Scalable Gaussian Processes on Physically Constrained Domains

Spatial Modelling with Intrinsic Geometry

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

Duke

At ISBA on June 28, 2022

Constrained Domains

- Groundwater

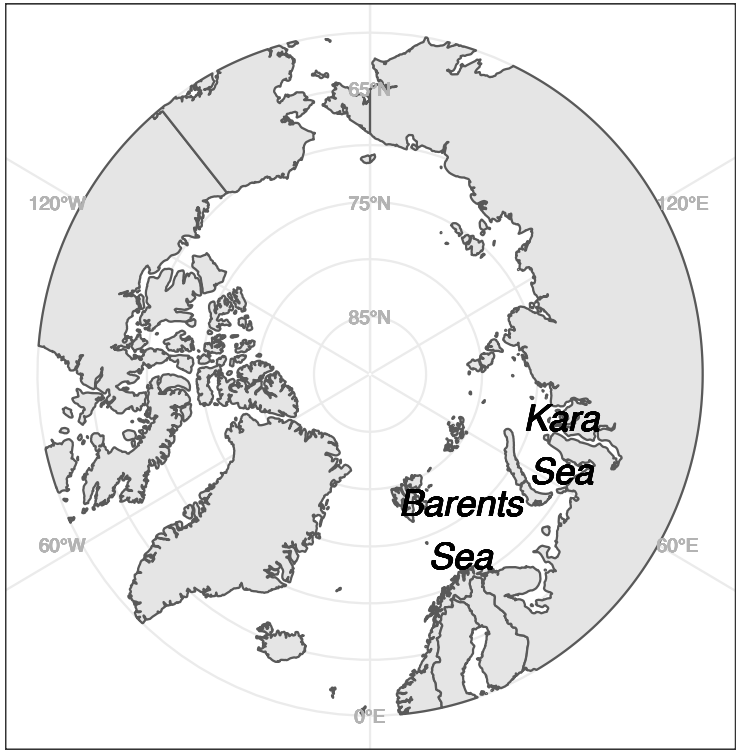
Duke

Constrained Domains

- Groundwater
- Neighborhood/wildlife disconnected by barriers

Duke

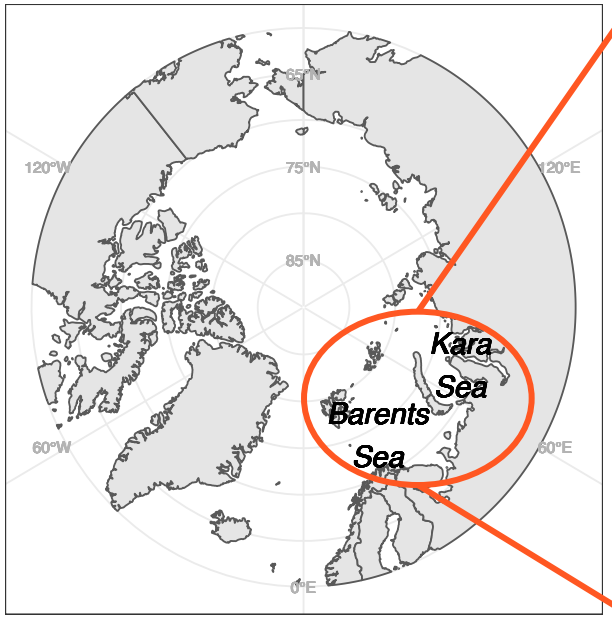
Constrained Domains



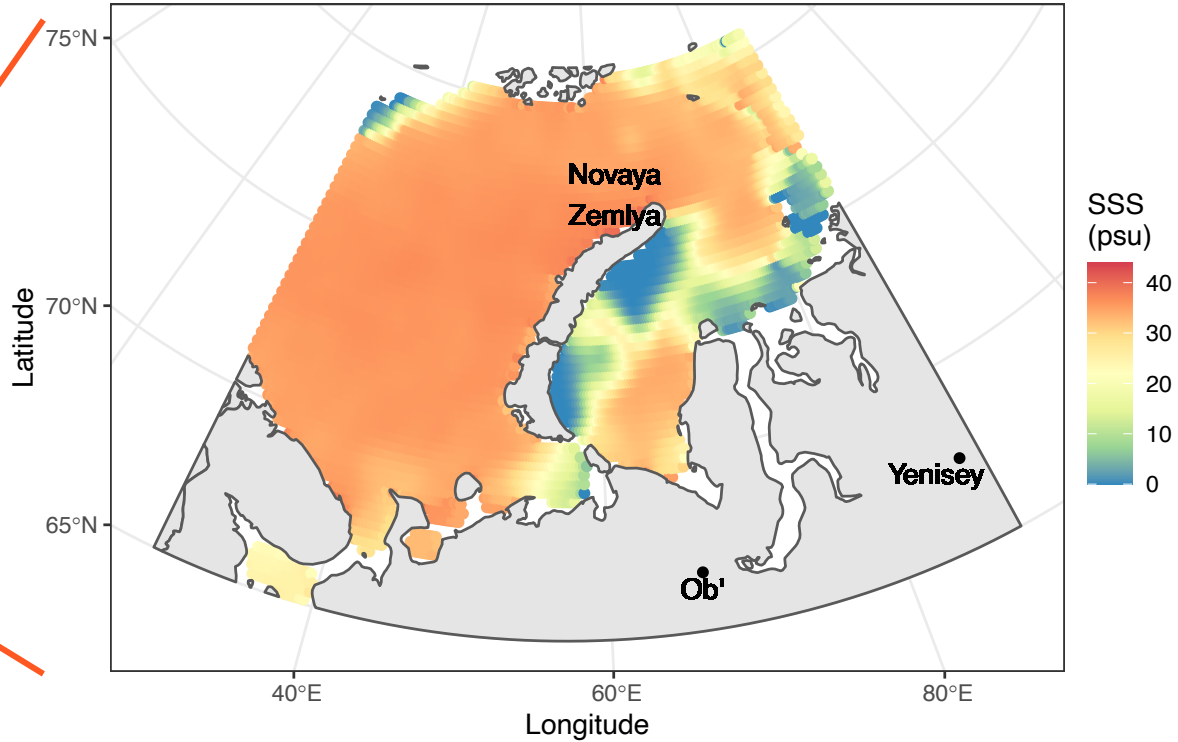
- Groundwater
- Neighborhood/wildlife disconnected by barriers
- Coastlines

Duke

Constrained Domains



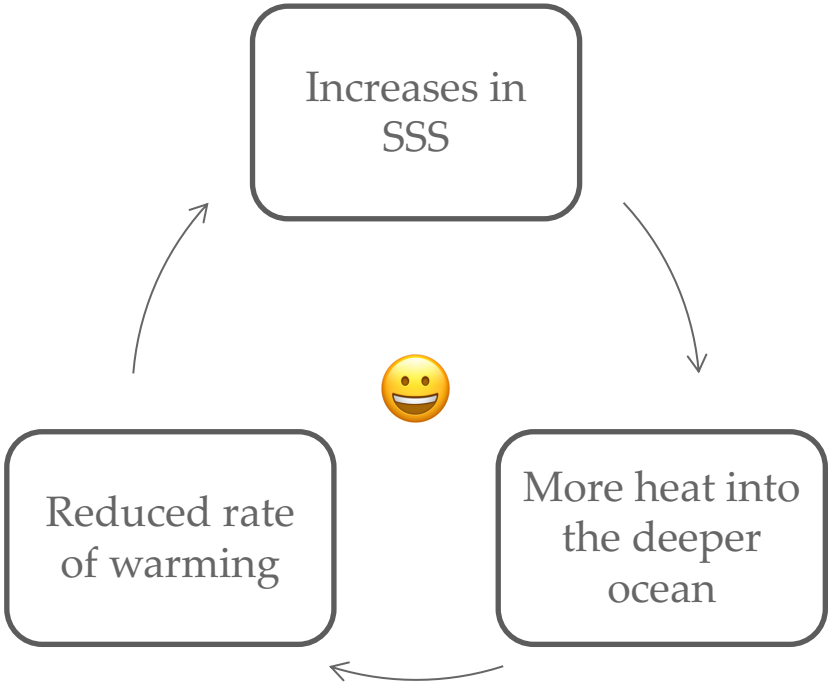
Sea Surface Salinity (SSS) over July 20 – July 31, 2017



