mathcal{O}(n^3)$.
  - Prohibitive already if $n > 50{,}000$.
  - Recent & active development of scalable GP models whose computational cost grows linearly with n.



Duke

Goals

- Aim to construct a new **scalable** GP model that incorporates **constrained domains**.
- In particular, we want the new GP to
  - Produce physically sensible kriging,
  - Induce physically sensible covariance behavior,
  - Mimic geodesic distance-based covariance without loss of scalability.
    - Typical geodesic distance estimation is also an expensive operation, $\mathcal{O}(n^3)$.

Duke

Methods

- Bayesian hierarchical spatial regression

$$y = X\beta + w + \epsilon, \qquad \epsilon \sim N(0, \tau^2 I_n)$$
$$w \sim GP(0, C(\cdot,\cdot\,|\theta))$$

  - Computational bottleneck in computing the joint density of $w_{obs} = (w_1, \cdots, w_n)^T$

$$p(w_1, \cdots, w_n|\theta) = N(0, C_\theta) \propto |C_\theta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} {w_{obs}}^T C_\theta^{-1} w_{obs}\right)$$

Costly! $\mathcal{O}(n^3)$ Costly! $\mathcal{O}(n^3)$

    - $w_i = w(s_i)$
    - n is the number of spatial locations.

Duke

Methods

- Vecchia (1988)
  - Approximate the joint density with lower-dimensional conditional densities to ease computational burden.

$$p(w_1, \cdots, w_n) = \prod_{i=1}^{n} p(w_i | w_1, \cdots, w_{i-1})$$

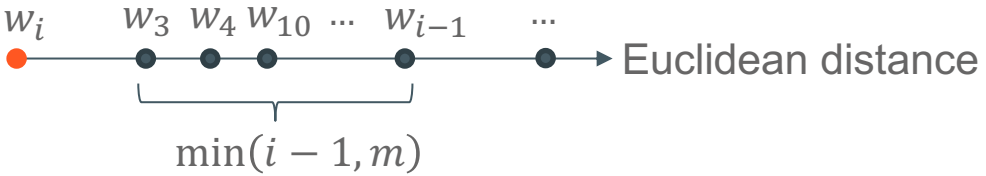
  - Lots of redundant information in the conditioning set $\{w_1, \cdots, w_{i-1}\}$ for large i's.
  - $p(w_i | w_1, \cdots, w_{i-1}) \approx p(w_i | w_{im})$ where w_{im} is a conditioning set of **at most** m $(\ll n)$ elements from $\{w_1, \cdots, w_{i-1}\}$.

$$p(w_1, \cdots, w_n) \approx \tilde{p}(w_1, \cdots, w_n) = \prod_{i=1}^{n} p(w_i | w_{im})$$

Duke

Methods

- Vecchia (1988) to Nearest Neighbor Gaussian Process (NNGP) (2016)
  - One possible choice of w_{im}



  - $w_{im} = \{w_1, w_2, w_3, \cdots, w_{i-2}, w_{i-1}\}$

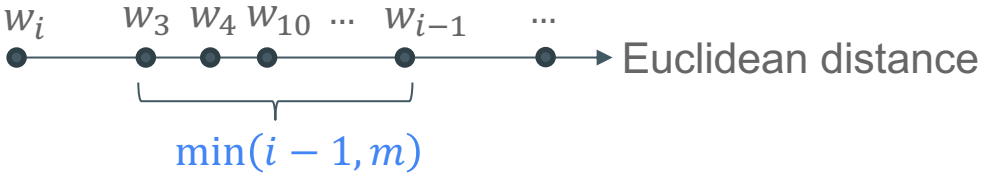
➡ conditional independence

➡ zeros in a precision matrix – sparsity

Duke

Methods

- Vecchia (1988) to **N**earest **N**eighbor **G**aussian **P**rocess (NNGP) (2016)
  - One possible choice of w_{im}



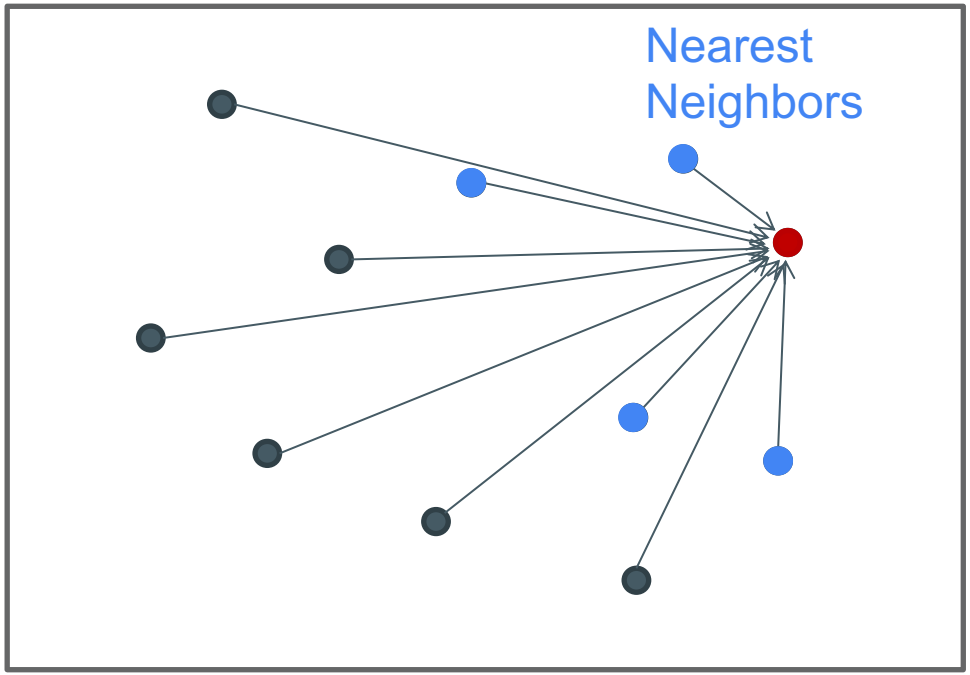
  - NNGP extends finite-dimensional sparsity to sparsity-inducing spatial processes.
  - Constructs a scalable and valid spatial process based on nearest neighbors.
  - Empirically shows that the above-mentioned w_{im} performs better than other alternatives in a very wide range of scenarios.

