Deep Spatial Learning for Forensic Geolocation with Microbiome Data

Neal S. Grantham
North Carolina State University

JSM Chicago
August 3rd, 2016

Joint work with

- ☞ Brian Reich (NCSU Statistics)
- ☞ Eric Laber (NCSU Statistics)

In collaboration with

- ☞ Rob Dunn (NCSU Biology)

What is a microbiome?

Community of microbial organisms occupying an ecological niche.

Next-generation sequencing technologies make possible efficient identification of these microbes at affordable cost[1].

Huge interest in understanding microbiomes as they relate to human health, diet, agriculture, environment, forensics, etc.

[1] Metzker. (2010) Sequencing technologies—the next generation. Nature Reviews Genetics.

The White House Launches the National Microbiome Initiative[2]

Half a billion dollars pledged, with three major goals:

- ☞ collaboration,
- ☞ developing better tools for studying microbiomes, and
- ☞ recruitment.

[2] The Atlantic article by Ed Yong, photo by Jim Young / Reuters

Statistical challenges

Microbiome data are...

☞ high-dimensional,

☞ sparse,

☞ over-dispersed, and

☞ possess complex dependence structure.

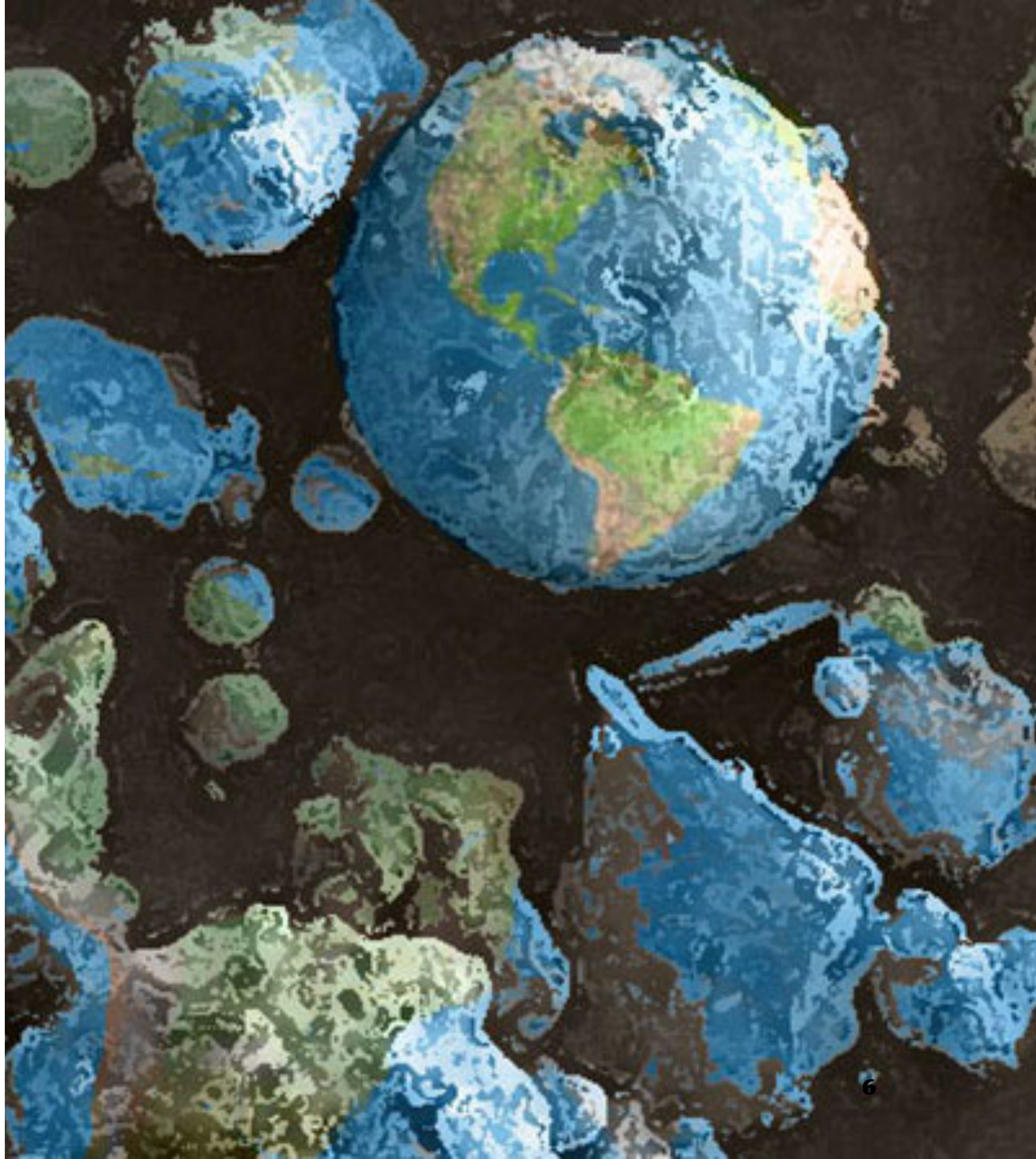
Many exciting opportunities for research with microbiologists.

Motivating dataset

Wild Life of Our Homes, a public science project by Rob Dunn Lab.

- ☞ Dust samples collected from outer door frames of $n \approx 1,000$ homes across the continental U.S.
- ☞ DNA sequencing revealed $p \approx 50,000$ fungi species[3].

[3] Well, operational taxonomic units (OTUs) to be precise.



Research question:

Does the microbiome composition of an ambient dust sample inform its geographic origin?

Our approach:

Build a model to estimate the unknown origin $\boldsymbol{s}$ of a dust sample conditional on its known microbiome composition $\boldsymbol{x}$.

Initial model and direction

Naive Bayes discriminant analysis estimates origin of dust samples with a median prediction error of 230 kilometers[4].

Here, we develop a new model based on

① spatial point pattern theory, and

② deep learning.

[4] Grantham et al. (2015) Fungi identify the geographic origin of dust samples. PLOS One.

1. Spatial point pattern theory

Assume a spatial point pattern[5] intensity surface $\boldsymbol{\lambda}(\boldsymbol{s} \mid \boldsymbol{x})$ has non-homogeneous Poisson process likelihood

$$\mathcal{L}\left[\lambda(\boldsymbol{s} \mid \boldsymbol{x}); \{(\boldsymbol{s}_i, \boldsymbol{x}_i)\}_{i=1}^n\right] = \exp\left[-\int_{\mathcal{D}} \lambda(\boldsymbol{s} \mid \boldsymbol{x}) d\boldsymbol{s}\right] \prod_{i=1}^n \lambda(\boldsymbol{s}_i \mid \boldsymbol{x}_i).$$

May select a parametric model for $\boldsymbol{\lambda}(\boldsymbol{s} \mid \boldsymbol{x})$.

[5] Gelfand et al. (2010) Handbook of Spatial Statistics. CRC Press.

1. Spatial point pattern theory

Liang et al.[6] propose a log Gaussian process (GP),

$$\boldsymbol{\lambda(s \mid x) = \pi(s) \exp[\beta' x + w(s)]}$$

with $\boldsymbol{w(s) \sim GP}$, $\boldsymbol{\pi(s)}$ a population offset, and $\boldsymbol{\beta}$ unknown.

However, no closed-form solution to integral in $\boldsymbol{\mathcal{L}}$.

[6] Liang et al. (2008) Analysis of Minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. The Annals of Applied Statistics.

1. Spatial point pattern theory

Numerical approximation possible with a Monte Carlo algorithm using a knot-based predictive process[6].

Effective, but difficult to implement in practice:

☞ Requires careful knot construction.

☞ Performance suffers for large $\boldsymbol{n}$, larger $\boldsymbol{p}$.

[6] Liang et al. (2008) Analysis of Minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. The Annals of Applied Statistics.

2. Deep learning

Well-suited for high-dimensional data with complex structure[7].

Rich, GPU-enabled software libraries available:

☞ Theano (Python)

☞ Torch7 (Lua)

☞ Tensorflow (C++)

[7] LeCun et al. (2015) Deep learning. Nature.

2. Deep learning

Let $\mathcal{P} = \{P_k\}_{k=1}^{K}$ denote a partition of spatial domain $\mathcal{D}$.

Represent every $\boldsymbol{s}$ by the region to which it belongs.

Two major benefits:

☞ Avoids costly approximation of integral in $\mathcal{L}$.

☞ Reframes estimation as a supervised classification problem.

2. Deep learning

For regions $k = 1, \ldots, K$,

$$Pr(s \in P_k \mid x) = \frac{\exp[f_k(x)]}{\sum_{l=1}^{K} \exp[f_l(x)]} \text{ where } f_k : \mathcal{X} \to \mathbb{R}.$$

Estimate $f_1(\cdot), \ldots, f_K(\cdot)$ by training a deep neural network (DNN) on $\{(s_i, x_i)\}_{i=1}^{n}$ with categorical cross-entropy cost function

$$C = -\sum_{i=1}^{n} \sum_{k=1}^{K} \log[Pr(s_i \in P_k \mid x_i)] I(s_i \in P_k).$$

Our "Deep Space" algorithm

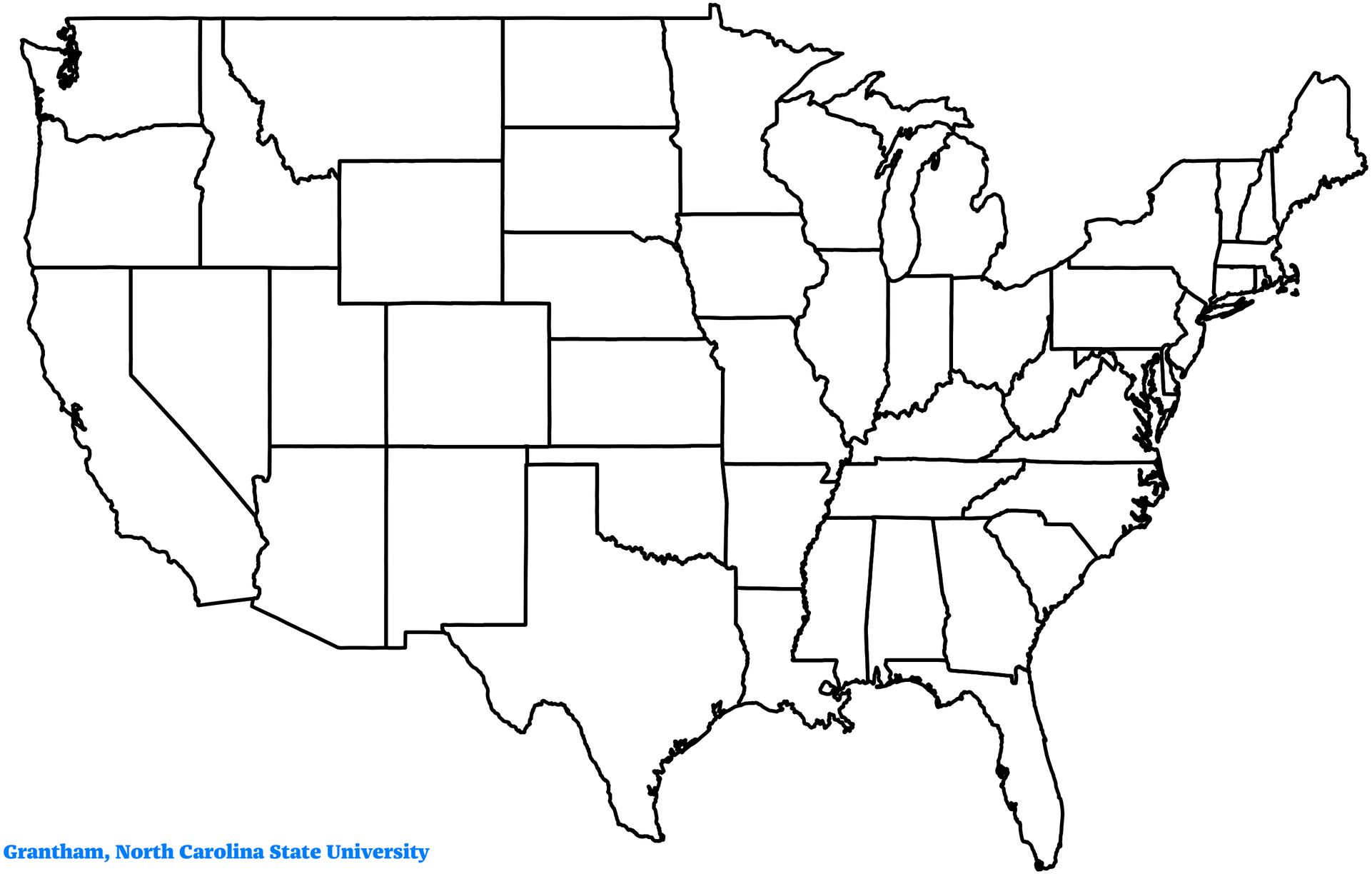
① Generate random Voronoi partition $\mathcal{P}$ over $\mathcal{D}$.