Duke

Background

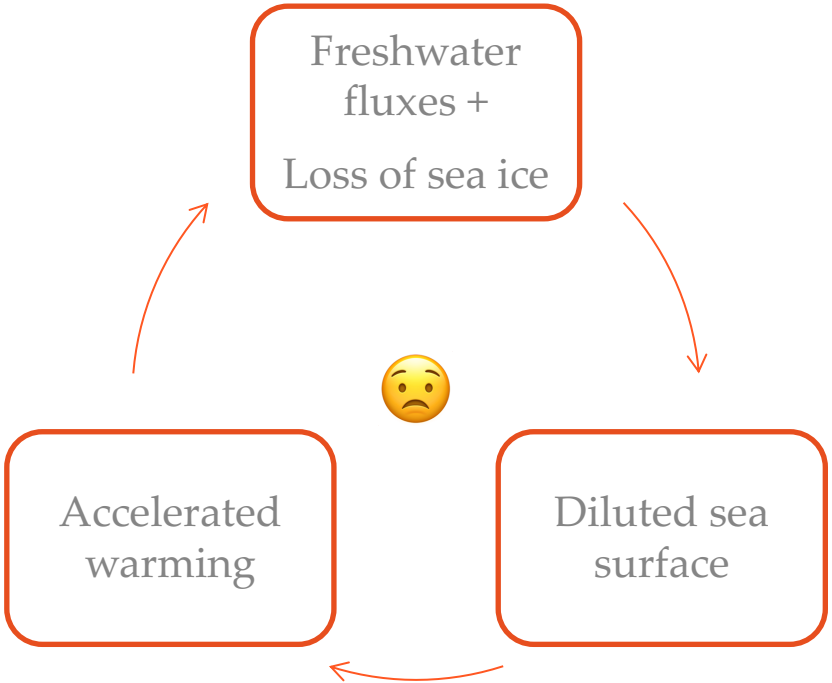
- SSS helps understand changes in the Arctic.



Duke

Background

- SSS helps understand changes in the Arctic.
- Kara Sea is one of the regions with a vulnerable feedback loop.
- Accurate prediction of SSS given unique domains is important to take actions before such regions fall into a vicious cycle of warming.

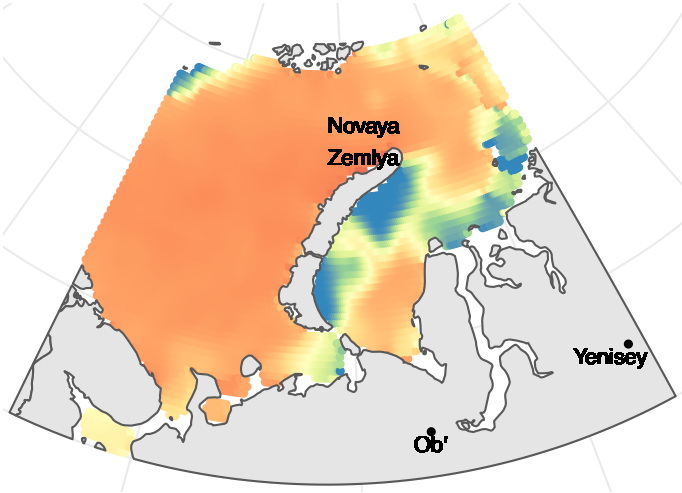


Duke

Background

Traditional spatial GP models

- Ignore the unique geometry of the domain.
  - Inappropriate smoothing over physical barriers
  - Likely produce sub-optimal results

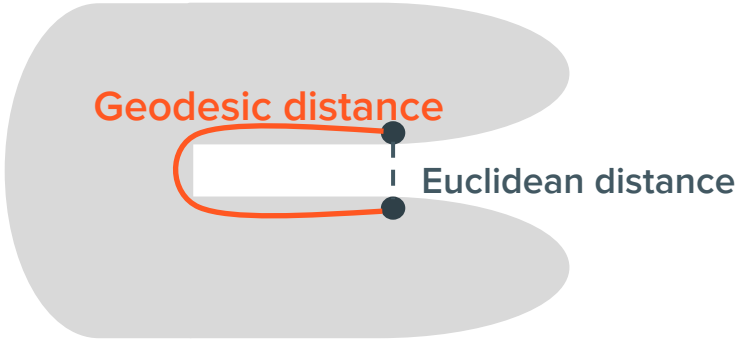


- Have high computational cost $\mathcal{O}(n^3)$.
  - Prohibitive already if $n > 50{,}000$.
  - Recent & active development of scalable GP models whose computational cost grows linearly with n
    e.g., Vecchia (1988); Nearest Neighbor GP (NNGP) – Datta et al. (2016)

Duke

Literature

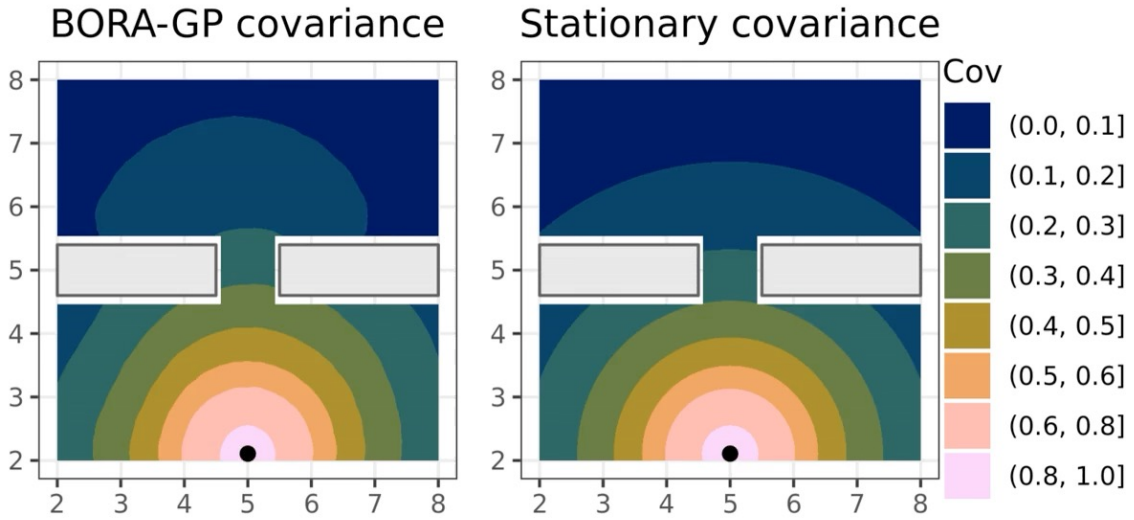
- Spline regression with partial differential equation (PDE) regularizations: Ramsay (2002); Wood et al. (2008); Sangalli et al. (2013)
- Based on stochastic PDEs (SPDEs): Bakka et al. (2019)
- Heat kernel-based GP approximations: Niu et al. (2019); Dunson et al. (2021)
- Geodesic distance in a covariance function of a scalable GP: Dai et al. (2021)