Duke

Methods

- Visualization of sparsity using directed acyclic graphs (DAGs)

$m = 4$

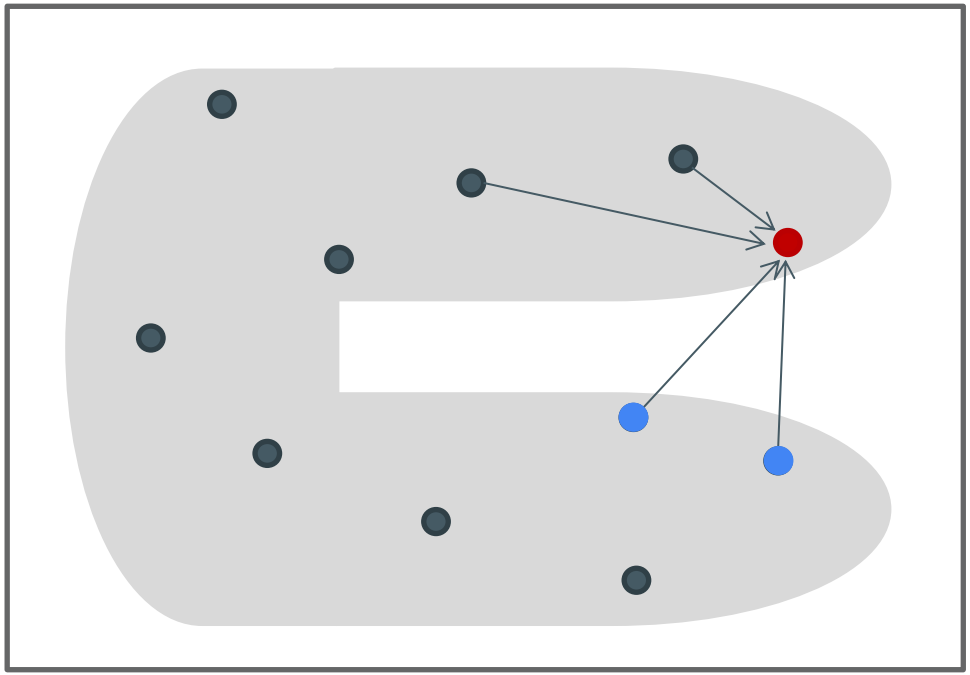


Duke

Methods

- What if we know that our measurements lie in a constrained domain?

$m = 4$

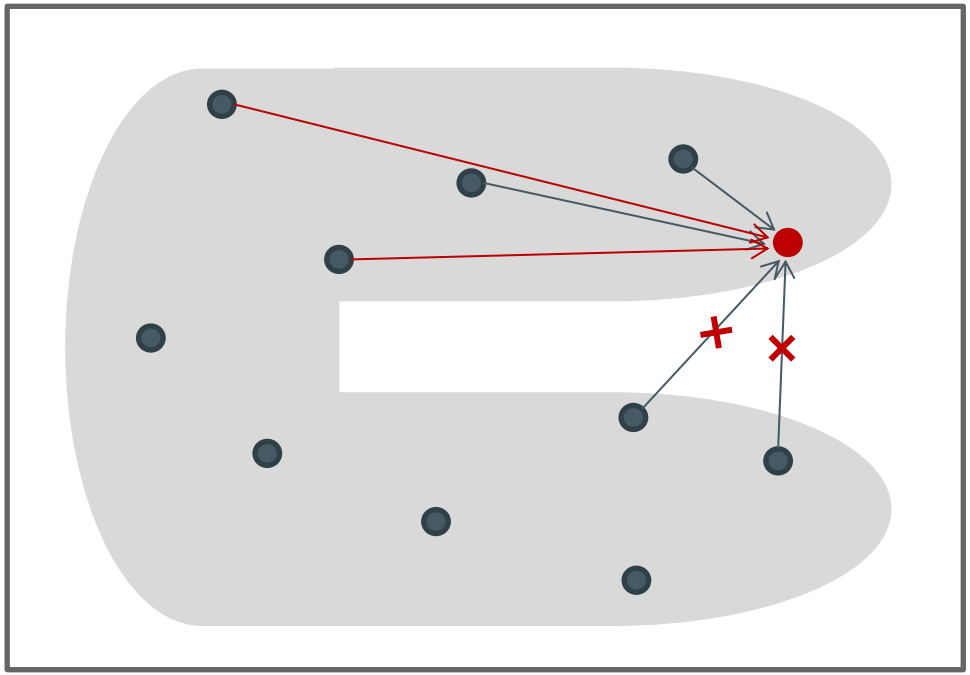


Duke

Methods

- Barrier Overlap-Removal Acyclic Directed Graph Gaussian Process (BORA-GP)