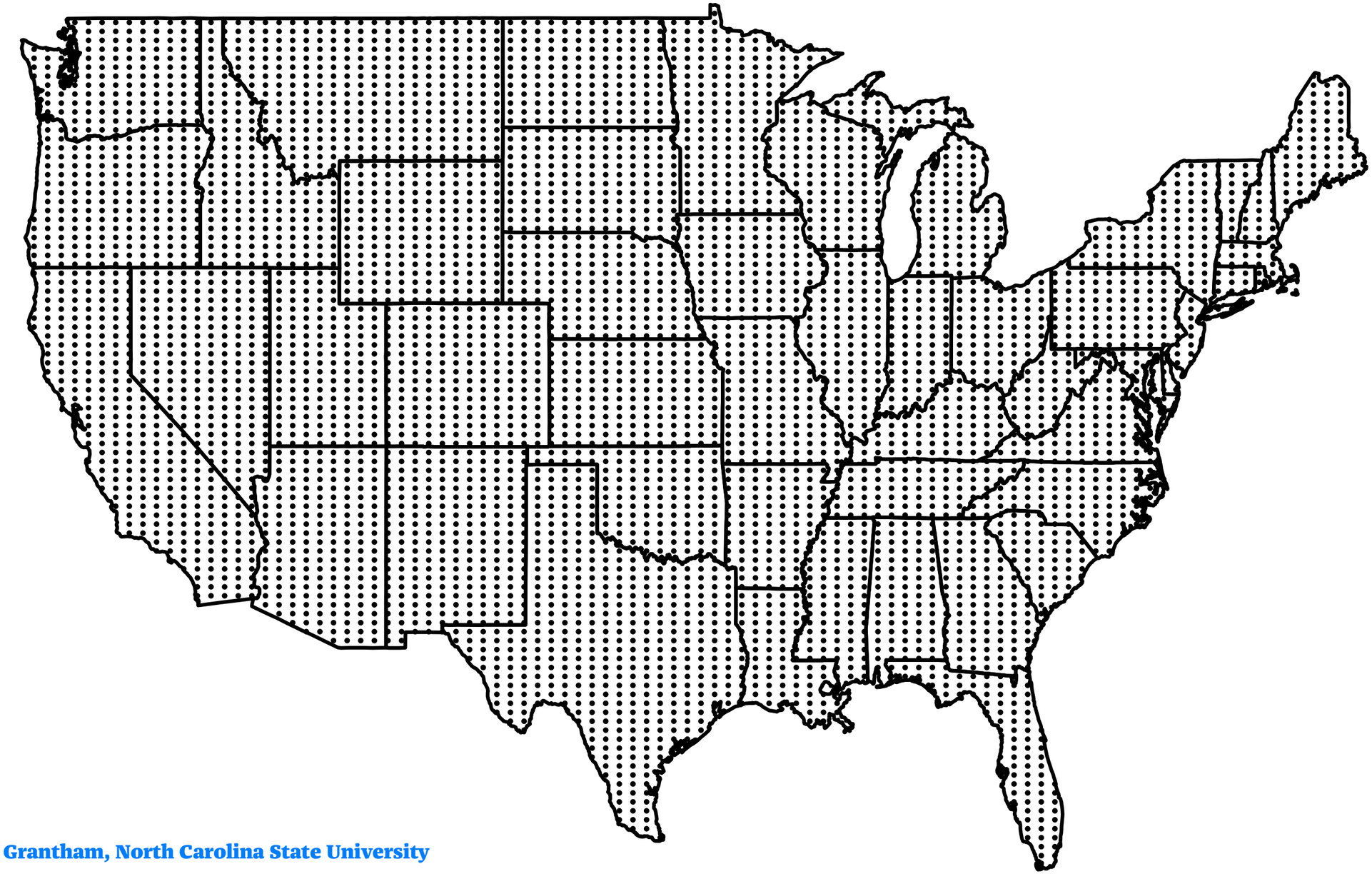
② Train DNN on available data from these regions.

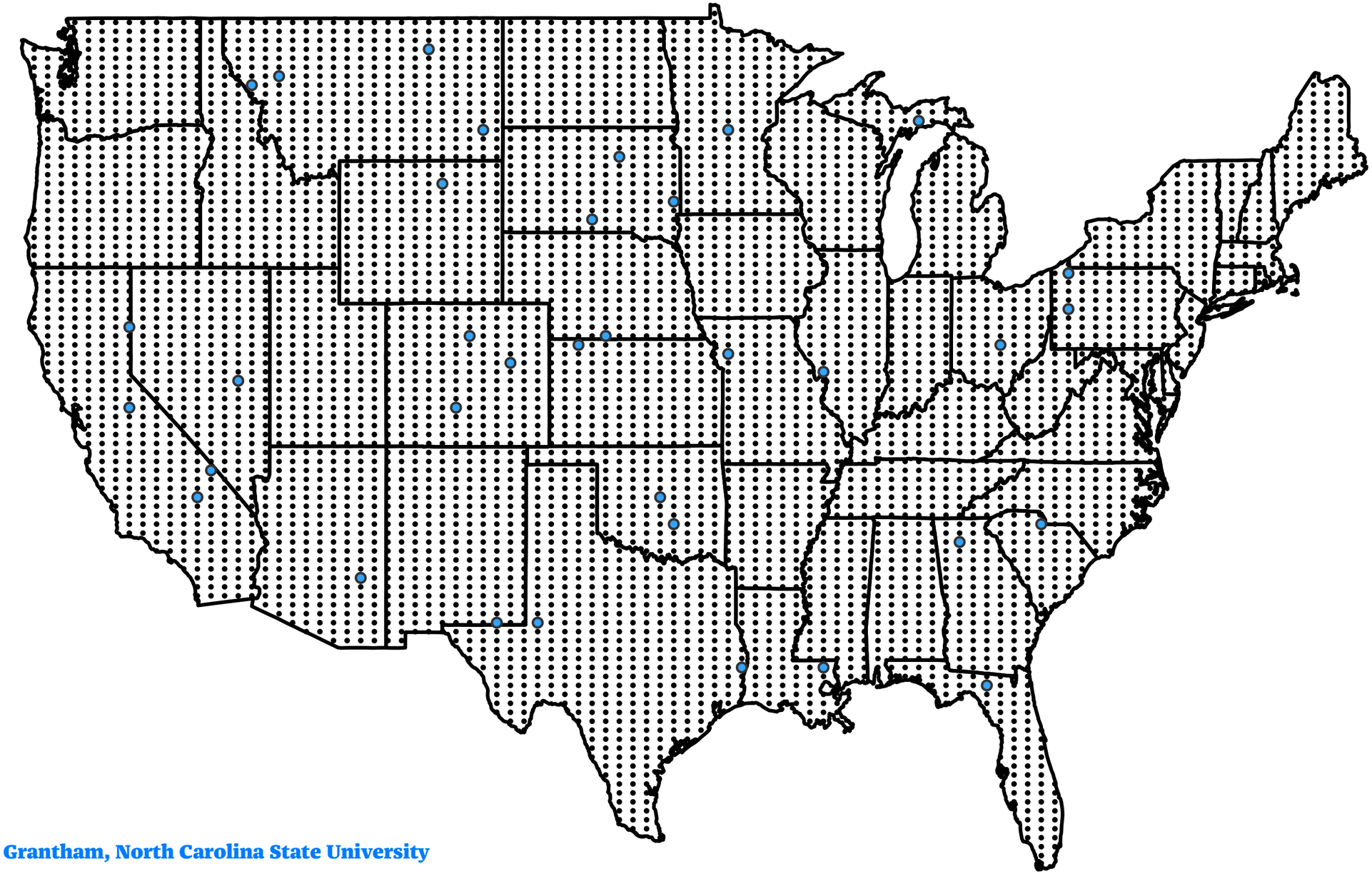
③ Repeat steps 1 & 2 N times to develop a diverse collection, $\mathcal{M}$, of trained DNNs.

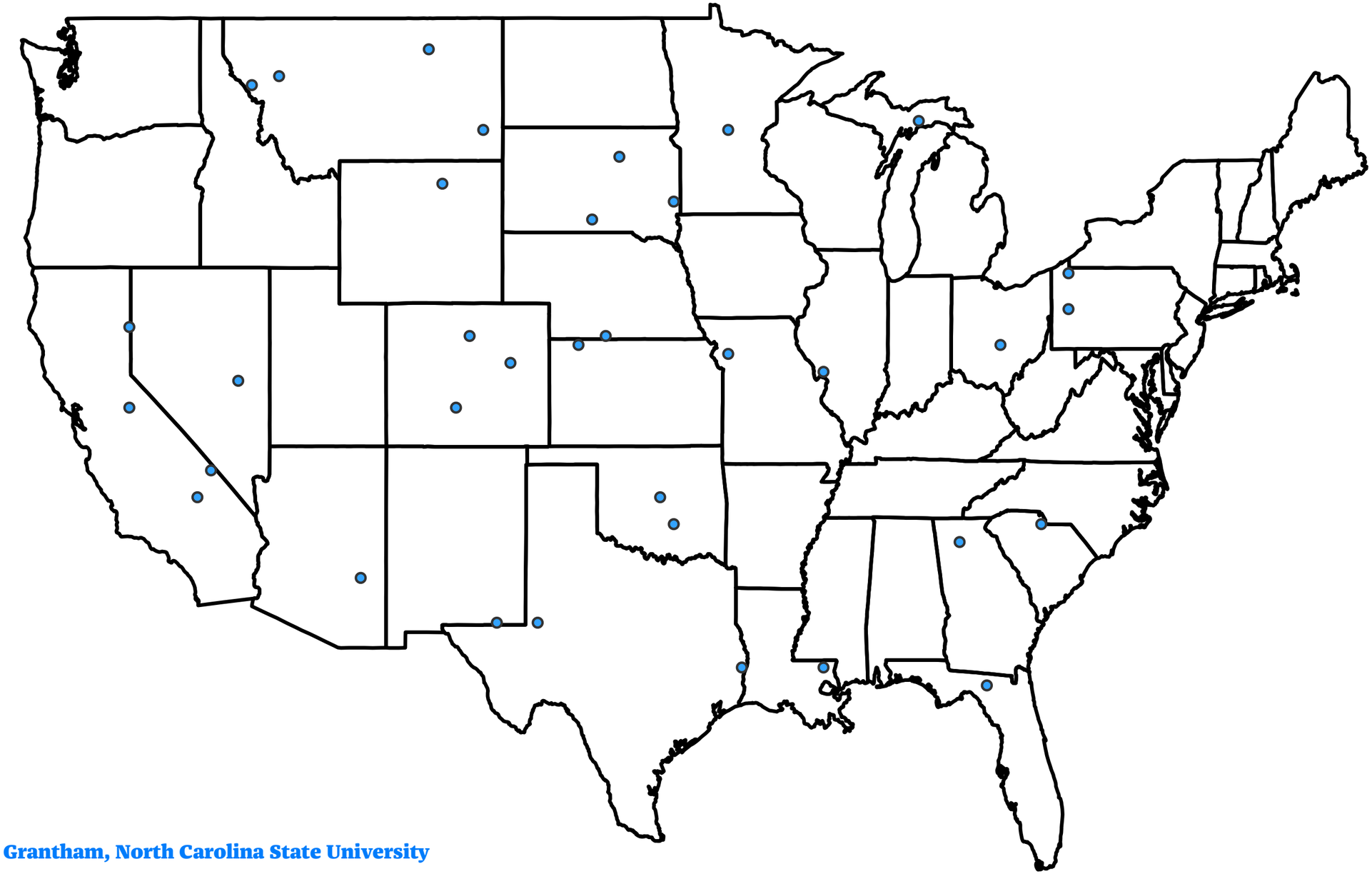
④ Predict most likely origin $\hat{s}$ by averaging over DNNs in $\mathcal{M}$.

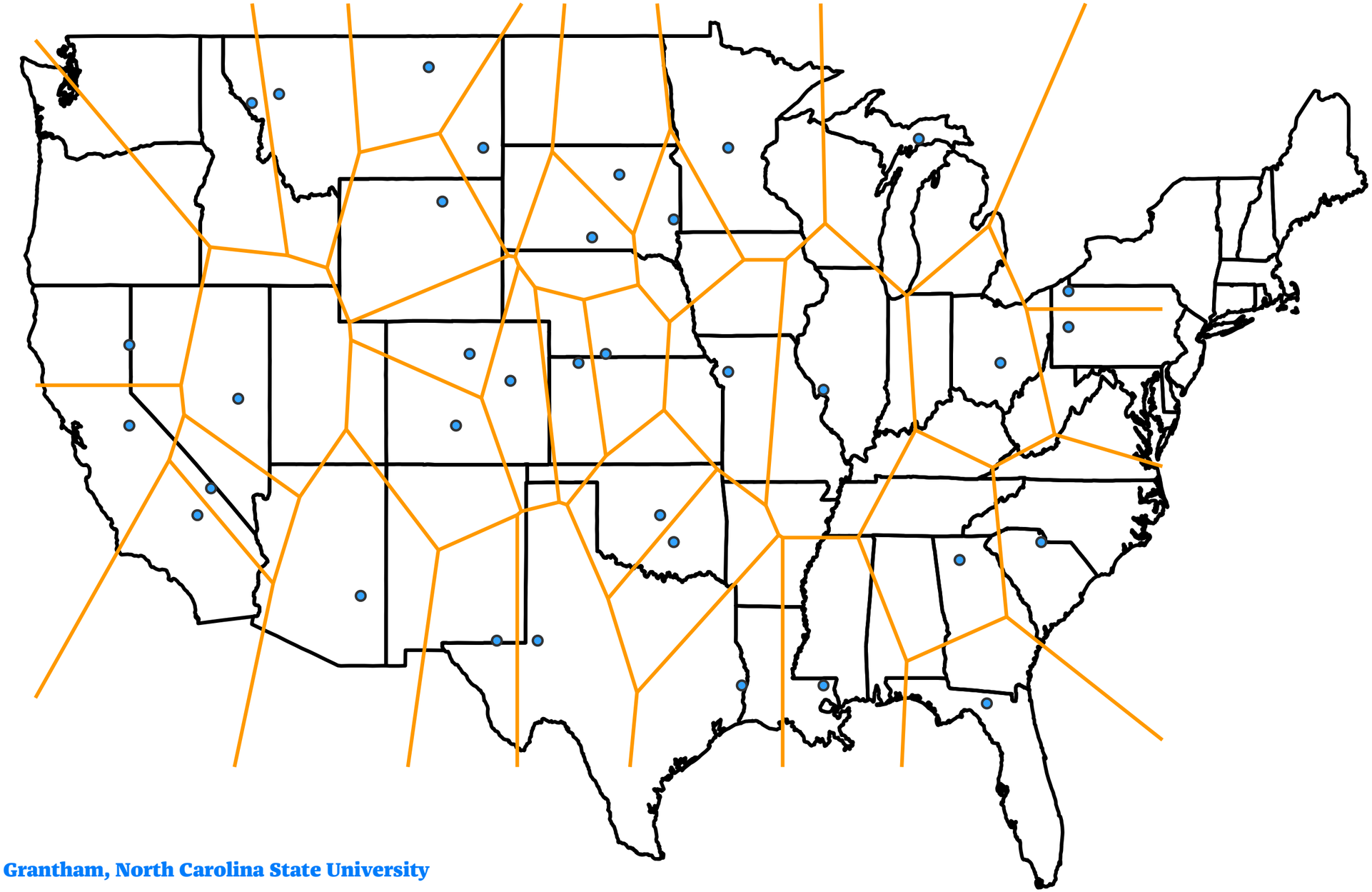
1. Generate random Voronoi partition $\mathcal{P}$ over $\mathcal{D}$.