Duke

Goals

- Aim to construct a new **scalable** **nonstationary** GP that incorporates **constrained domains**.
  - Physically sensible kriging
  - Physically sensible covariance behaviors



Duke

Existing Scalable Methods

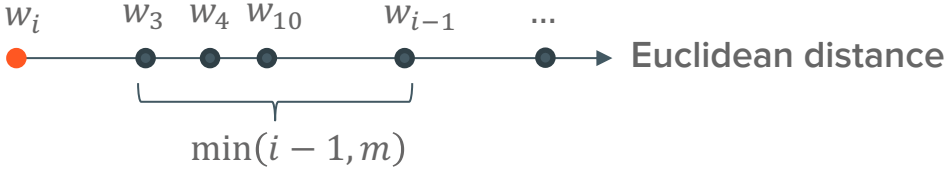
- Bayesian hierarchical spatial regression

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

  - Vecchia (1988) cheaply approximates the likelihood.

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

    - Nearest neighbors w_{im}:



    - $w_{im} = \{w_1, w_2, w_3, \cdots, w_{i-2}, w_{i-1}\}$ ➡ Sparse precision matrix ➡ Scalability

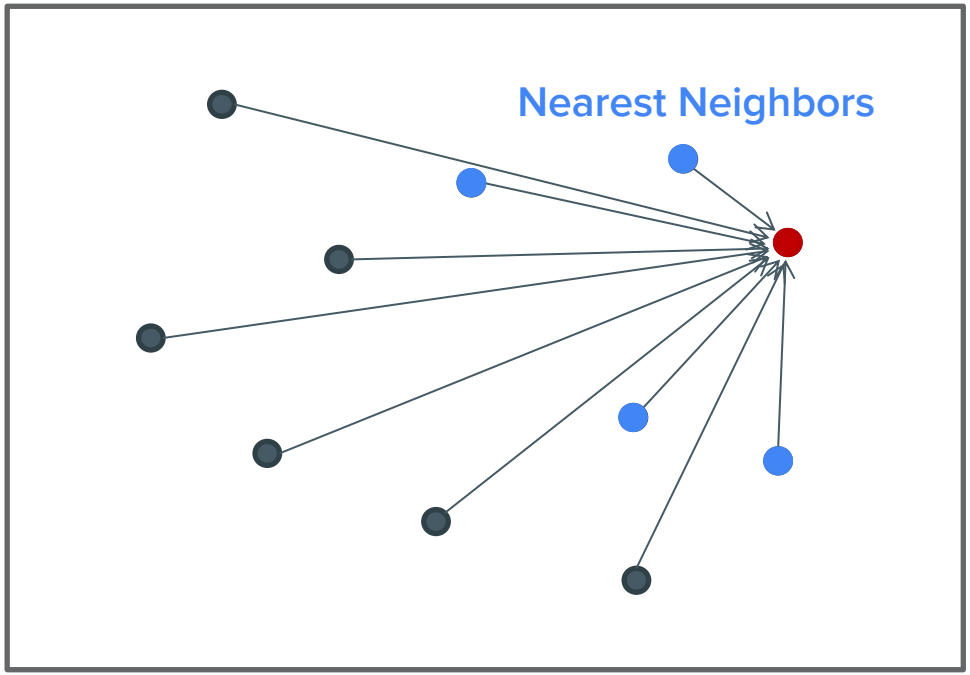
  - NNGP (Datta et al. 2016) extends it to a process based on directed acyclic graphs (DAG).

Duke

Existing Scalable Methods

- Visualization of sparsity using DAGs

$m = 4$

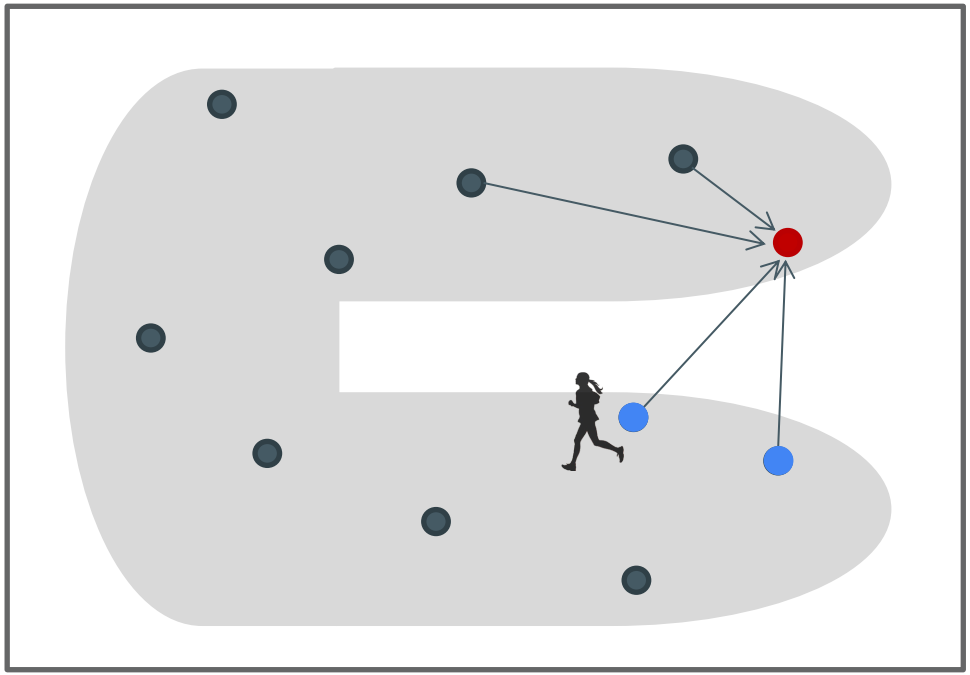


Duke

Existing Scalable Methods

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

$m = 4$

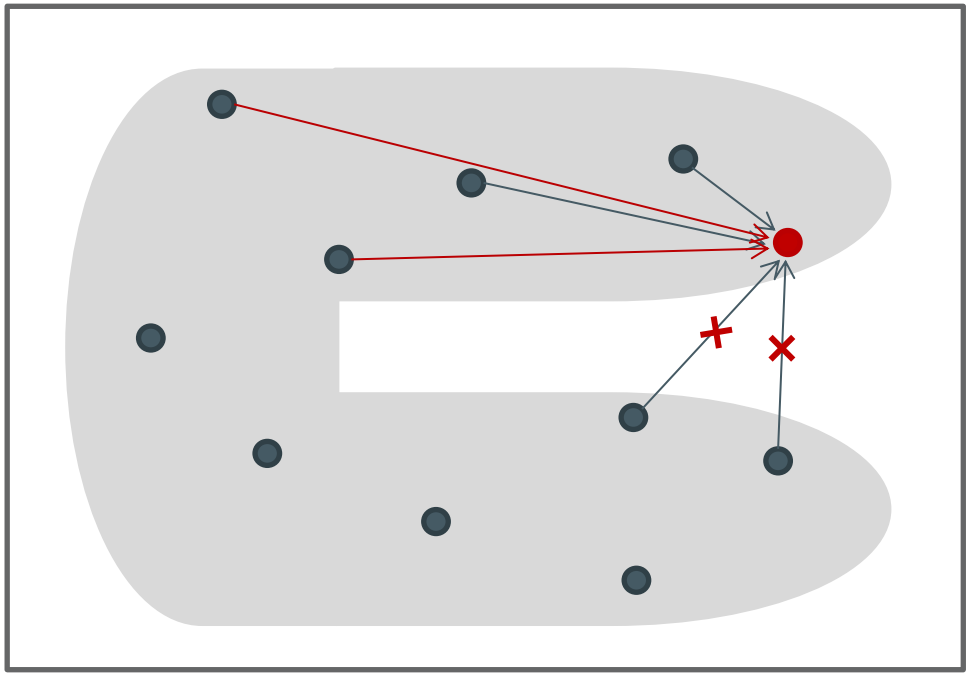


Duke

Our Proposal

- **B**arrier **O**verlap-**R**emoval **A**cyclic directed graph **G**aussian **P**rocess (BORA-GP*)