$m = 4$



Duke

Methods

- BORA-GP
  1. Specify a multivariate normal distribution over a fixed finite set $S = \{s_1, \cdots, s_k\} \subset \mathcal{D}$

$$w_S = \left(w(s_1), \cdots, w(s_k)\right)^T \sim N\left(0, \tilde{C}_S\right) = \prod_{i=1}^{k} N(w(s_i); H_{s_i} w_{N(s_i)}, R_{s_i})$$

  - $H_{s_i} = C_{s_i, N(s_i)} C_{N(s_i)}^{-1}$, $R_{s_i} = C_{s_i} - C_{s_i, N(s_i)} C_{N(s_i)}^{-1} C_{N(s_i), s_i}$
  - C: base covariance function
  - $N(s_i)$: set of spatial locations in S whose straight line to s_i does not overlap barriers (physically sensible neighbors of s_i of size $\min(i, m)$)

Duke

Methods

- BORA-GP

  2. Extend it to the whole domain $\mathcal{D}$

$$\forall s \in \mathcal{D} \setminus S, \qquad w(s) \sim N(H_s w_{N(s)}, R_s)$$

  - $N(s) \subset S$ physically sensible neighbors of size m

  ➡ 1 & 2 define a valid scalable process

$$w \sim \text{BORA–GP}(0, \tilde{C}(\cdot,\cdot\,|\theta))$$

  - $\tilde{C}$ is non-stationary.

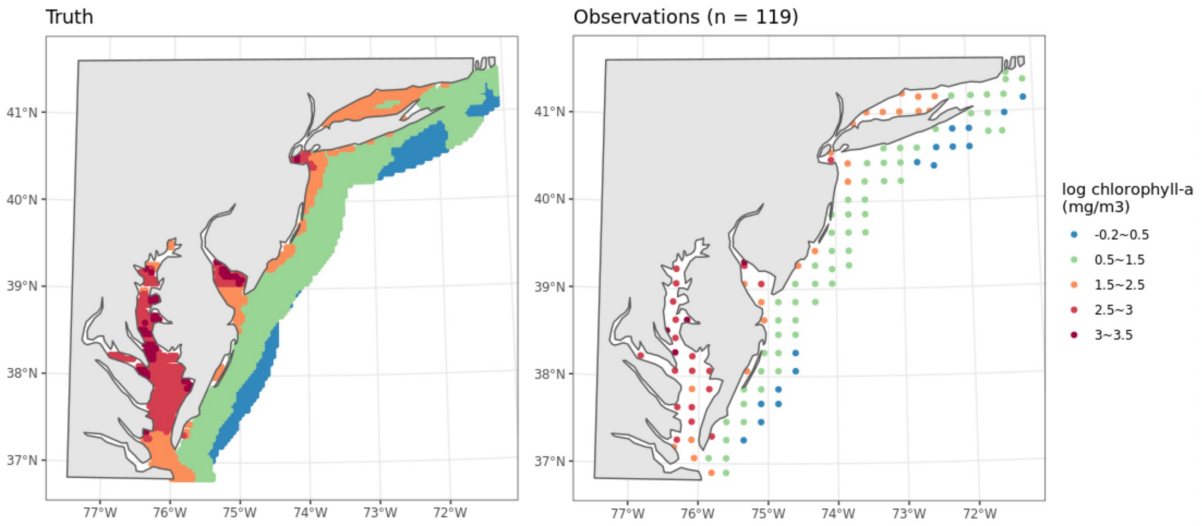
Duke

Analysis

- Univariate Bayesian spatial regression

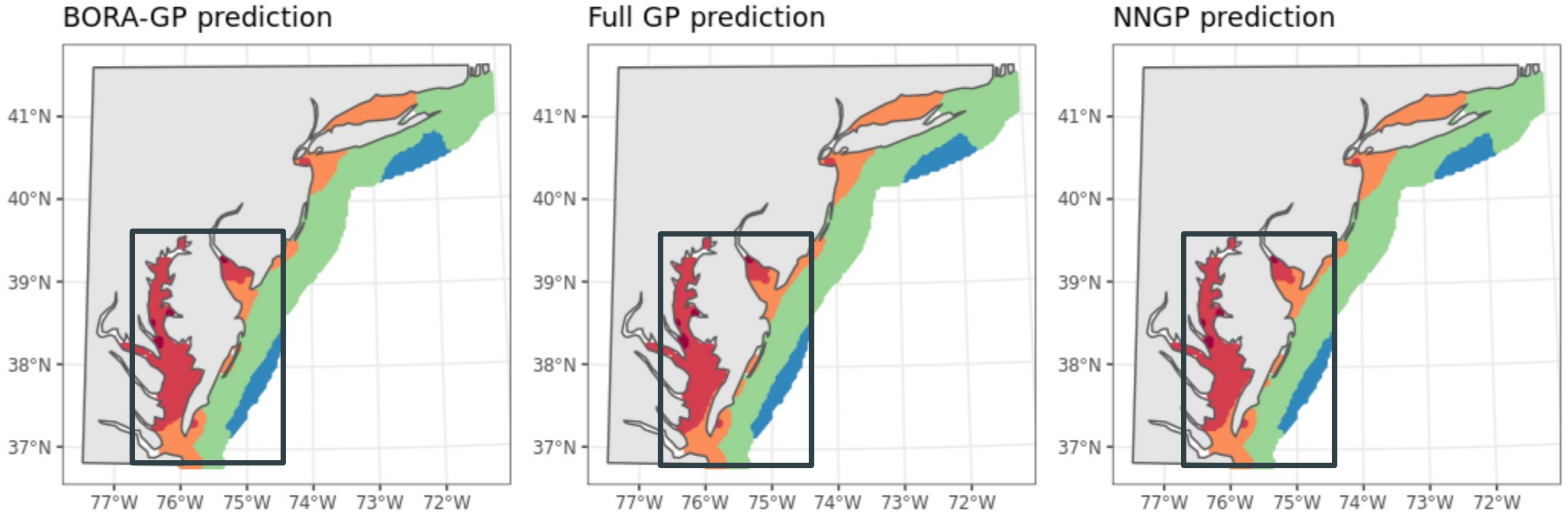
$$y = w + \epsilon,\ w \sim \text{BORA-GP}(0, \tilde{C}(\cdot,\cdot\,|\theta))$$