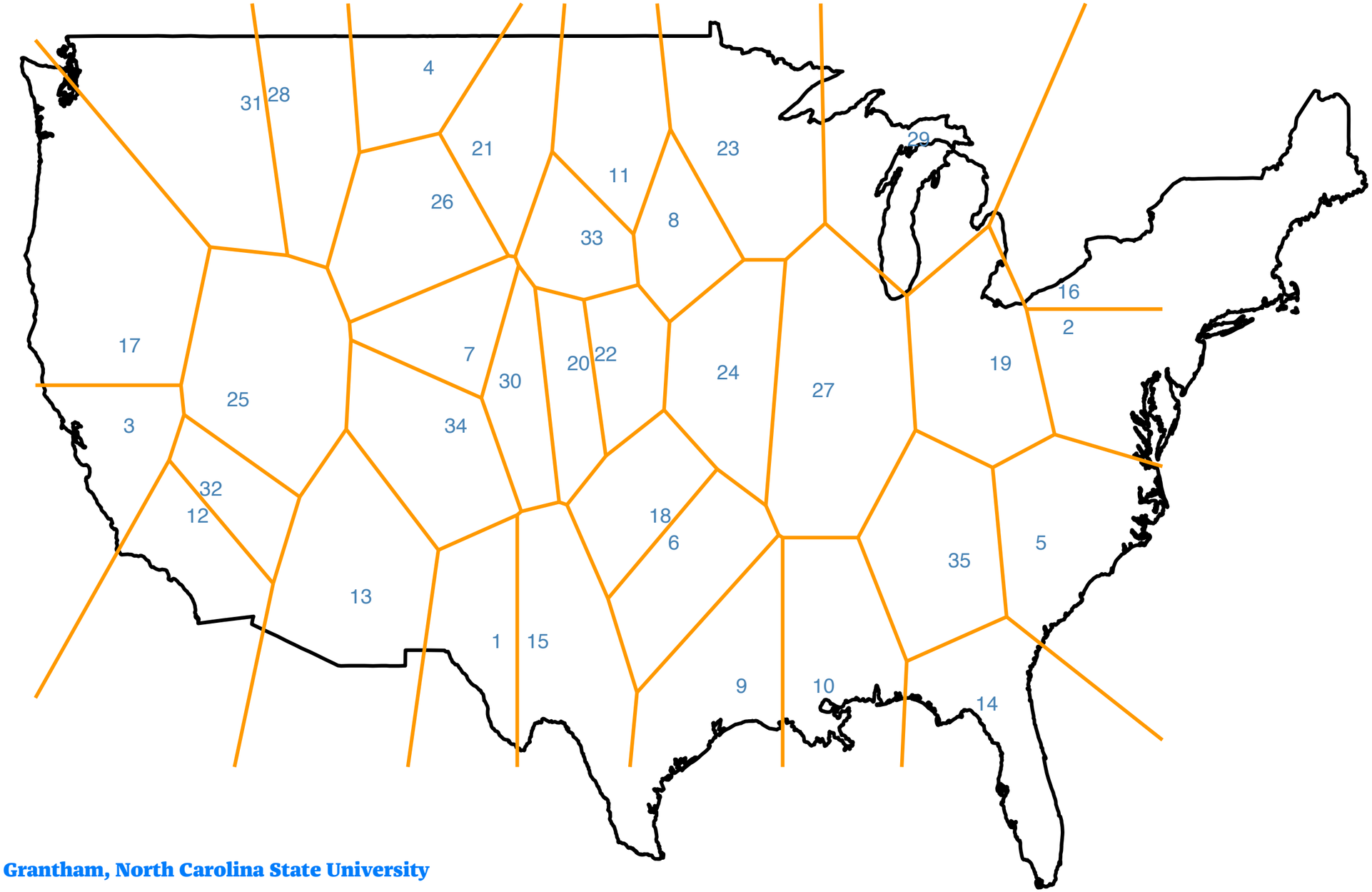




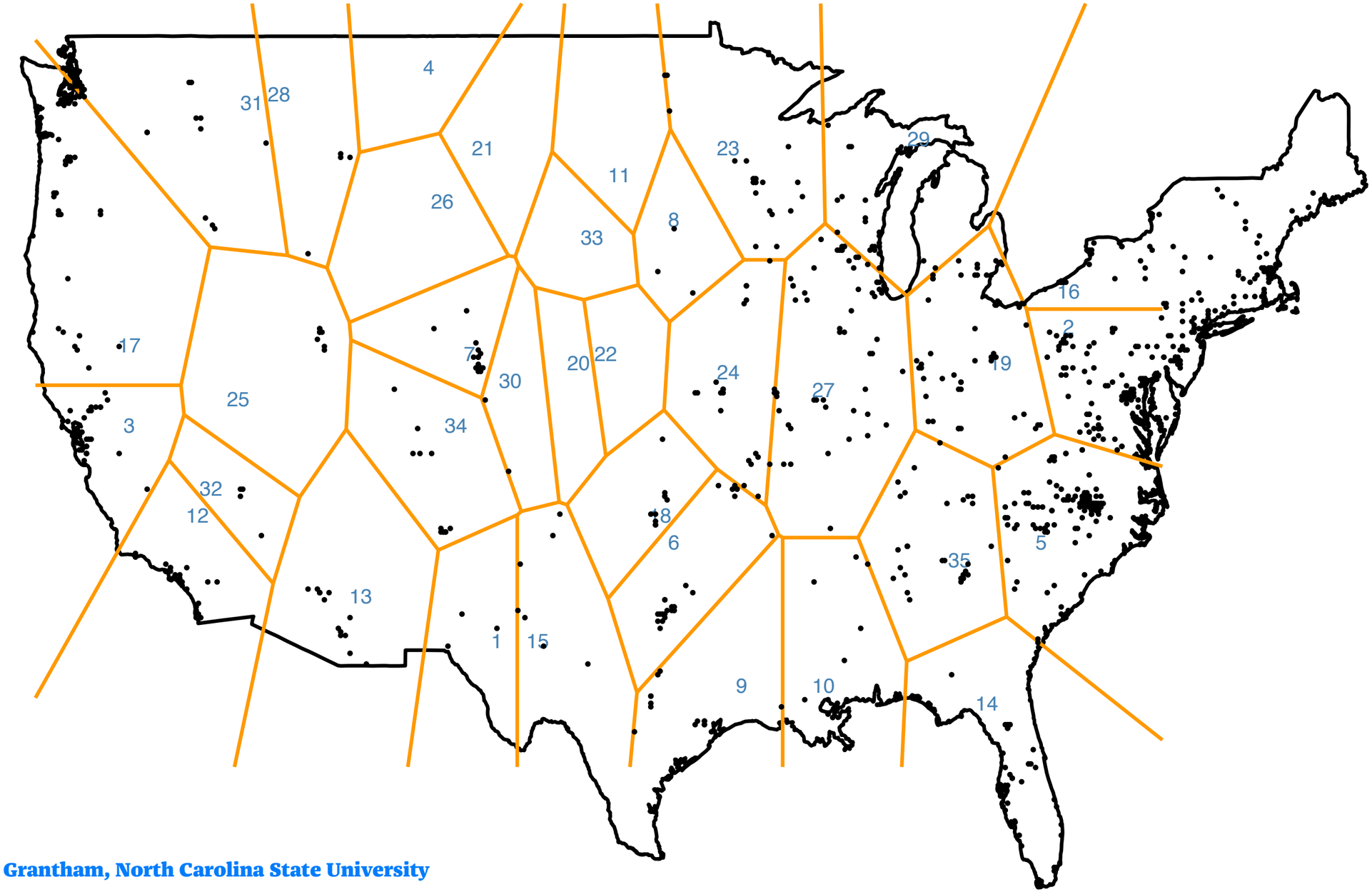


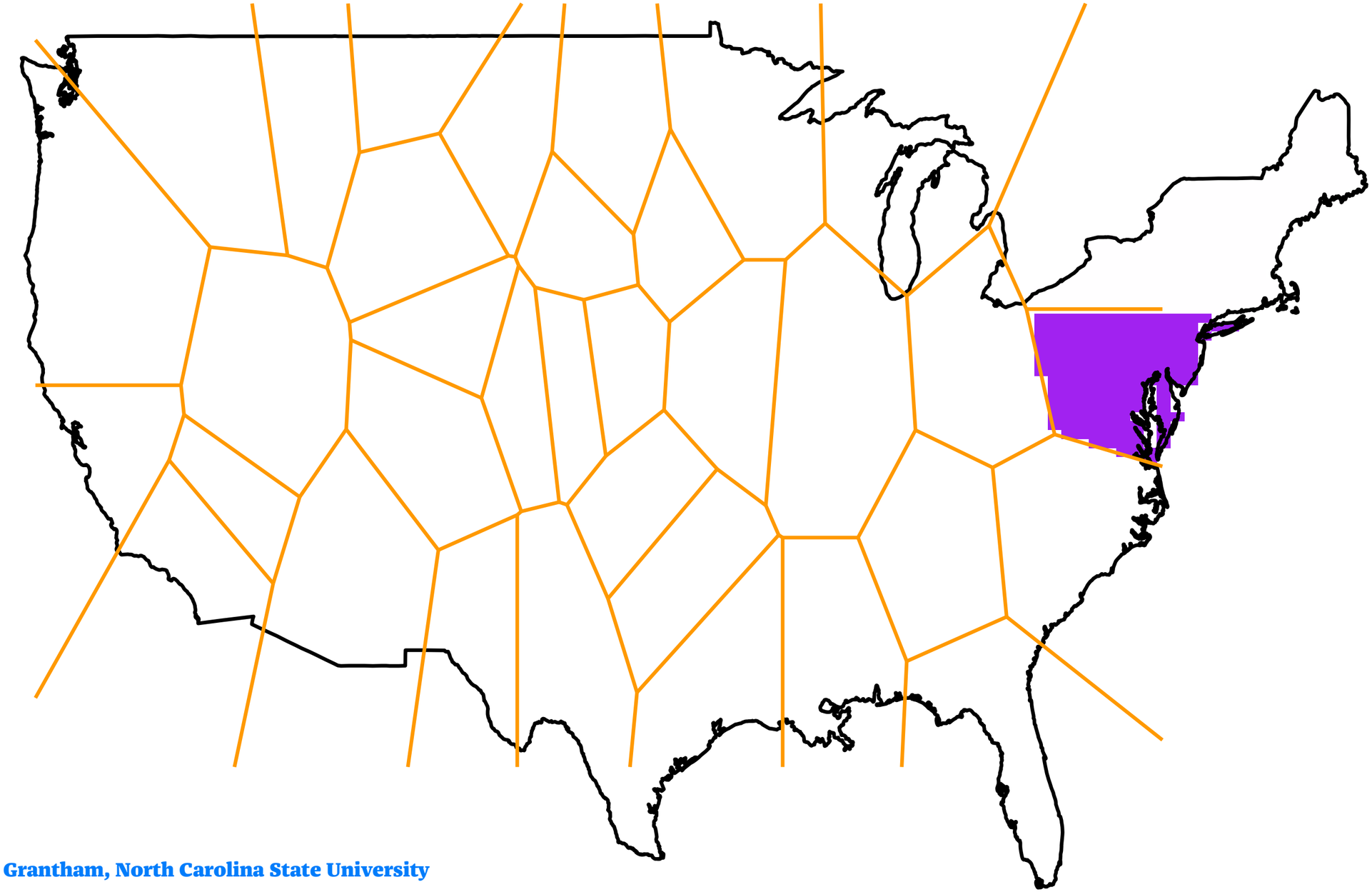




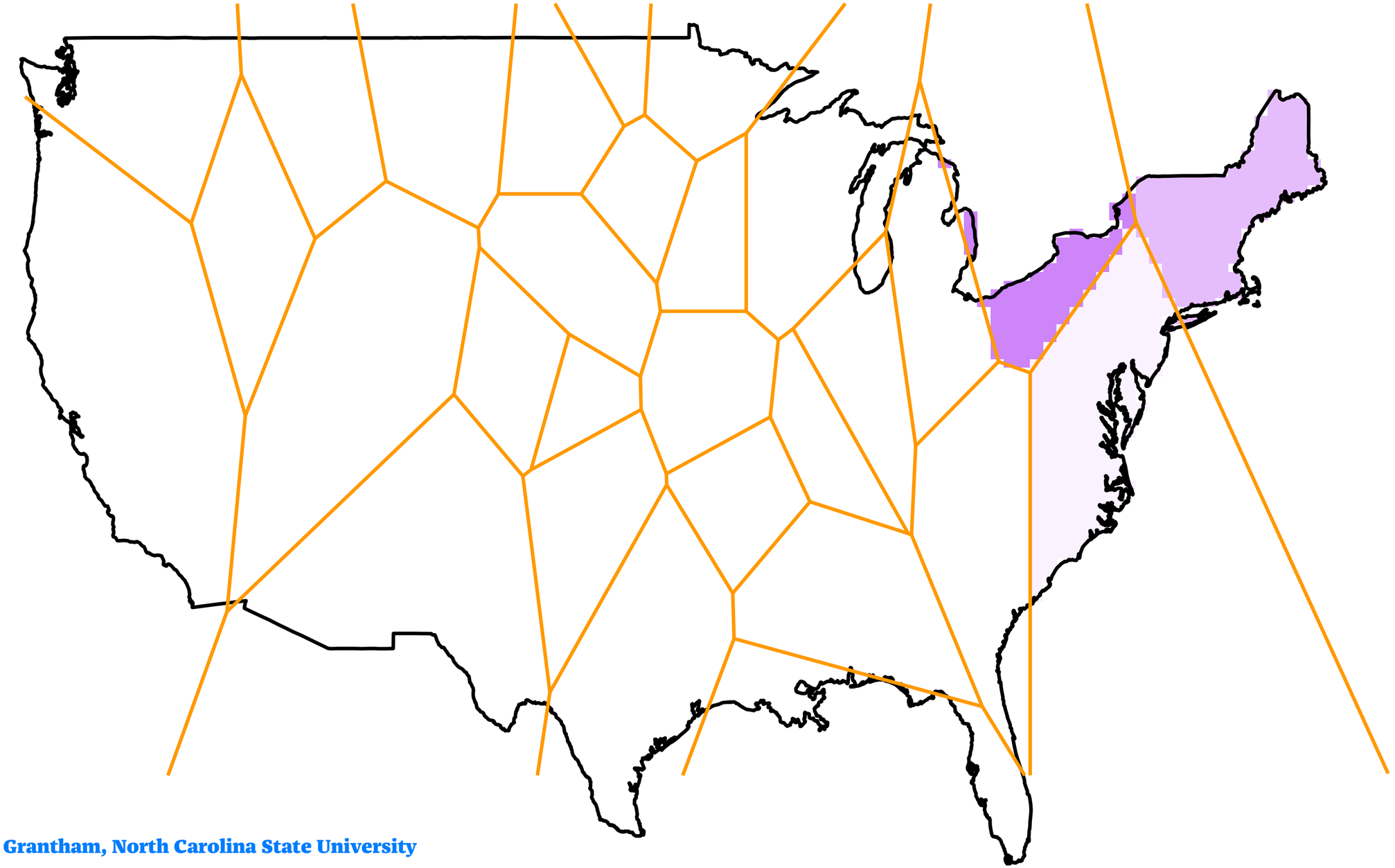

4
31
28
21
29
23
11
26
8
33
16
2
17
7
20
22
19
30
24
27
25
3
34
32
12
18
6
5
35
13
1
15
9
10
14