$m = 4$



*Credit: Michele Peruzzi

Duke

Our Proposal

1. Specify a multivariate normal distribution over a fixed finite set $R = \{r_1, \cdots, r_k\} \subset \mathcal{D}$

$$w_R = \big(w(r_1), \cdots, w(r_k)\big)^T \sim N\big(0, \tilde{C}_R\big) = \prod_{i=1}^{k} N(w(r_i); M_{r_i} w_{[r_i]}, V_{r_i})$$

- $M_{r_i} = C_{r_i,[r_i]} C_{[r_i]}^{-1}$, $V_{r_i} = C_{r_i} - C_{r_i,[r_i]} C_{[r_i]}^{-1} C_{[r_i],r_i}$
- C: base covariance function
- $[r_i]$: set of spatial locations in R whose straight line to r_i does not overlap barriers (physically sensible neighbors of r_i of size $\min(i-1, m)$)

Duke

Our Proposal

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

$$\forall u \in \mathcal{D} \setminus R,\ w(u)|w_R \sim N(M_u w_{[u]}, V_u)$$

- $[u] \subset R$ physically sensible neighbors of size m

➡ 1 & 2 define a valid scalable process

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

- $\tilde{C}$ is nonstationary: covariance between any two locations depends on R through their neighbor sets.

Duke

Neighbor Search Algorithm

Algorithm 1: Neighbor search for reference locations

Input: The number of neighbors m, the reference locations $\mathcal{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k\}$, and barriers $\mathcal{B}$

Output: Neighbor set $[\boldsymbol{r}_i] \subset \mathcal{R}$ of size $\min(m, i-1)$, $\forall \boldsymbol{r}_i \in \mathcal{R}$

Note: The reference locations are arranged in some ordering.

/* First-order neighbors */

for $i = 1, \ldots, m+1$ **do**
  $[\boldsymbol{r}_i] = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{i-1}\}^*$
for $i = m+2, \ldots, k$ **do**
  Sort† $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{i-1}\}$ by Euclidean distance to $\boldsymbol{r}_i$, enumerated by $\{\boldsymbol{r}_{\pi_1}, \ldots, \boldsymbol{r}_{\pi_{i-1}}\}$
  for $j = 1, \ldots, i-1$ **do**
    if $\overrightarrow{\boldsymbol{r}_{\pi_j}\boldsymbol{r}_i}$ does not intersect with $\mathcal{B}$ **then**
      $\boldsymbol{r}_{\pi_j} \in [\boldsymbol{r}_i]$
    if $\#([\boldsymbol{r}_i])^\ddagger = m$ **then**
      break

Duke

Neighbor Search Algorithm

```
   /* Second-order neighbors                                          */
10 for i such that 0 < #([r_i]) = m_1 < m  do
11     Sort [r_i]_2 := [[r_i]]^§ \[r_i] by sum of Euclidean distances¶
12     if #([r_i]_2) ≥ m − m_1 then
13         Include the first m − m_1 elements of [r_i]_2 in [r_i]
14     else
15         Include all elements of [r_i]_2 in [r_i] and repeat from step 10 as necessary.
```

Neighbor Search Algorithm

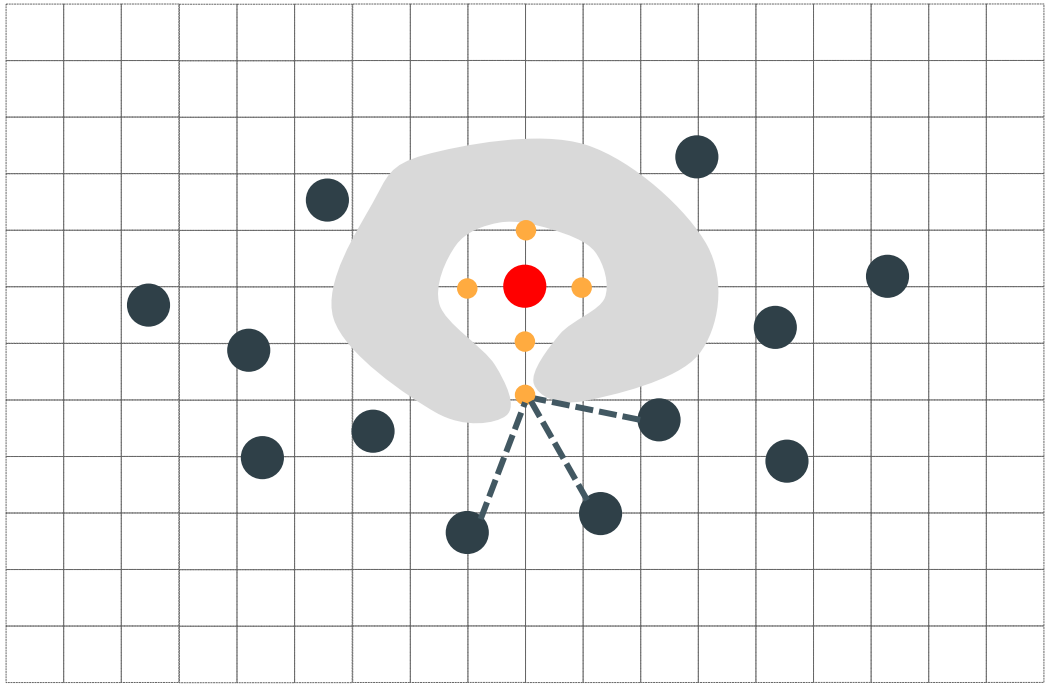
Duke

Neighbor Search Algorithm

1 **for** i such that $[\boldsymbol{r}_i] = \emptyset$ **do**
2 Sort m nearest grid points by Euclidean distance to $\boldsymbol{r}_i$, enumerated by $\{\boldsymbol{l}_{i1}, \ldots, \boldsymbol{l}_{im}\}$
3 **for** $t = 1, \ldots, m$ **do**
4 **if** $\overrightarrow{\boldsymbol{l}_{it}\boldsymbol{r}_i}$ does not intersect with $\mathcal{B}$ **then**
5 **for** $j = 1, \ldots, i-1$ **do**
6 **if** $\overrightarrow{\boldsymbol{r}_{\pi_j}\boldsymbol{l}_{it}}$ does not intersect with $\mathcal{B}$ **then**
7 $\boldsymbol{r}_{\pi_j} \in [\boldsymbol{r}_i]$
8 **if** $\#([\boldsymbol{r}_i]) > 0$ **then**
9 **break**