- y: log chlorophyll-a level (mg/m3) of March 2021
- Exponential covariance function as a base C
- Number of obs $n = 119$
- 2791 new spatial locations to predict
- $m = 10$



Duke

Analysis

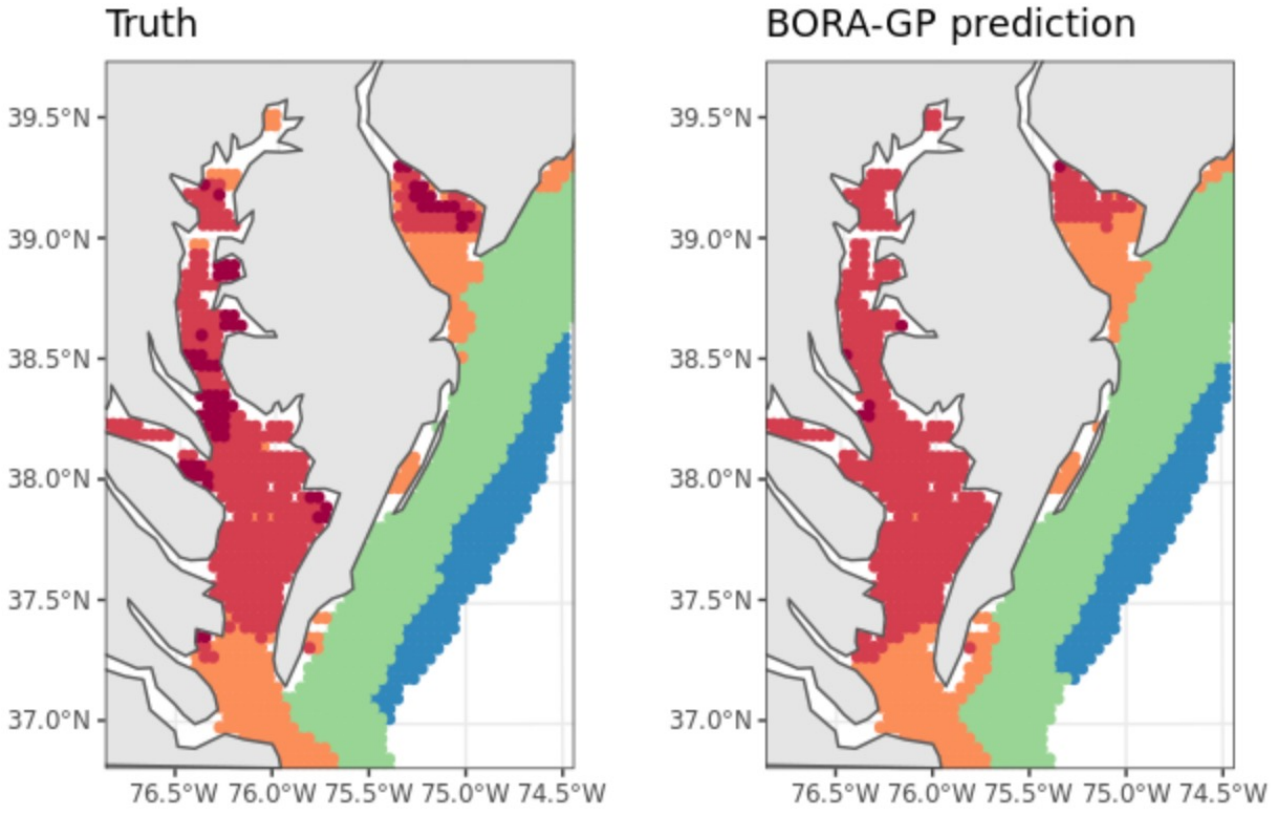


	BORA-GP	Full GP	NNGP
RMSPE*	**0.165**	0.172	0.178
95% CI coverage	98.57%	**98.42%**	98.42%
Mean 95% CI width	**1.110**	1.113	1.172