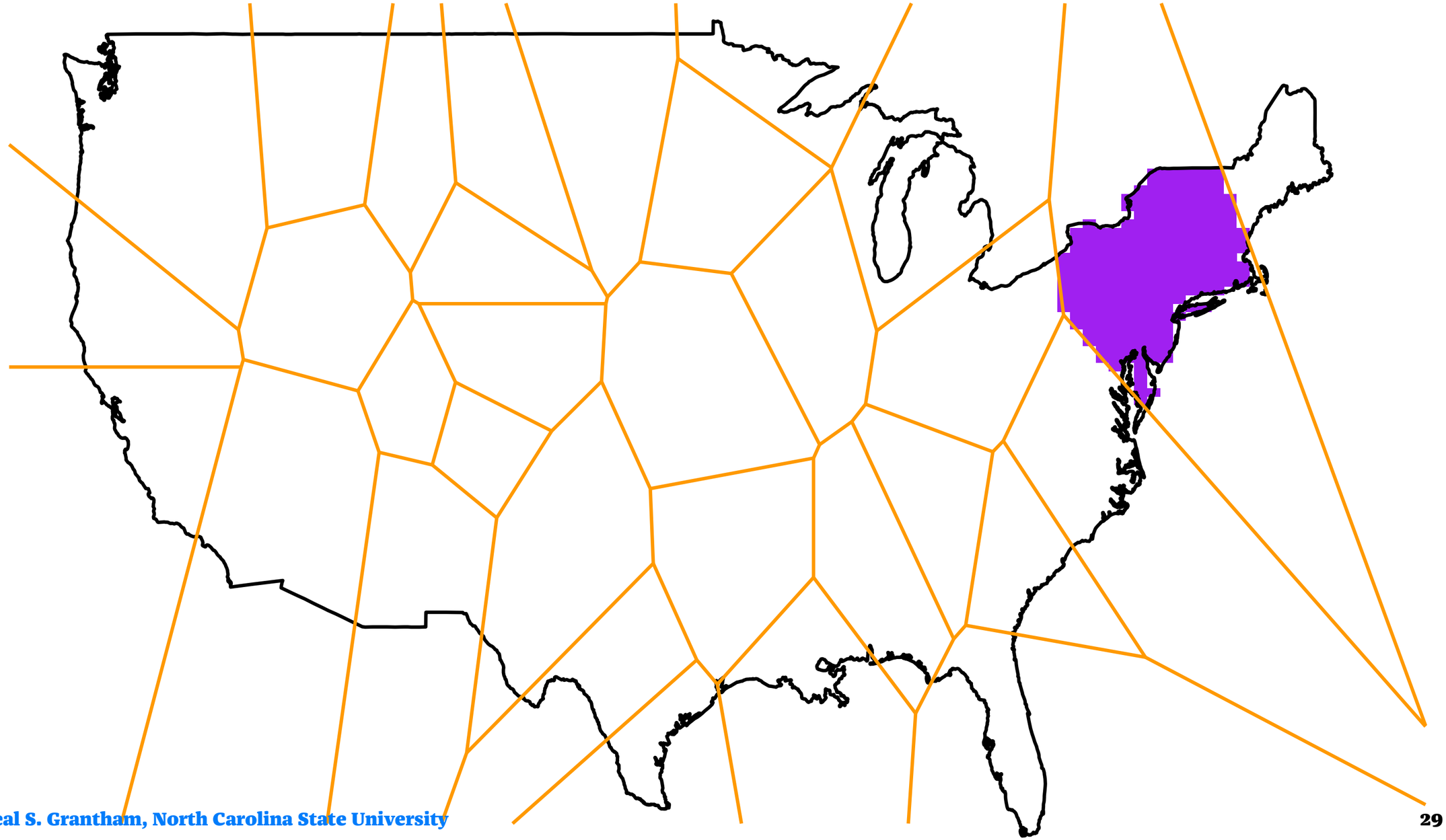
2. Train DNN on available data from these regions.

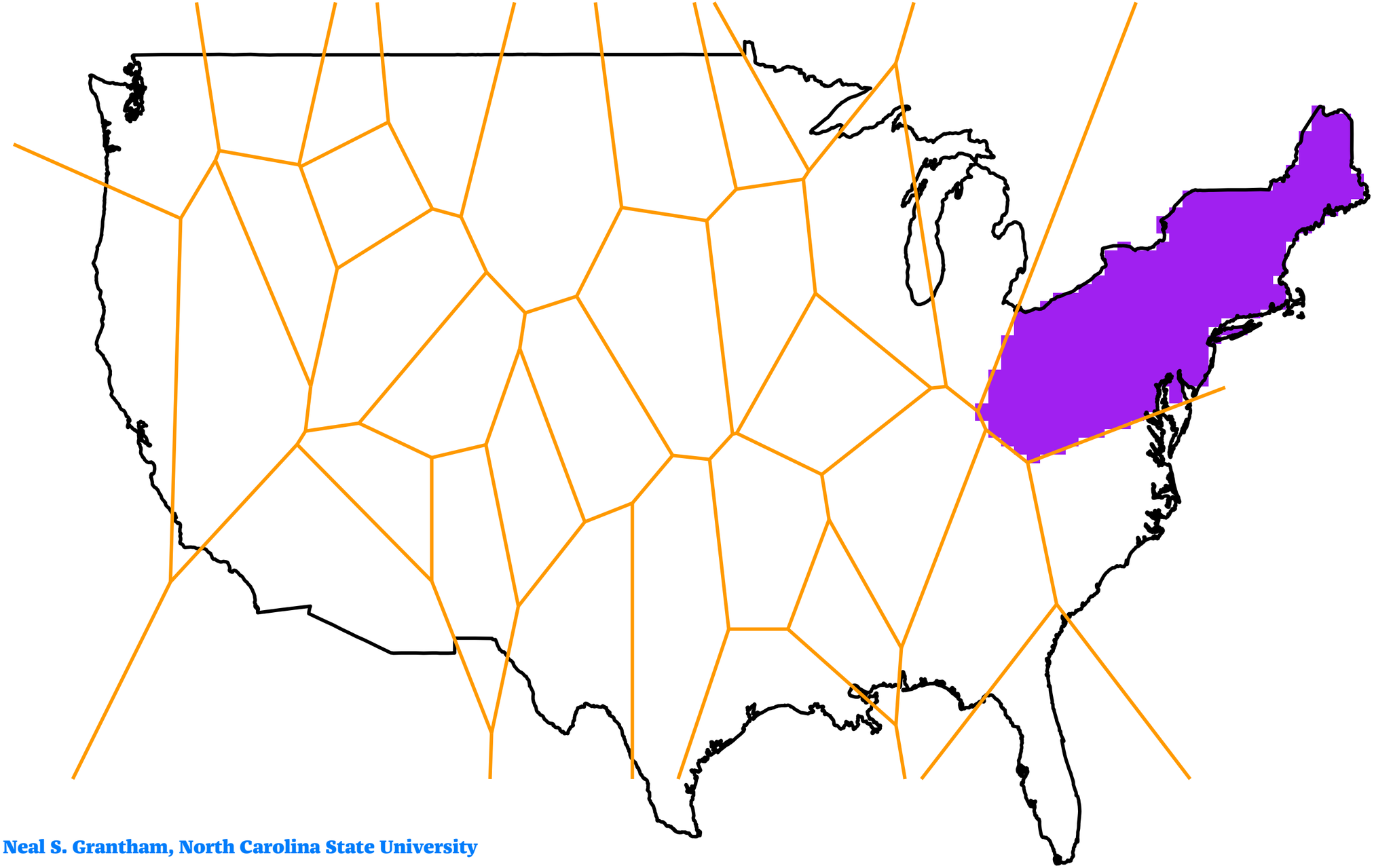

4
31
28
21
26
11
23
29
33
8
16
2
17
7
20
22
19
24
30
27
25
3
34
32
12
18
6
5
35
13
1
15
9
10
14

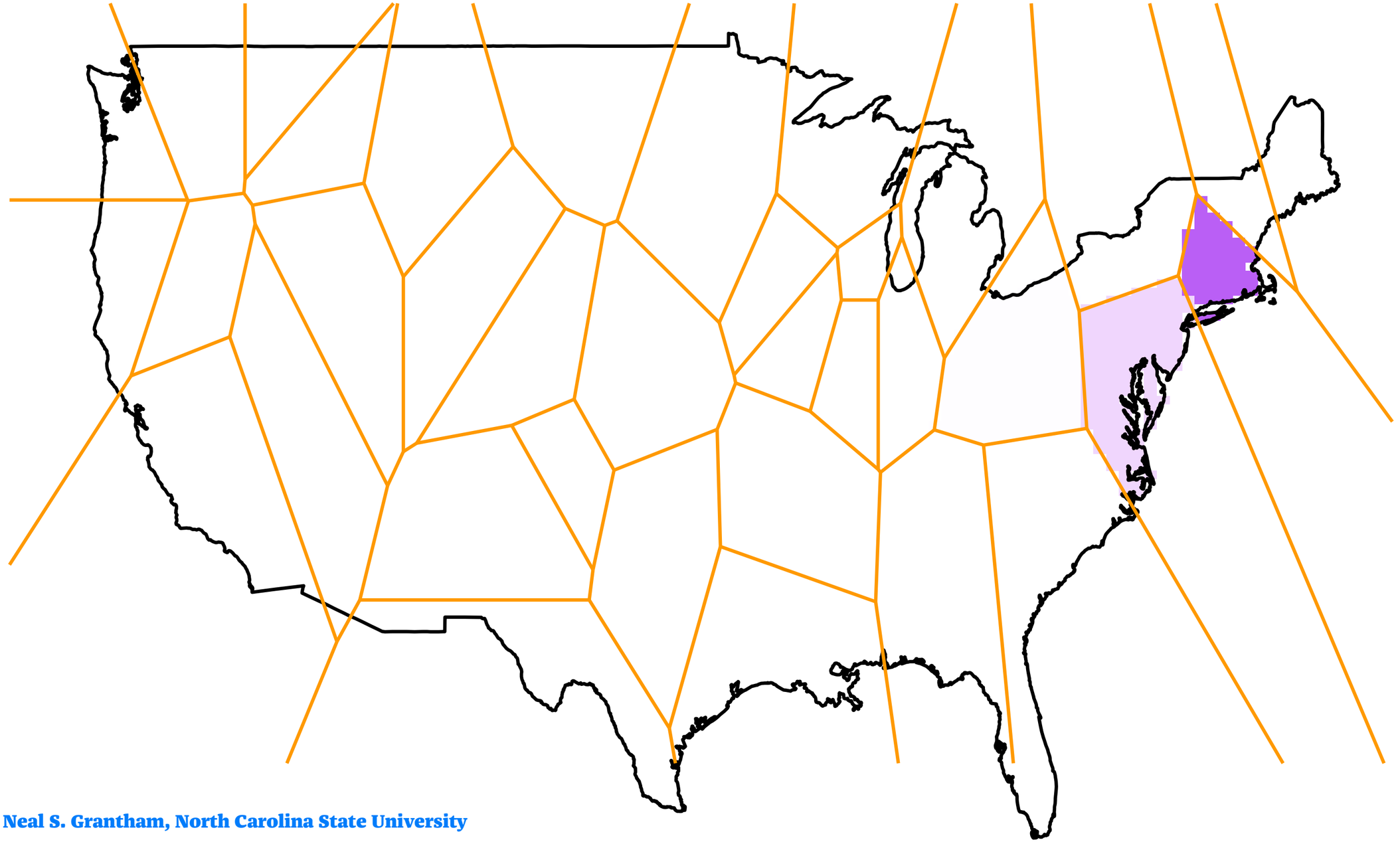


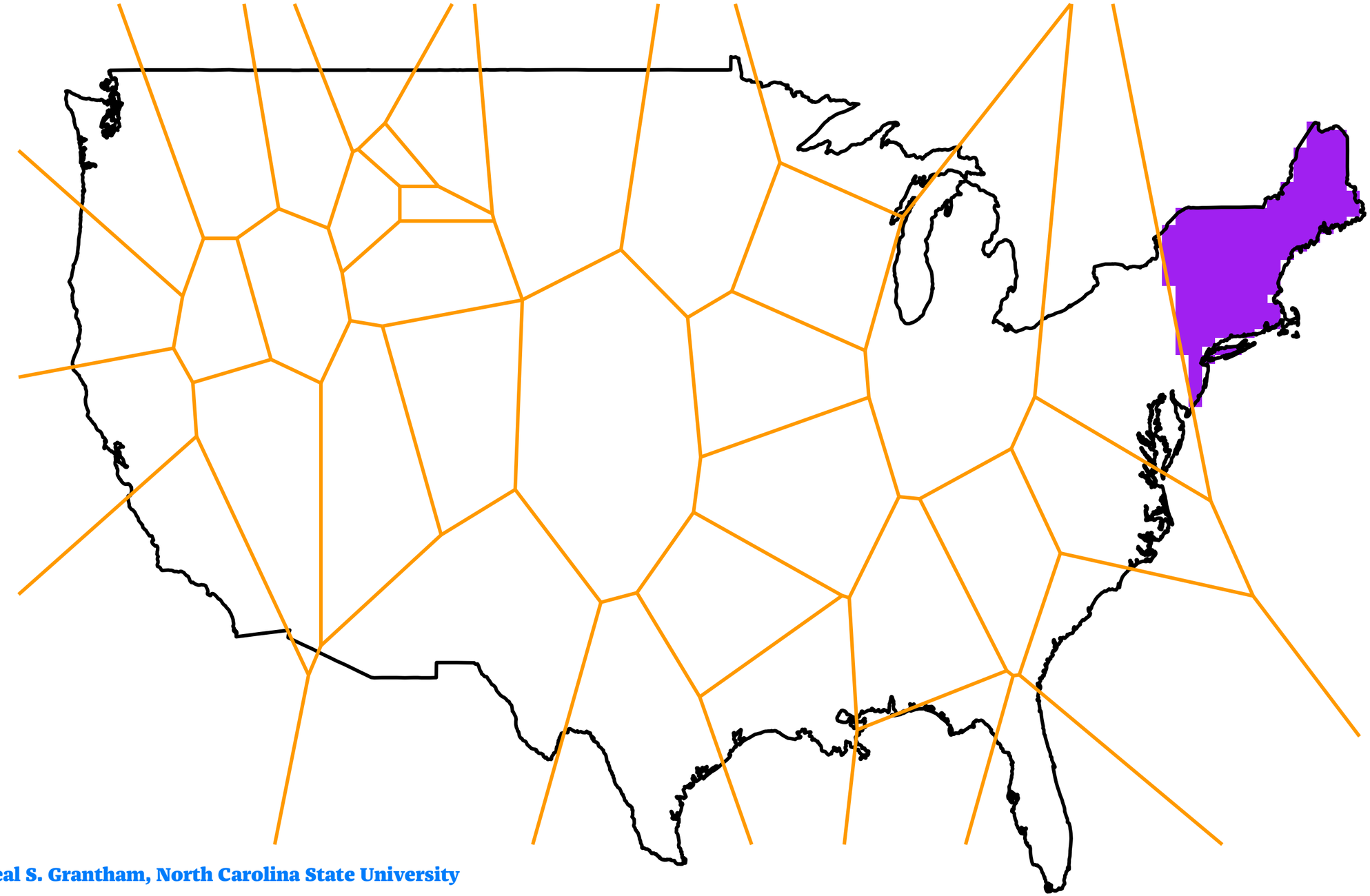
3. Repeat steps 1 & 2 N times to develop a diverse collection, $\mathcal{M}$, of trained DNNs.