Duke

Neighbor Search Algorithm



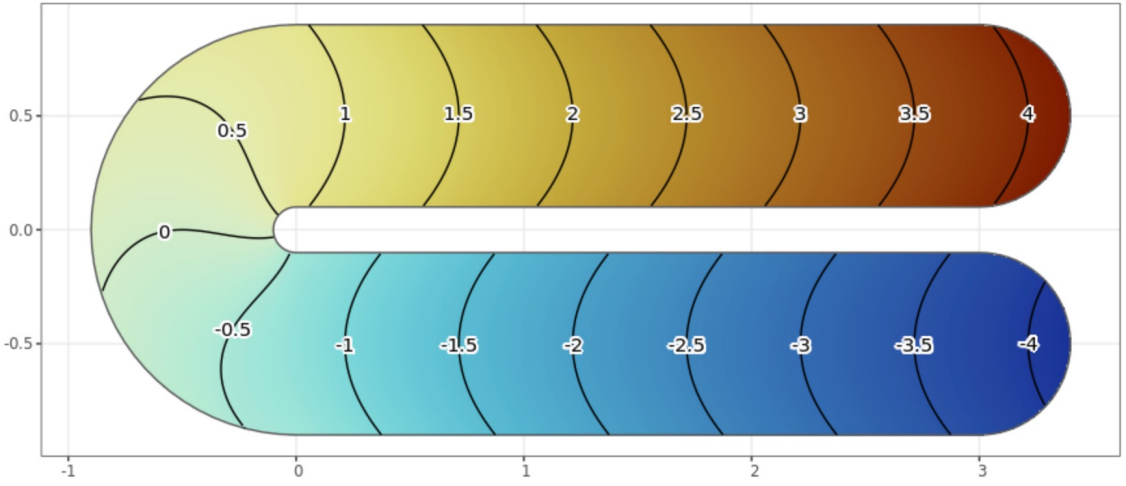
Duke

BORA-GP vs. NNGP

- BORA-GP proposes a new set of neighbors *conforming to barriers* in the NNGP framework.
- BORA-GP reduces to NNGP if no barriers.
- BORA-GP can enjoy nice properties of NNGP
  - ✓ Scalability
  - ✓ Nonstationarity
  - ✓ Standalone process
  - ✓ Versatility with any covariance functions
  - ✓ Straightforward extensions to multivariate, non-Gaussian data, etc.
- End-goal for BORA-GP vs. side product for NNGP

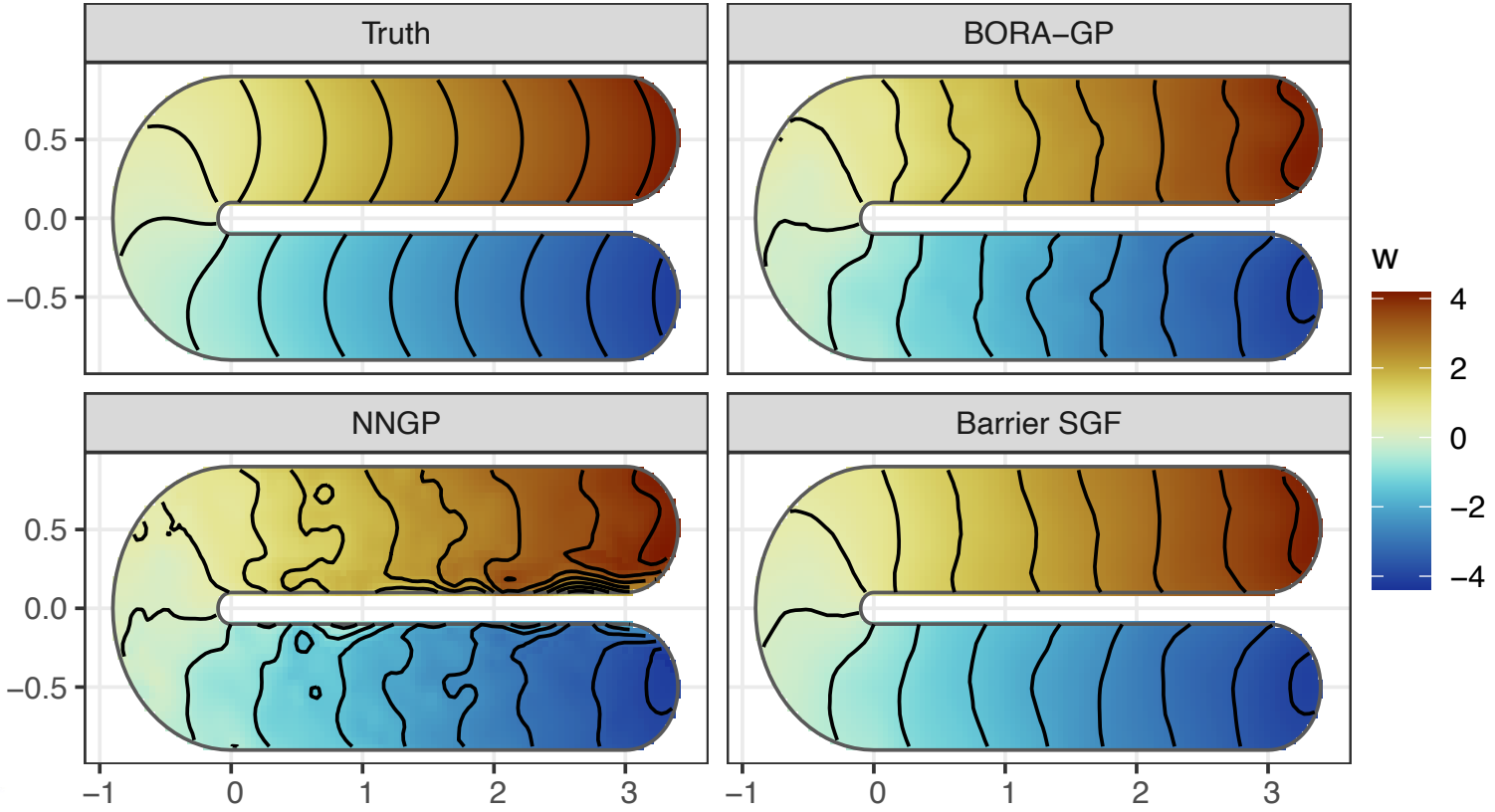
Modified Horseshoe

- $y(s) = w(s) + \epsilon(s),\ \epsilon(s) \sim N(0, \tau^2),\ \tau = 0.1$
- $w(s)$ generated as suggested in Wood et al. (2008)
- Compare BORA-GP with NNGP and Barrier SGF (Bakka et al. 2019)
- 30 replicates for each combination of n = 300, 600, 1000 and m = 10, 15, 20