*RMSPE: Root Mean Squared Prediction Error

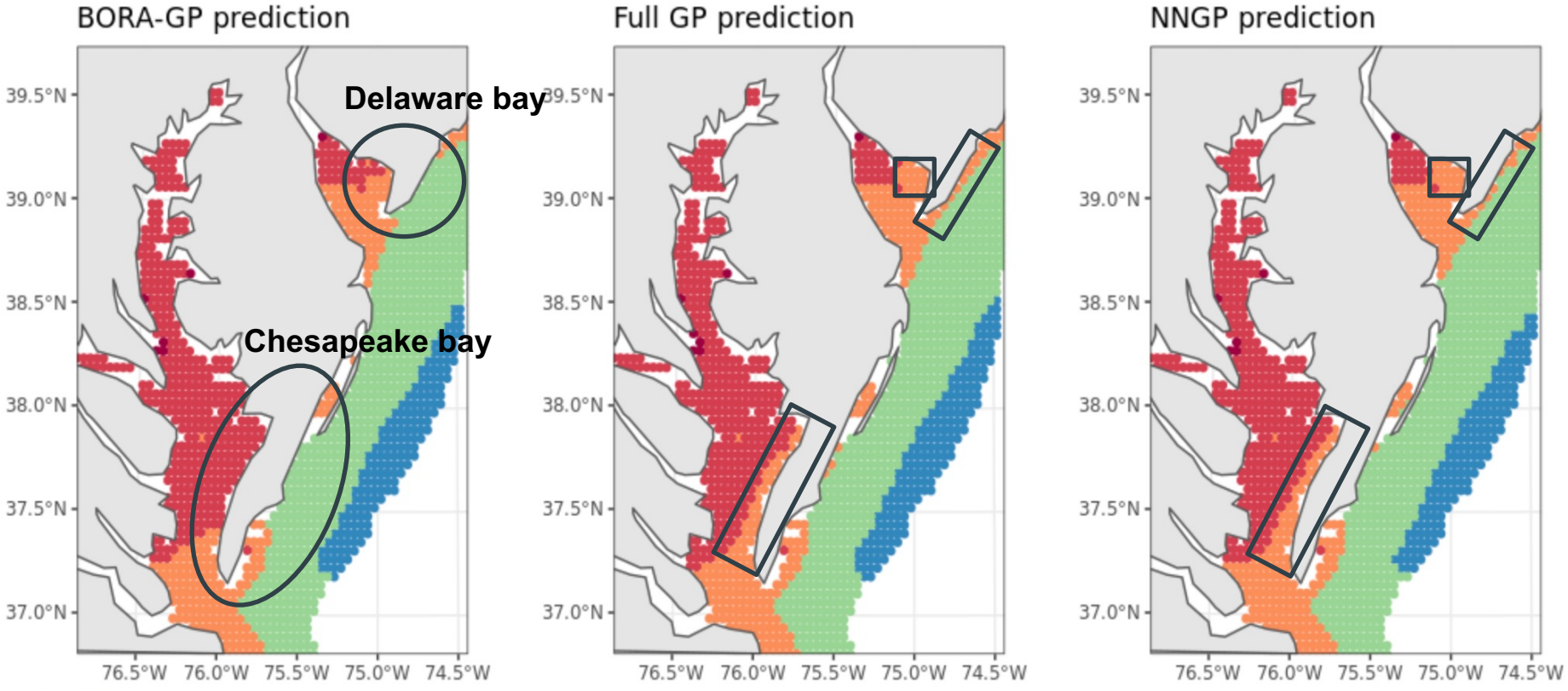
Duke

Analysis



Duke

Analysis


BORA-GP prediction
Full GP prediction
NNGP prediction
Delaware bay
Chesapeake bay
39.5°N
39.0°N
38.5°N
38.0°N
37.5°N
37.0°N
76.5°W
76.0°W
75.5°W
75.0°W
74.5°W