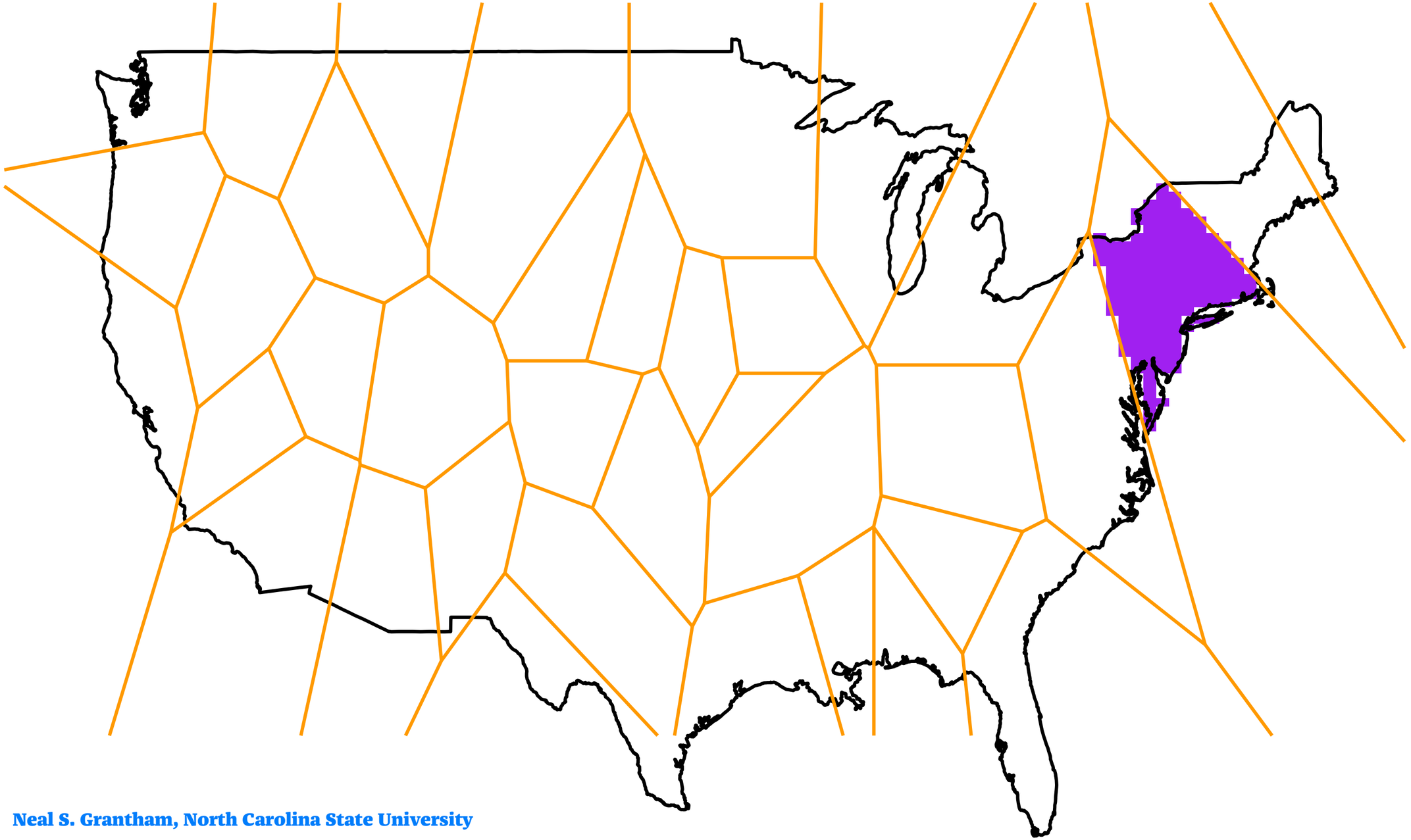


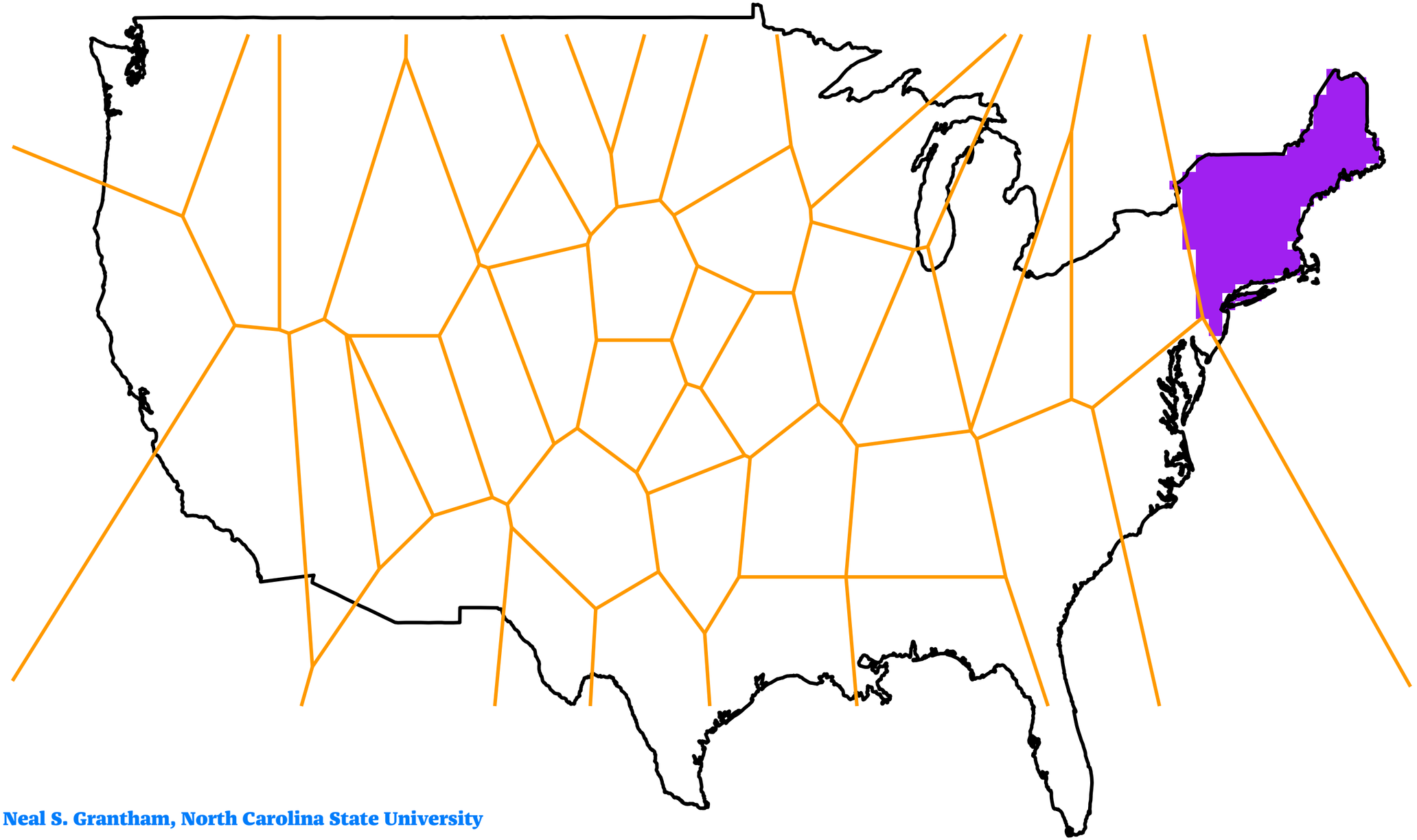


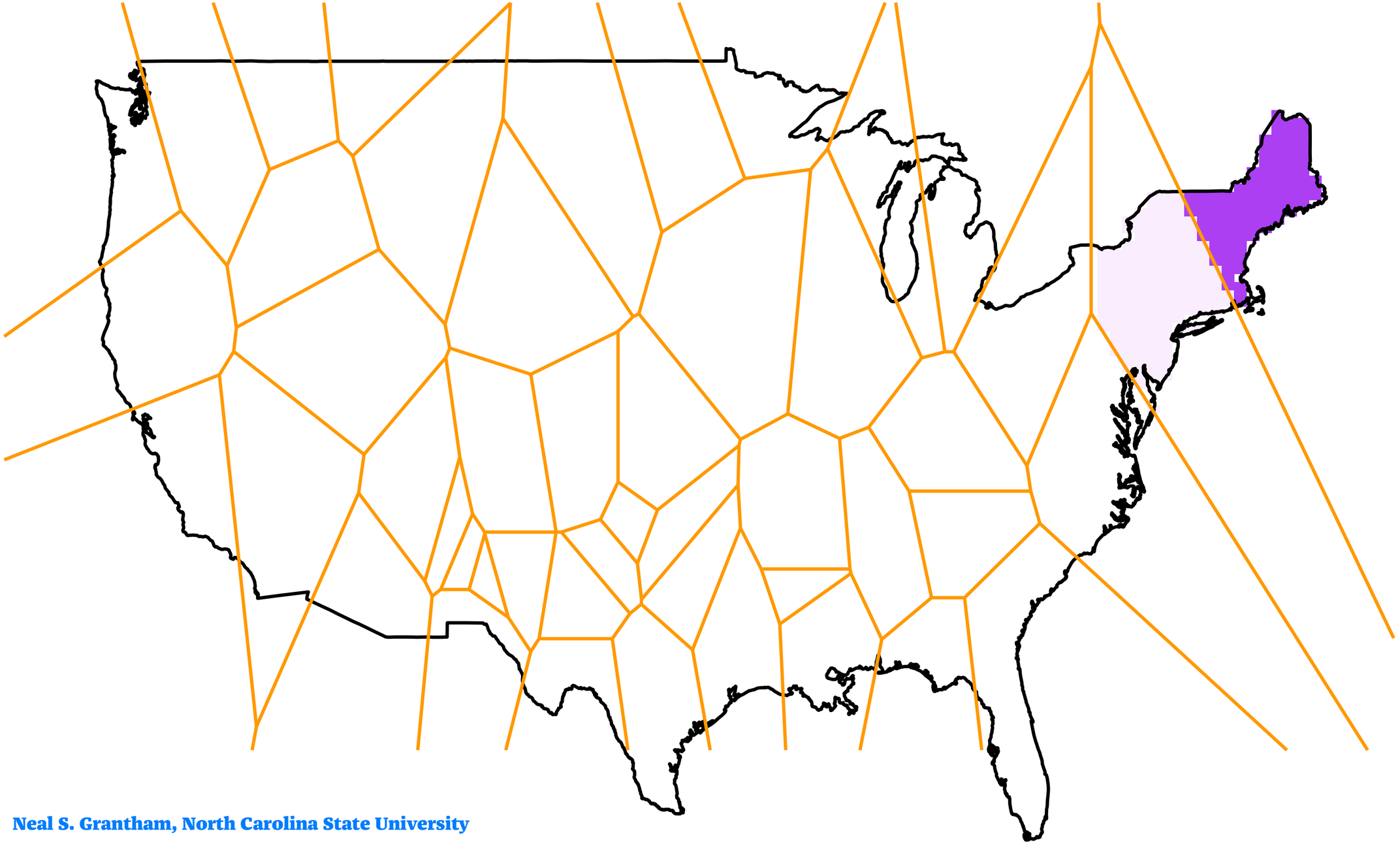


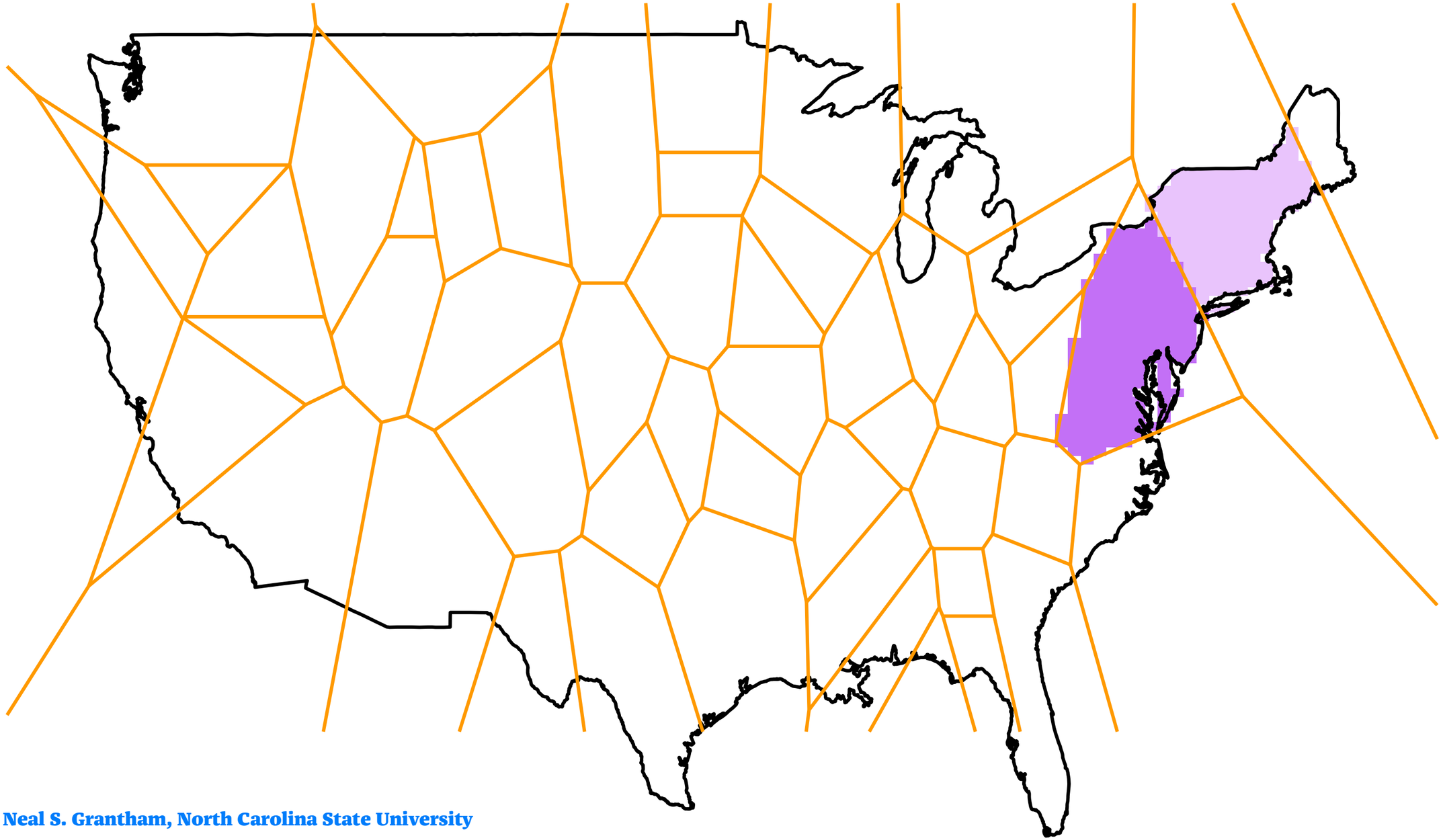












4. Predict most likely origin $\hat{s}$ by averaging over DNNs in $\mathcal{M}$.

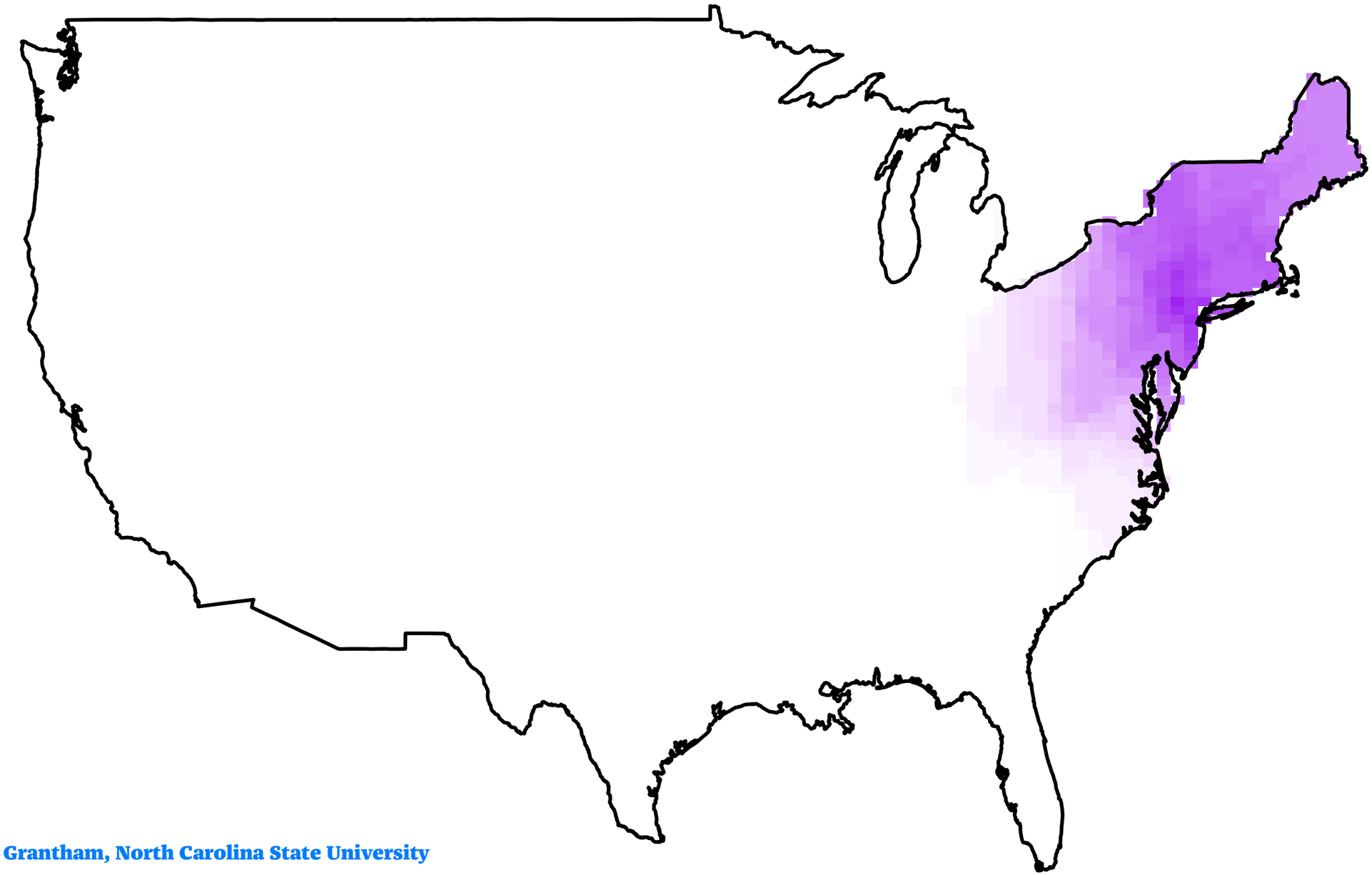