Duke

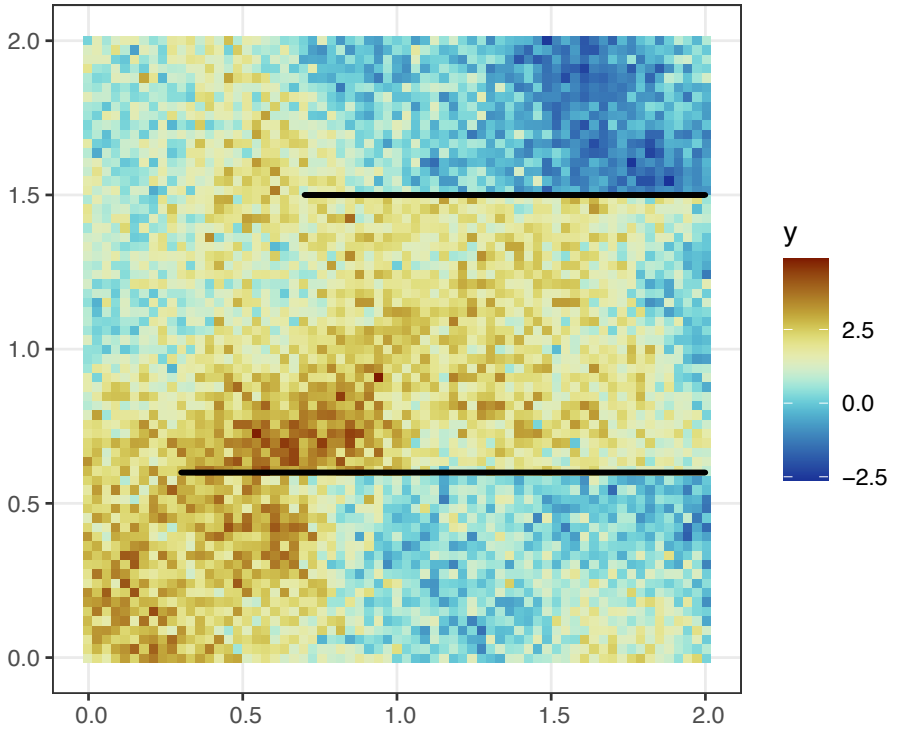
Modified Horseshoe



Duke

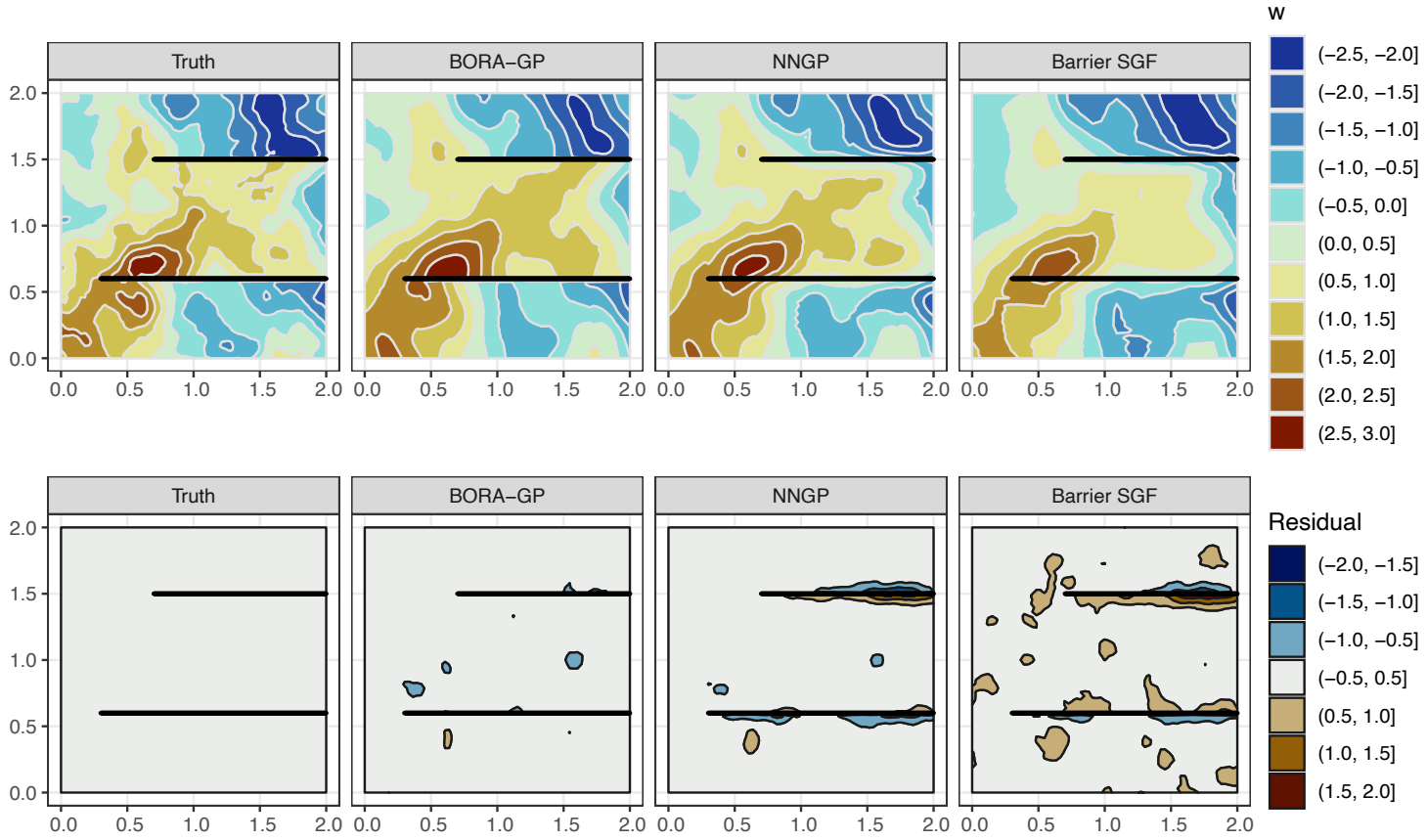
Domain with Faults

- $y(s) = \beta_0 + x(s)\beta_1 + w(s) + \epsilon(s)$, $\epsilon(s) \sim N(0, 0.1^2)$
- $\beta_0 = 1,\ \beta_1 = 0.5,\ x(s) \sim N(0,1)$
- $w(s)$ generated by BORA-GP
  - $m = 15$
  - Matérn base covariance function with $\sigma^2 = 1,\ \nu = 1.5,\ \phi = 4$
- 4,489 locations in a domain with two faults
- One example out of 30 replicates