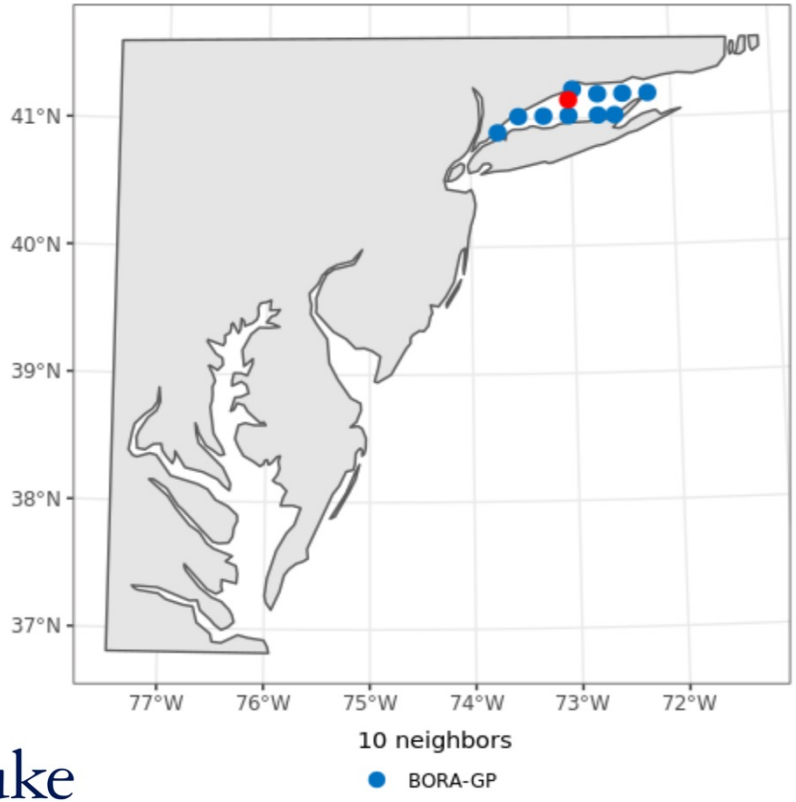
Analysis

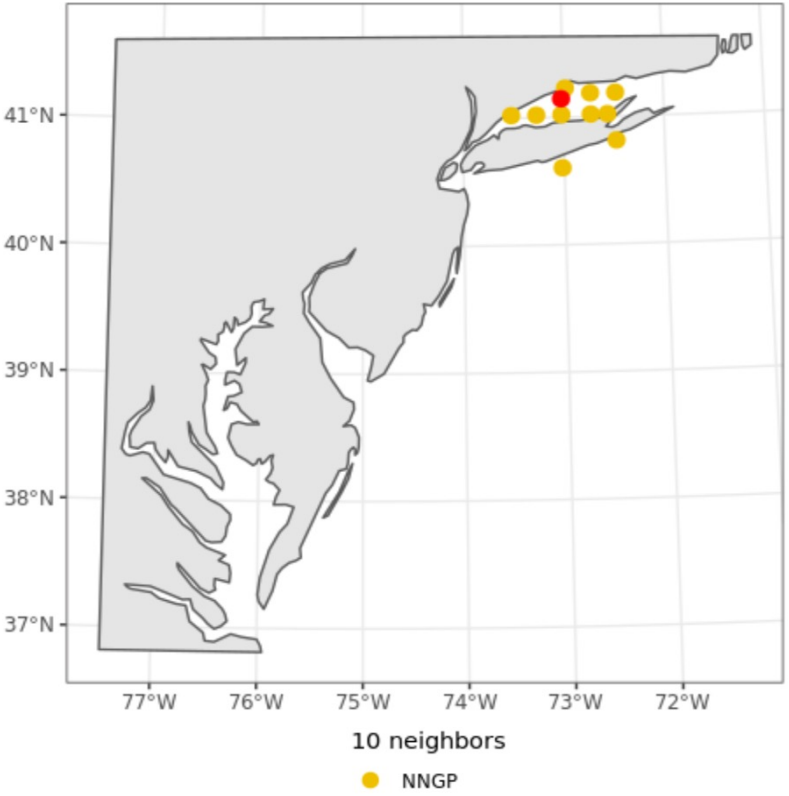
Delaware Bay	BORA-GP	Full GP	NNGP
RMSPE	**0.257**	0.346	0.356
95% CI coverage	**92.98%**	85.96%	87.72%
Mean 95% CI width	1.171	**1.150**	1.210

Chesapeake Bay	BORA-GP	Full GP	NNGP
RMSPE	**0.214**	0.296	0.300
95% CI coverage	98.18%	**95.45%**	95.45%
Mean 95% CI width	1.157	**1.131**	1.192