Geolocation

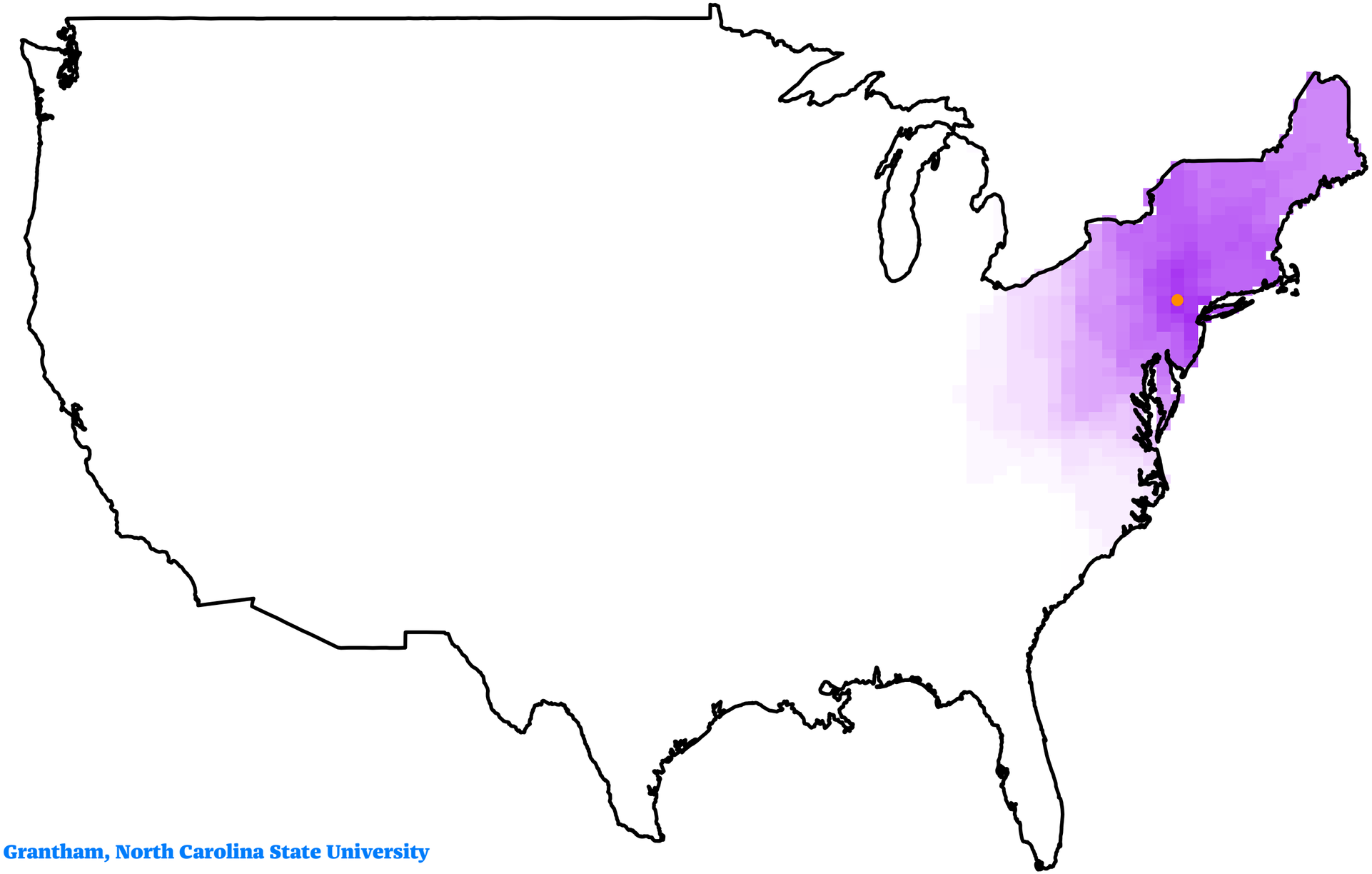
A sample with microbiome $\boldsymbol{x}$ is most likely to have originated from

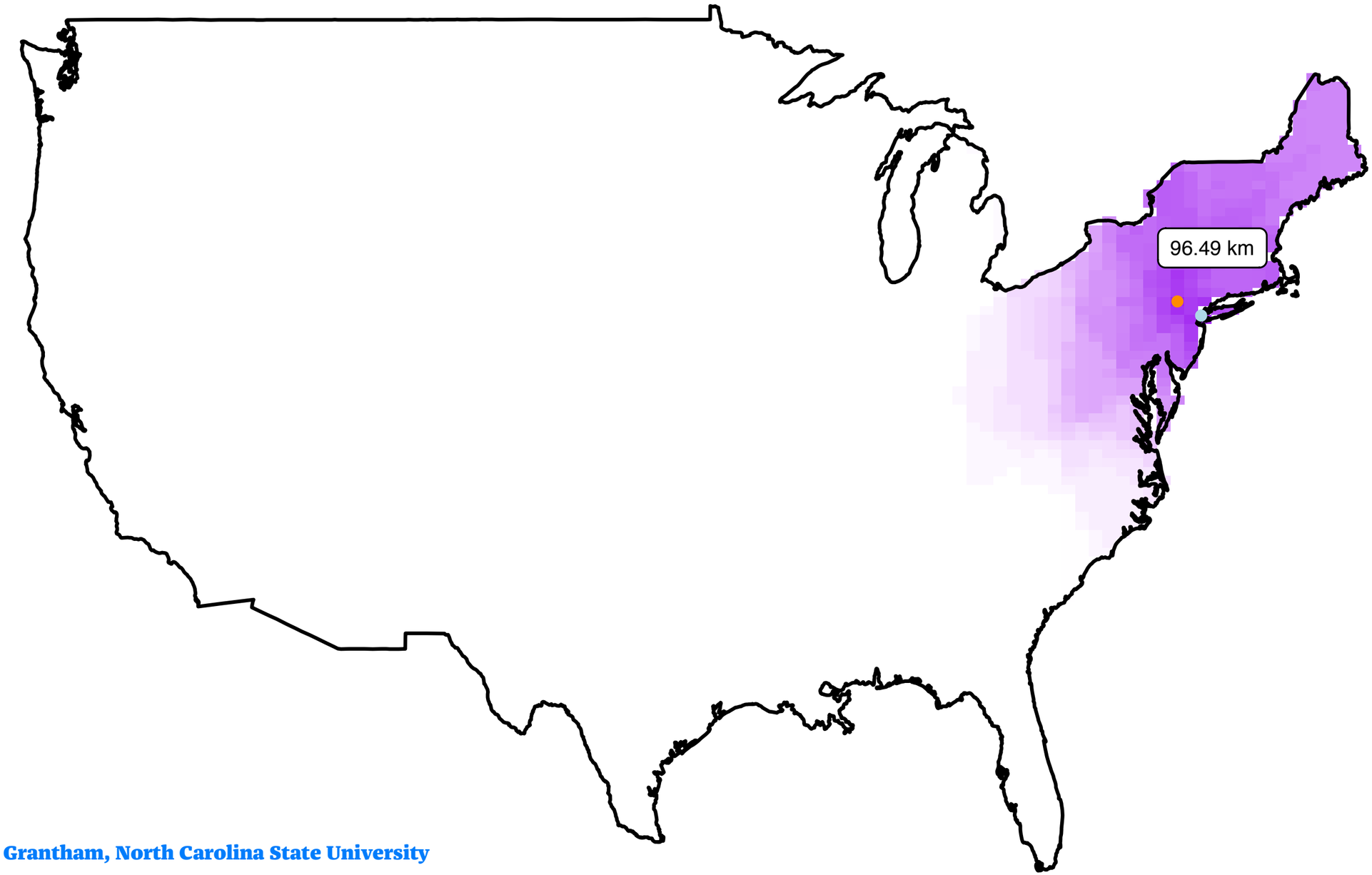
$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s} \in \mathcal{D}} g(\boldsymbol{s} \mid \boldsymbol{x}, \mathcal{M})$$