Duke

Domain with Faults



Duke

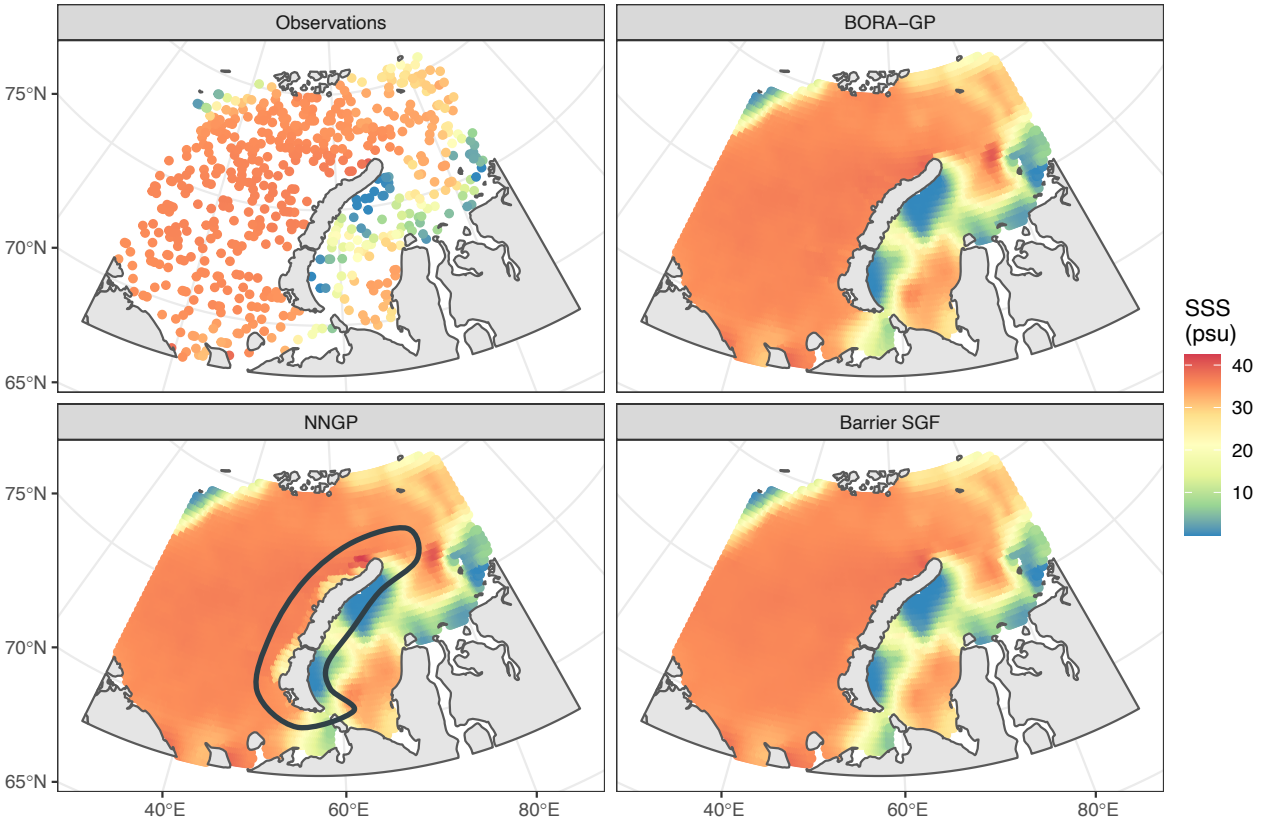
Analysis

- Bayesian spatial regression

$$y(s) = \beta_0 + w(s) + \epsilon(s),\ \ \epsilon(s) \sim N(0, \tau^2),$$
$$w(s) \sim \text{BORA-GP}(0, \tilde{C}(\cdot,\cdot\,|\theta))$$

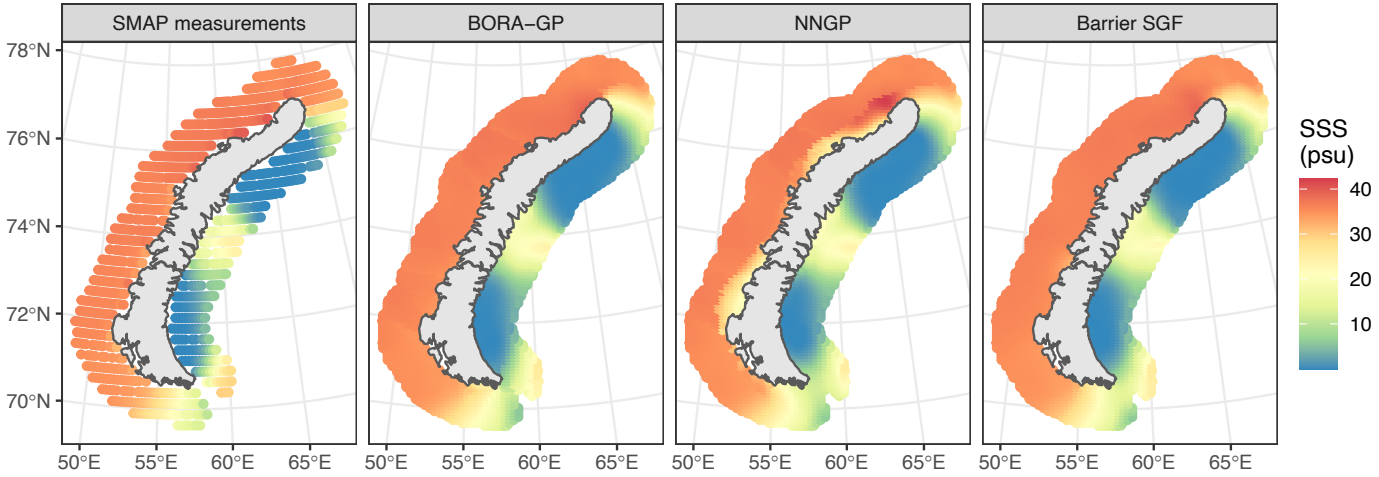
- $y(s)$: sqrt of SSS in late July of 2017
- Data from Soil Moisture Active Passive (SMAP) V5.0 at a 60km spatial resolution
- Number of obs n = 589 out of 8,418 measurements
- 10,315 new spatial locations to predict (7,829 labeled + 2,486 unlabeled)
- Matérn covariance function with $\nu = 1$ as a base C
- For BORA-GP & NNGP, m = 15

Analysis



Duke

Analysis

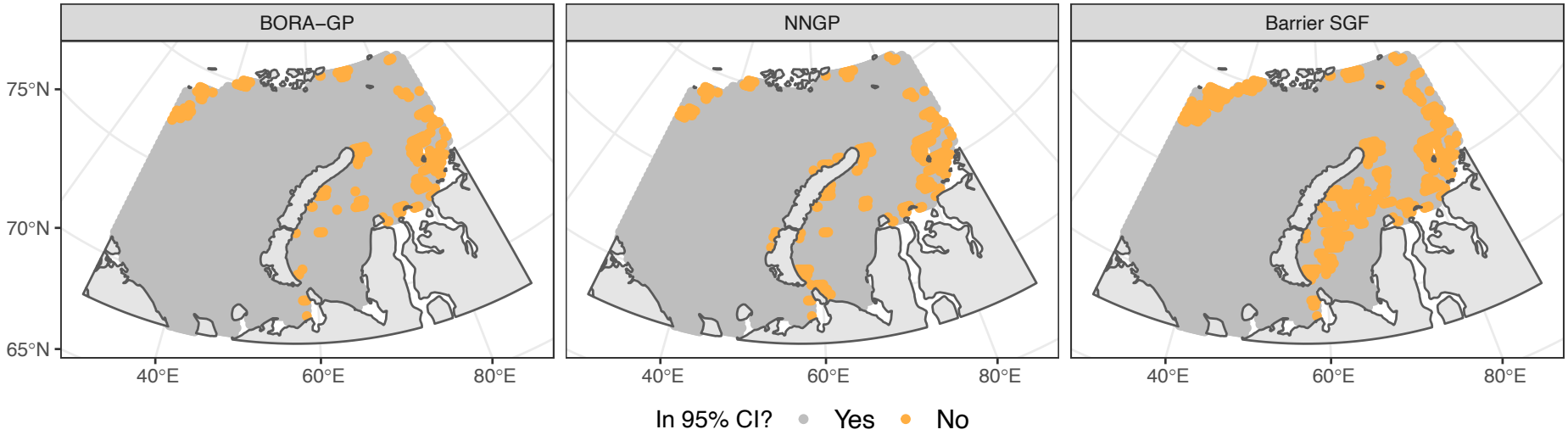


	BORA-GP	NNGP	Barrier SGF
RMSPE	**0.325**	0.412	0.332
MAPE	**0.147**	0.218	0.159
95% CI coverage	**0.950**	0.911	0.875
Mean 95% CI width	1.140	1.196	**0.876**