Duke

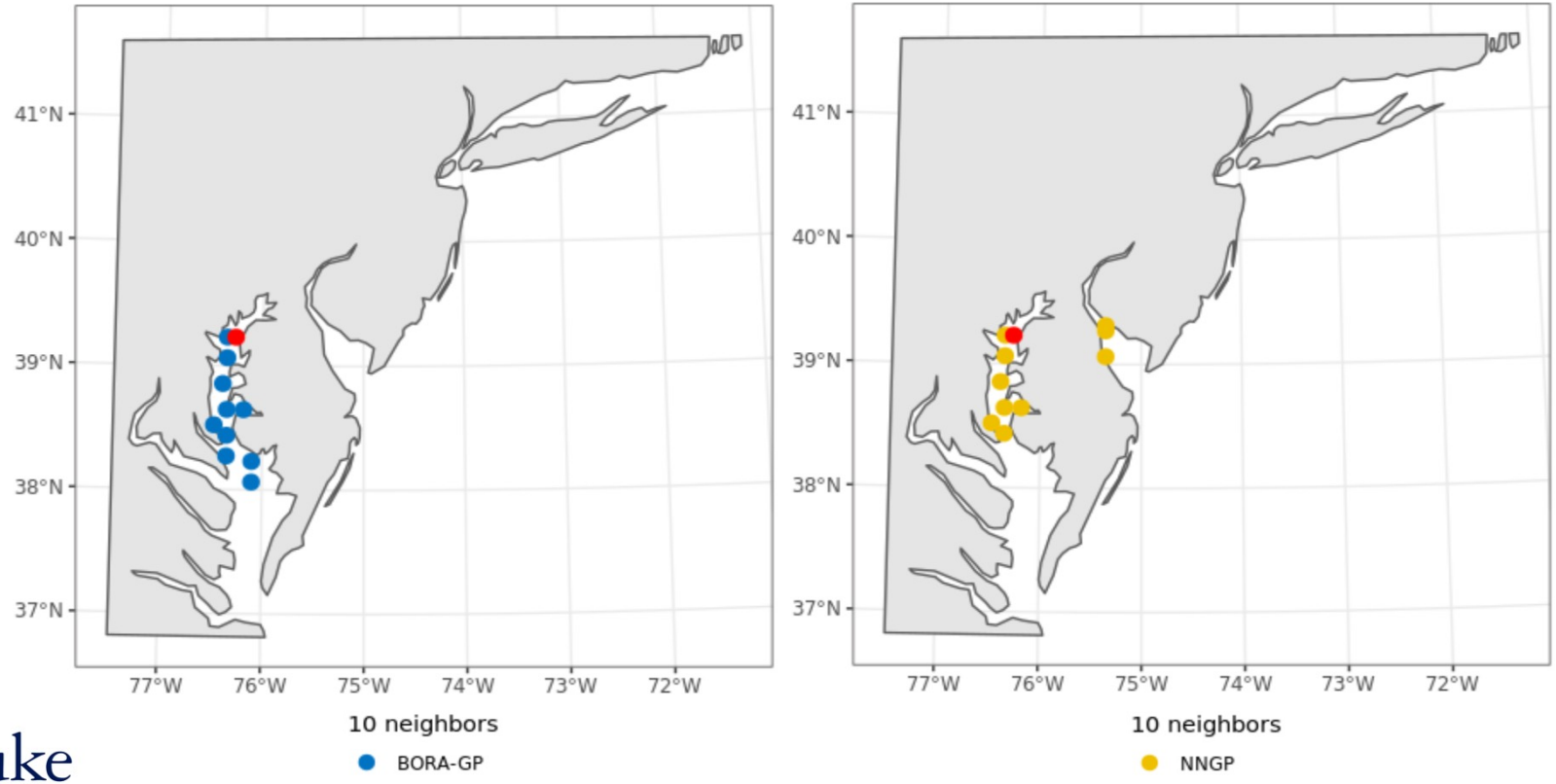
Analysis





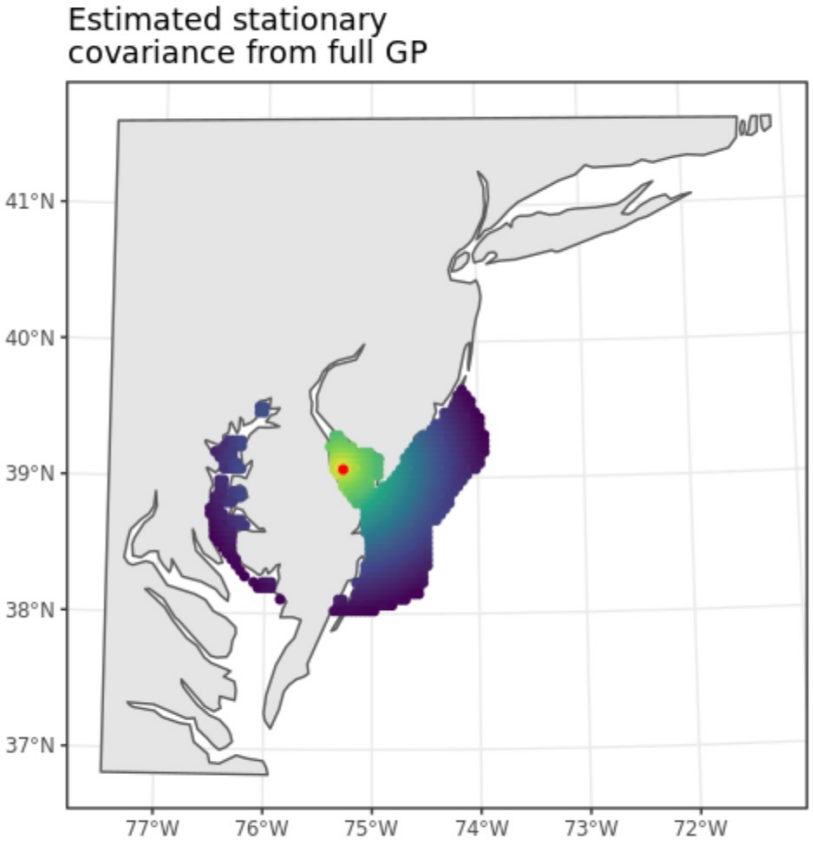
Duke

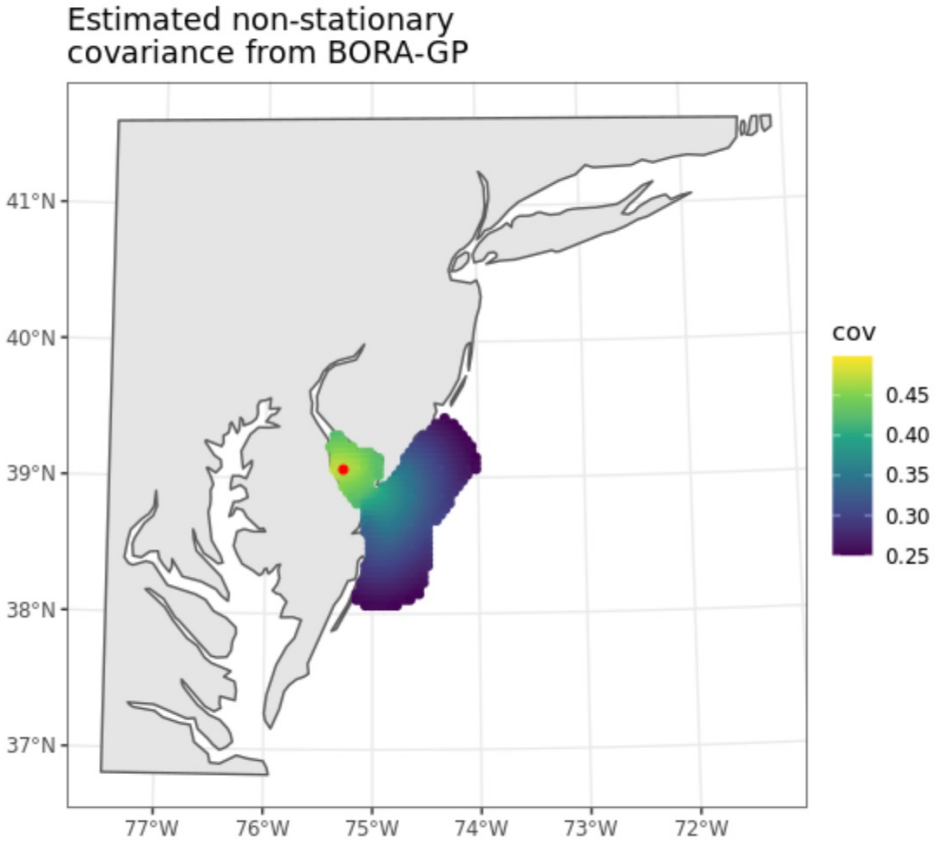
Analysis



Duke

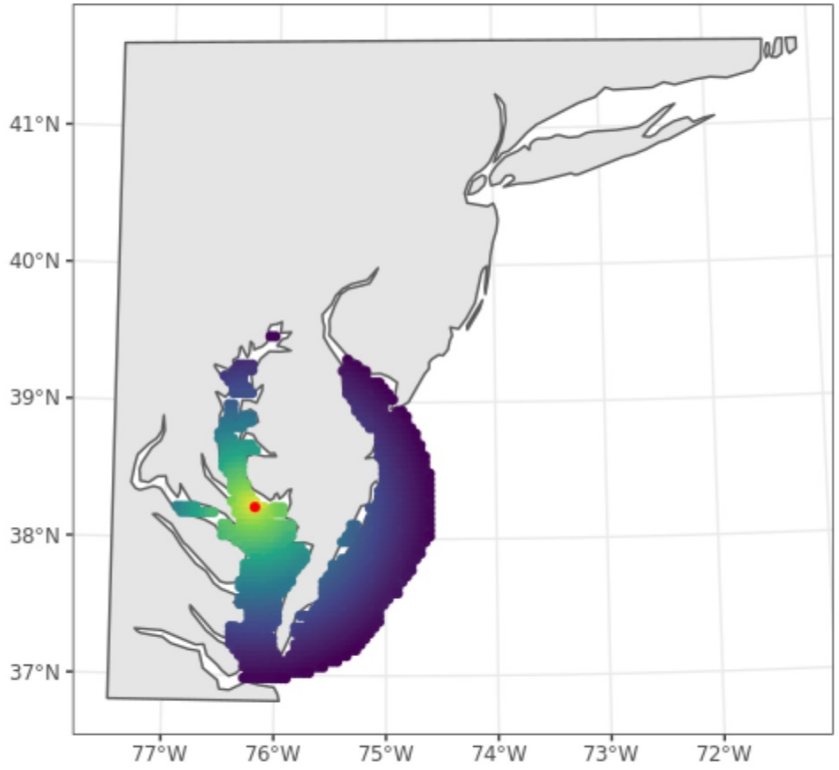
Analysis


Estimated stationary
covariance from full GP
41°N
40°N
39°N
38°N
37°N
77°W
76°W
75°W
74°W
73°W
72°W

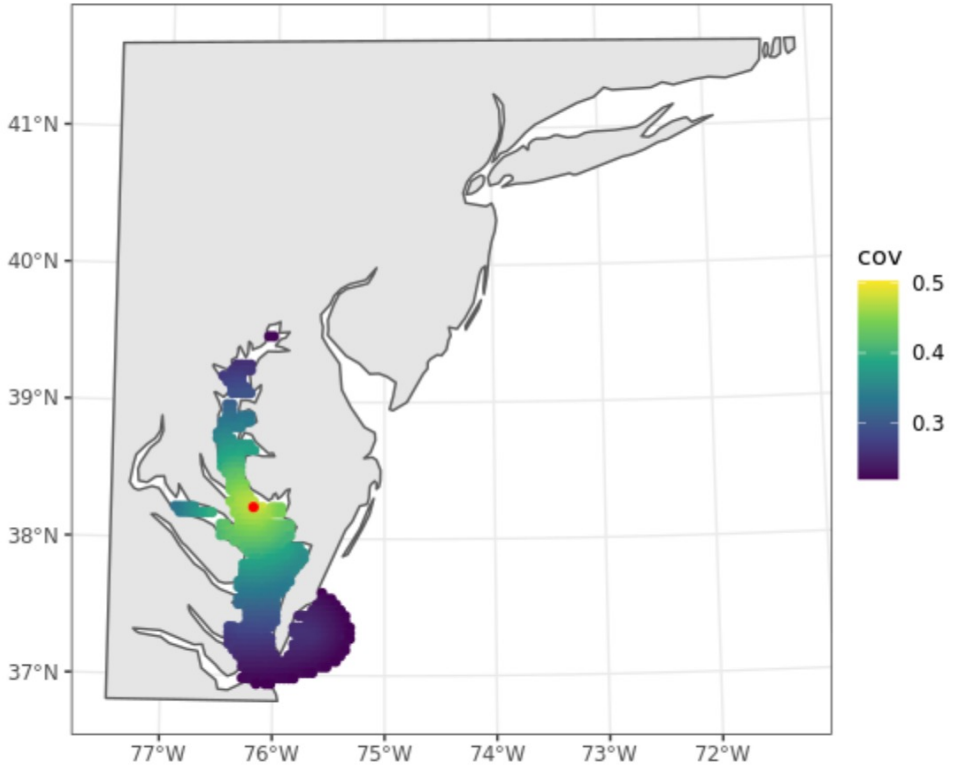

Estimated non-stationary
covariance from BORA-GP
41°N
40°N
39°N
38°N
37°N
77°W
76°W
75°W
74°W
73°W
72°W
cov
0.45
0.40
0.35
0.30
0.25

Analysis

Estimated stationary covariance from full GP



Estimated non-stationary covariance from BORA-GP



Conclusions

- Does BORA-GP achieve all the goals?
  - ✓ Scalable
    $m \ll n$ neighbors only (Sparsity-inducing DAG)
  - ✓ Account for the domain
    - ✓ Physically sensible kriging
      Using physically sensible neighbors
    - ✓ Physically sensible covariance behavior
      The implied non-stationary covariance does not go beyond barriers
    - ✓ Mimicking geodesic distance-based covariance without loss of scalability
      No need to estimate geodesic distance
      Covariance resembles water flow

Duke

References & Acknowledgement

- Datta, A., Banerjee, S., Finley, A.O., and Gelfand, A.E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, *111*(514), 800-812.
- Vecchia, A.V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *50*(2), 297-312.

- BORA-GP soon to be available in arXiv.

- Thanks, coauthors: Amy H. Herring and David Dunson!
- Thanks for the support from NIEHS R01ES027498 and R01ES028804!

Duke