where $\boldsymbol{g}(\cdot)$ is the geolocation function given by

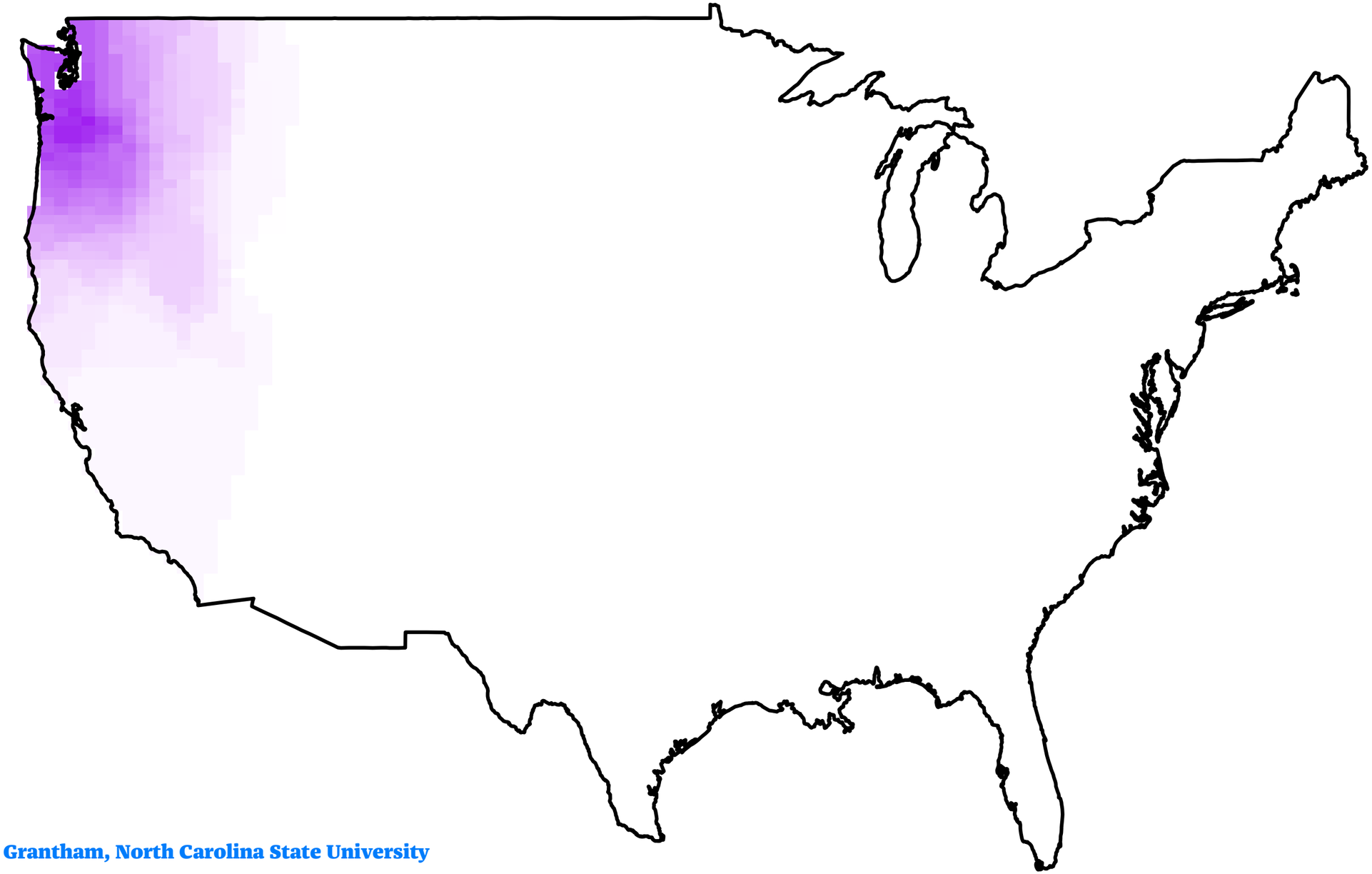
$$g(\boldsymbol{s} \mid \boldsymbol{x}, \mathcal{M}) = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{K_j} \frac{1}{|P_{jk}|} Pr(\boldsymbol{s} \in P_{jk} \mid \boldsymbol{x}) I(\boldsymbol{s} \in P_{jk}).$$