>48%

Duke

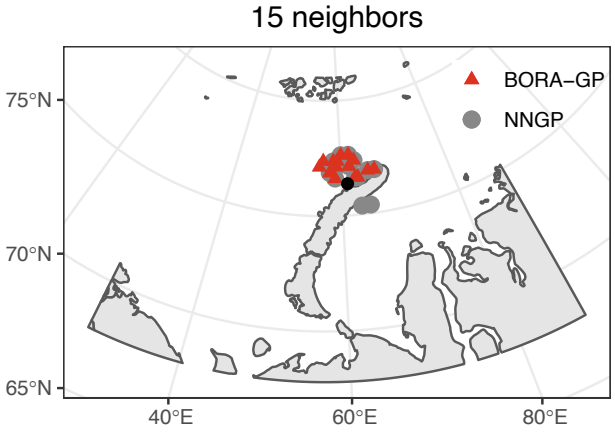
Analysis

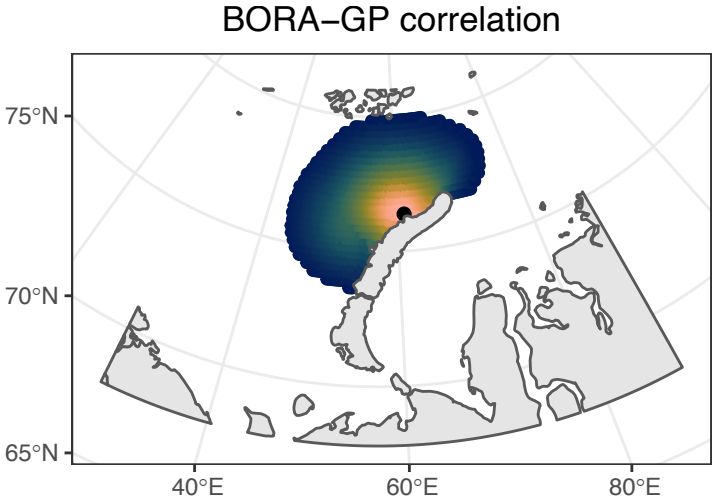


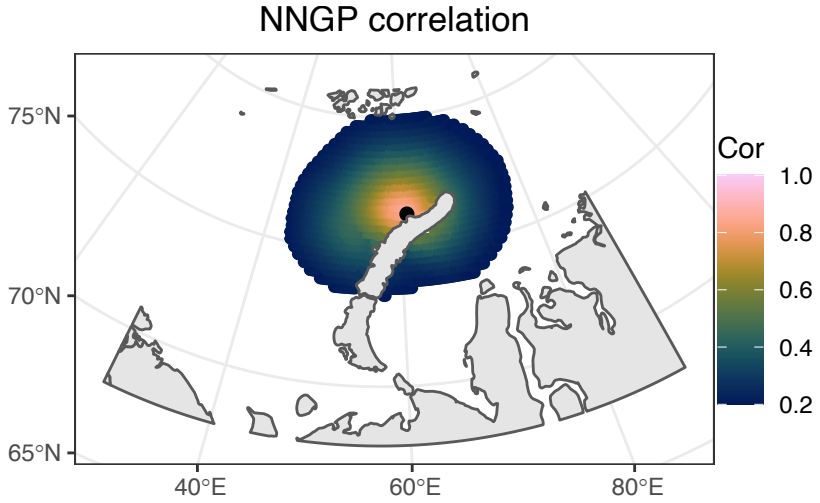
- Low salinity is a key to understand various ocean processes!
  - Strong stratification: less circulation of nutrients → limited productivity of phytoplankton
  - Acidification: negative impacts on marine species, marine food-webs, reefs in coastal protection against storms

Duke

Analysis

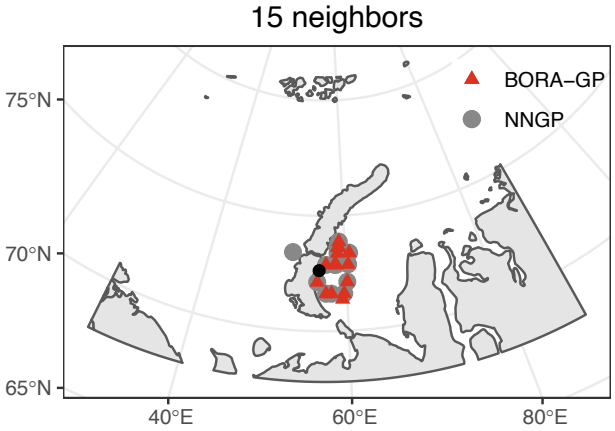


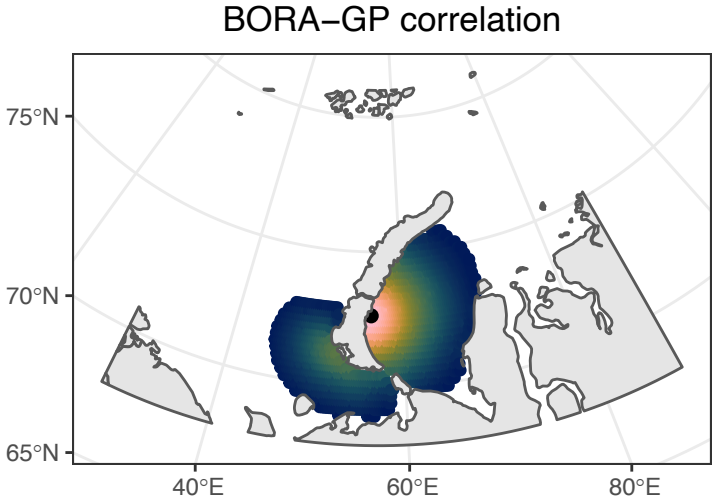


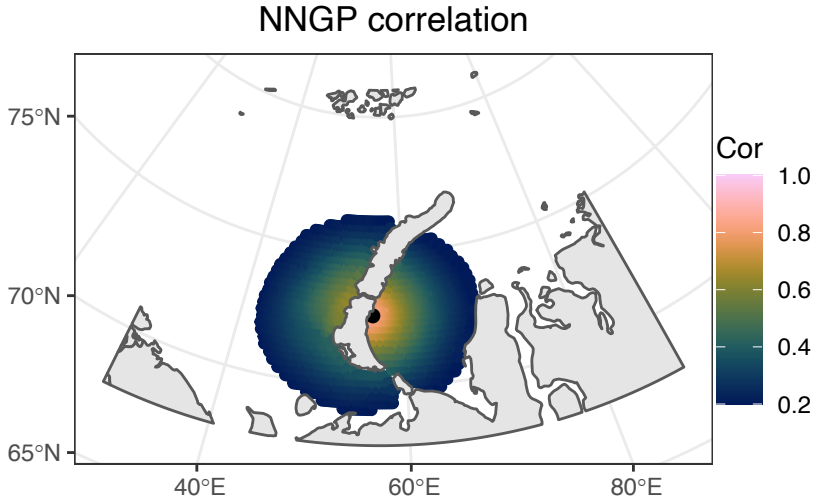


Duke

Analysis







Duke

Conclusions & Discussions

- BORA-GP considers geometric features of domains and produces
  - physically sensible neighbors,
  - plausible nonstationary covariance behaviors,
  - physically sensible prediction.
- Users familiar with NNGP can use BORA-GP seamlessly.
- Assume hard barriers (Horseshoe, sliding doors, line barriers, complex coastlines and islands, etc.)
- Extend it to a soft barrier
  - For any location s, each reference location r receives a weight that decreases as either Euclidean distance between s and r increases or a directed edge from r to s overlaps barriers.
  - Locations crossing barriers can still receive a small but positive probability.

Duke

Thank you!

Paper and an R package coming soon

bora.jin@duke.edu

We are grateful for the financial support from the National Institute of Environmental Health Sciences through grants R01ES027498 and R01ES028804 and from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506)

Duke