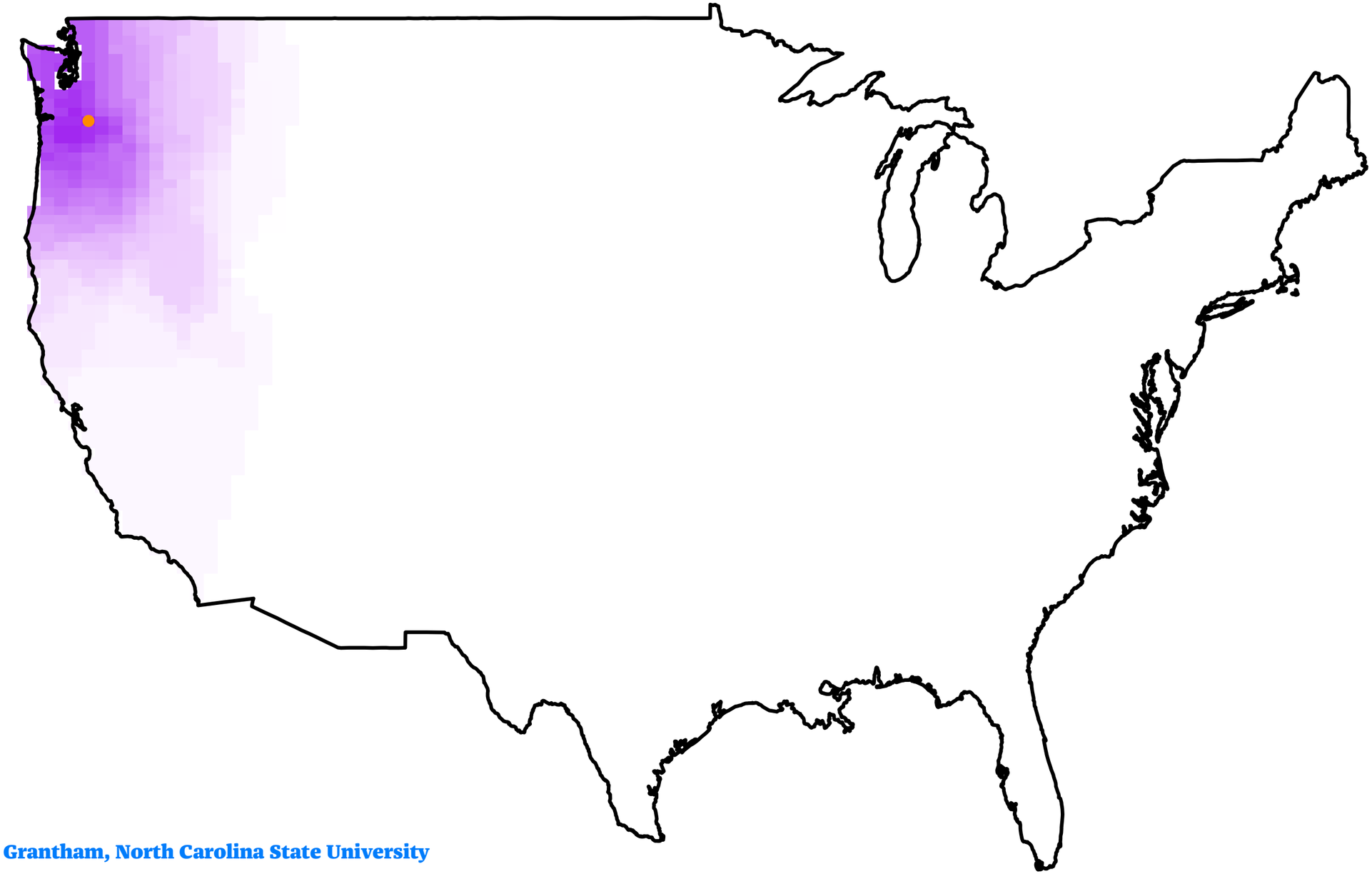


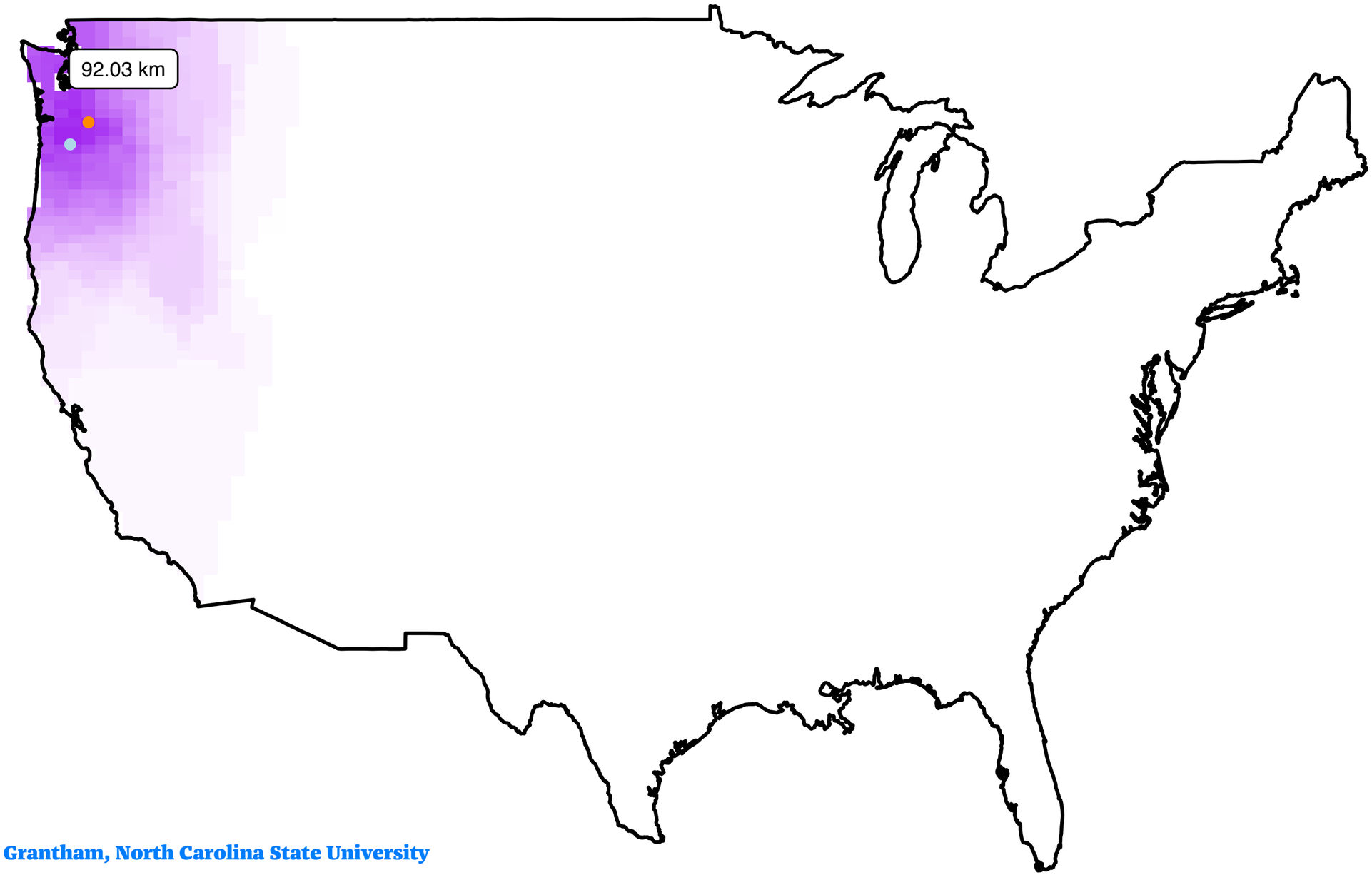


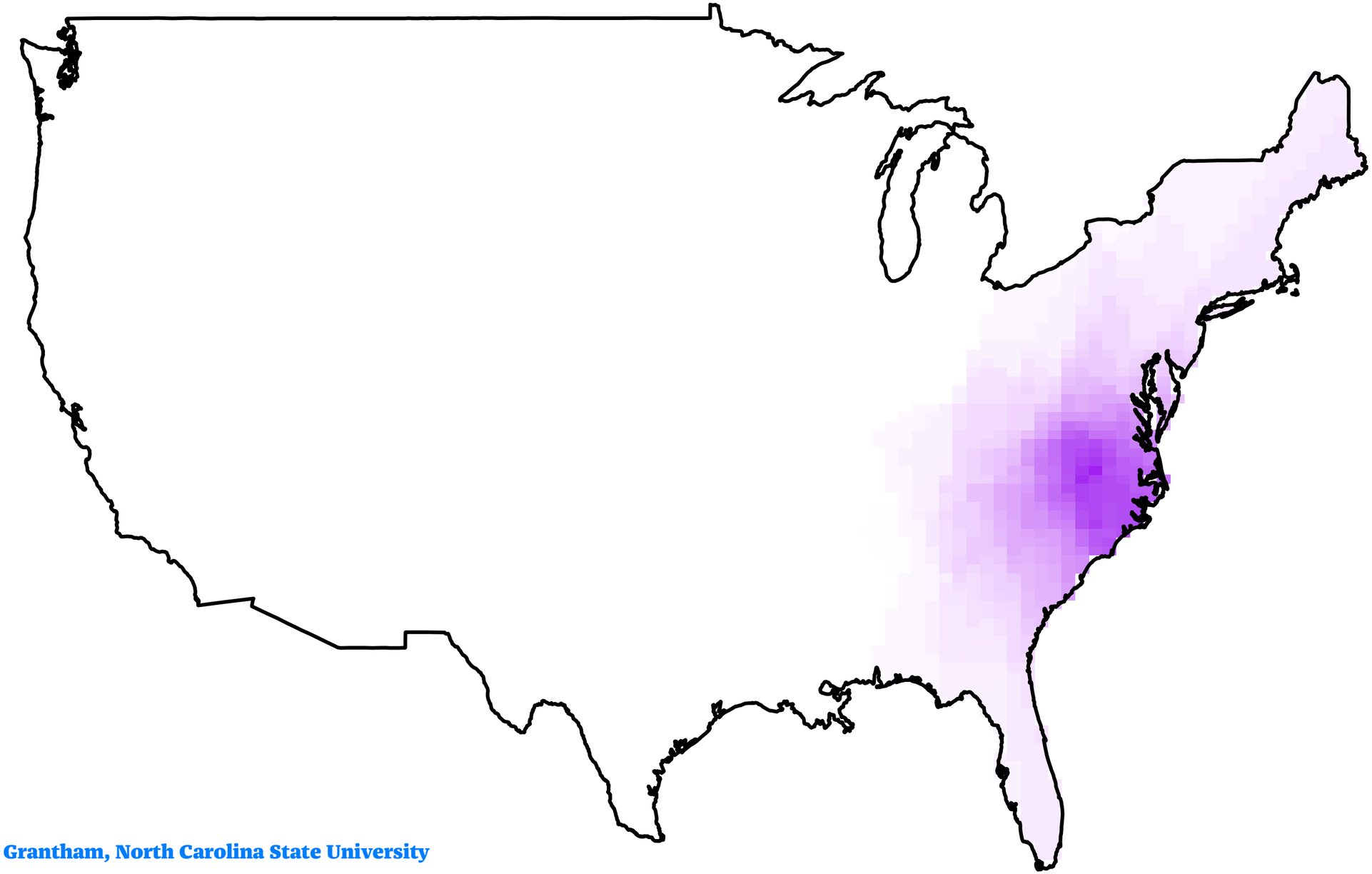

96.49 km

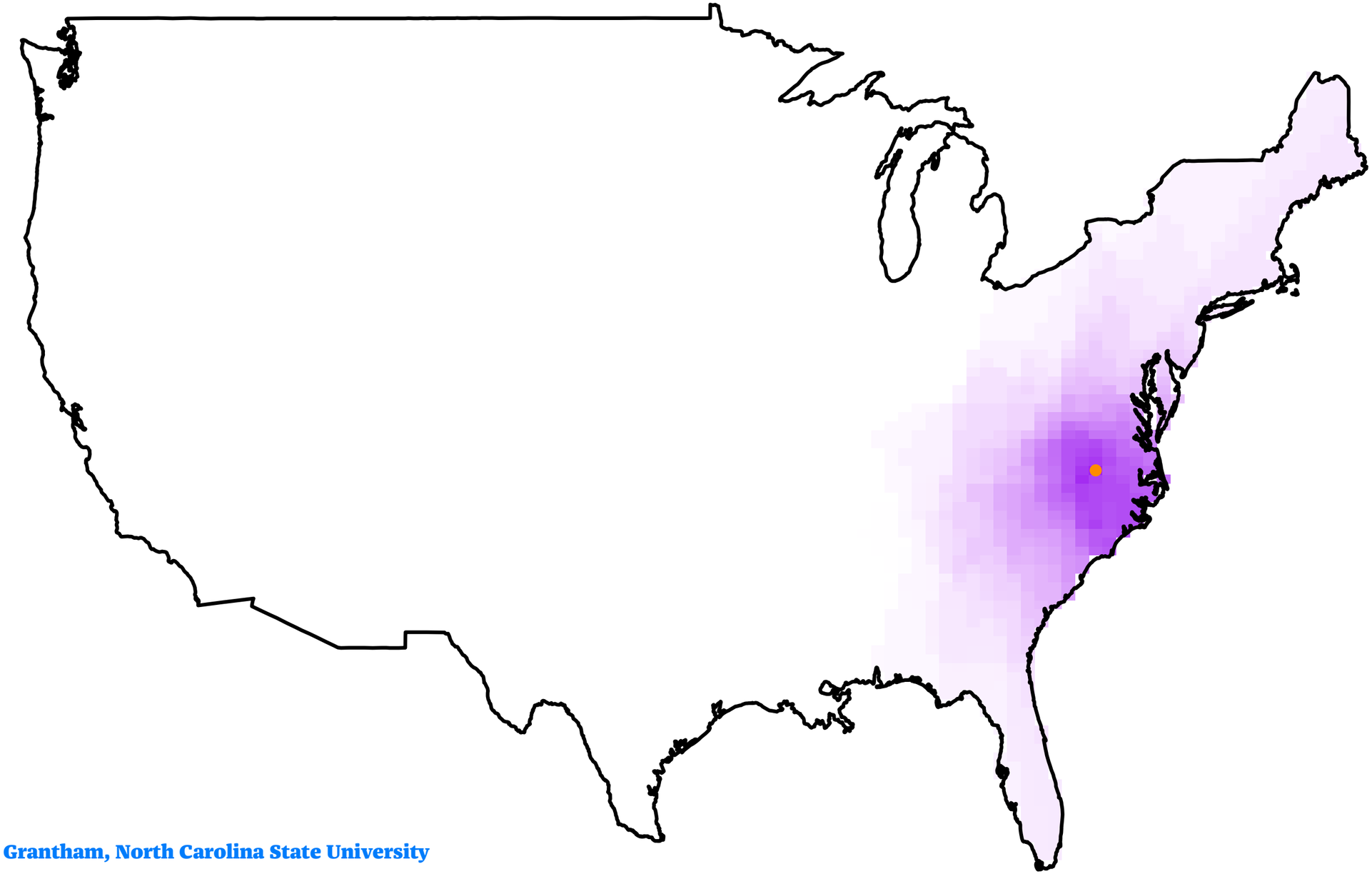
More test sample predictions…

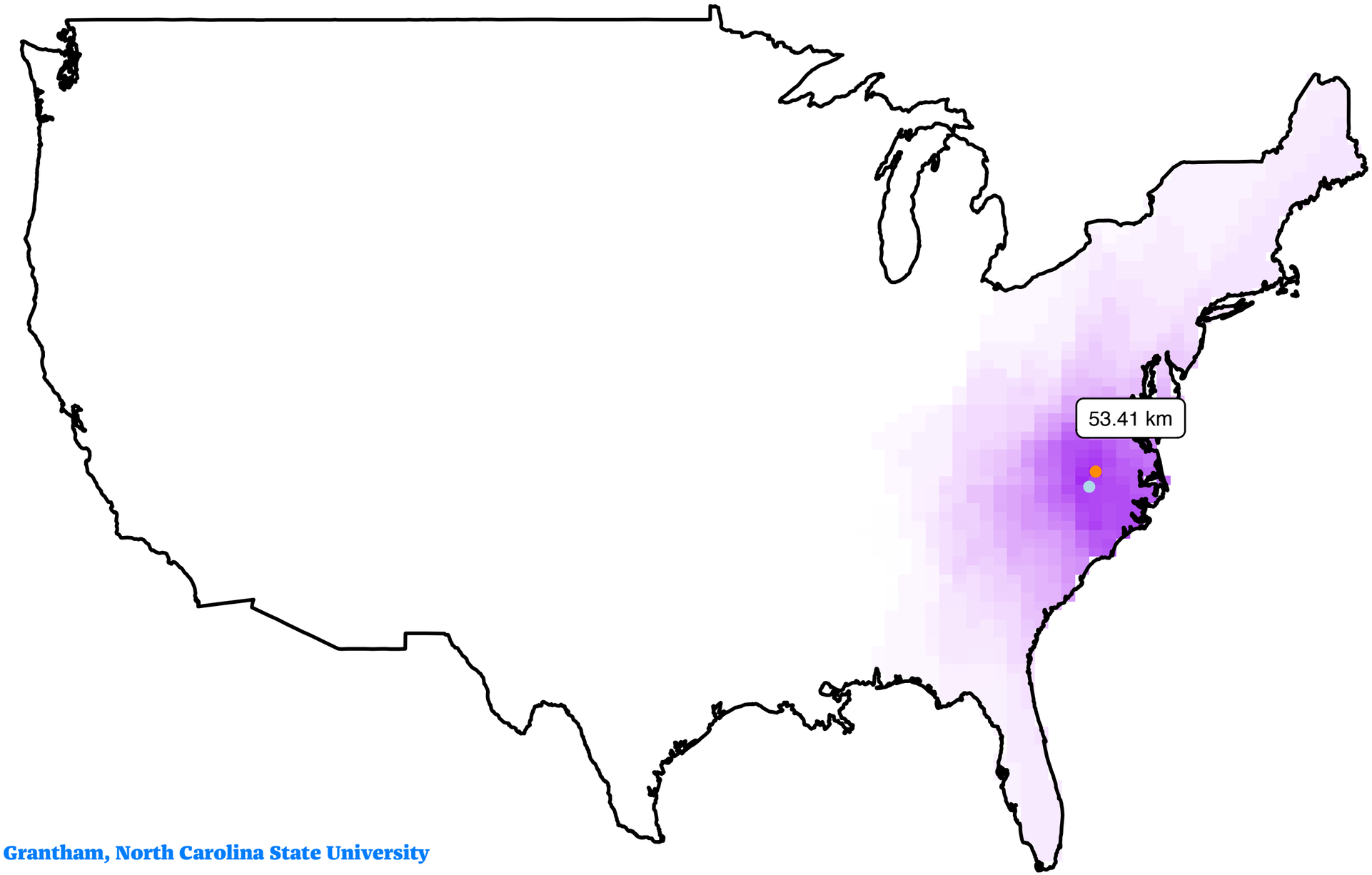


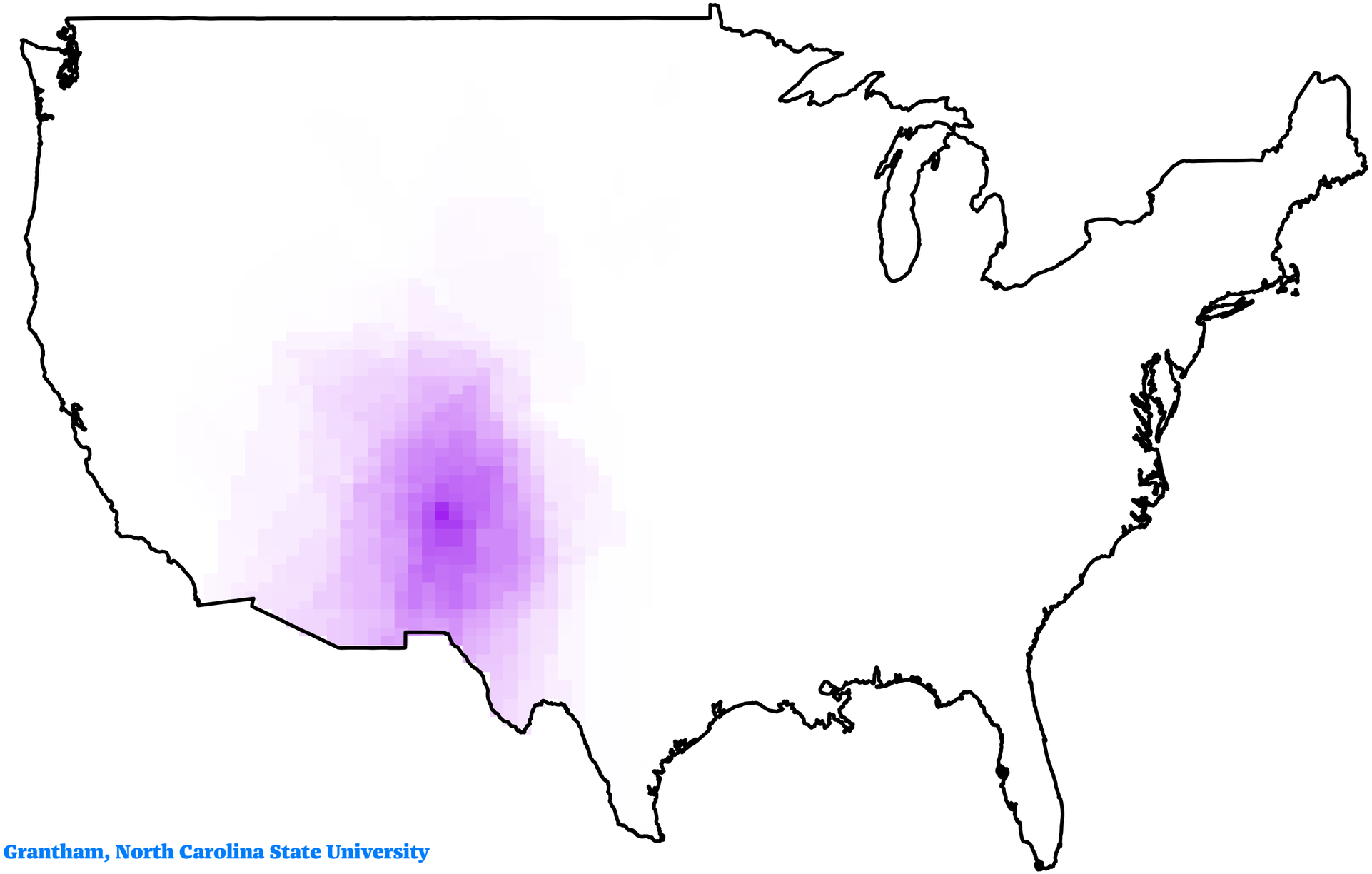


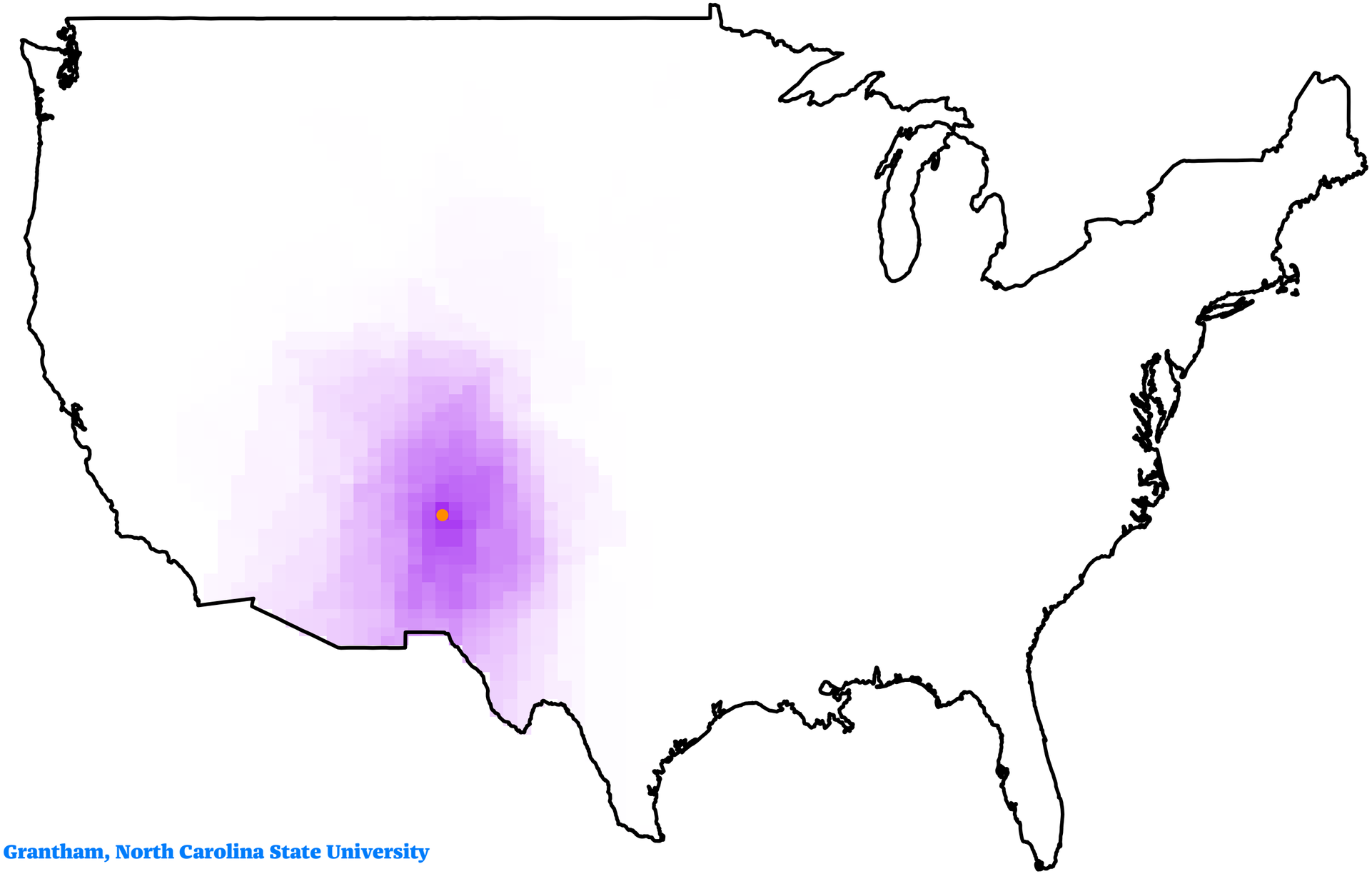

92.03 km

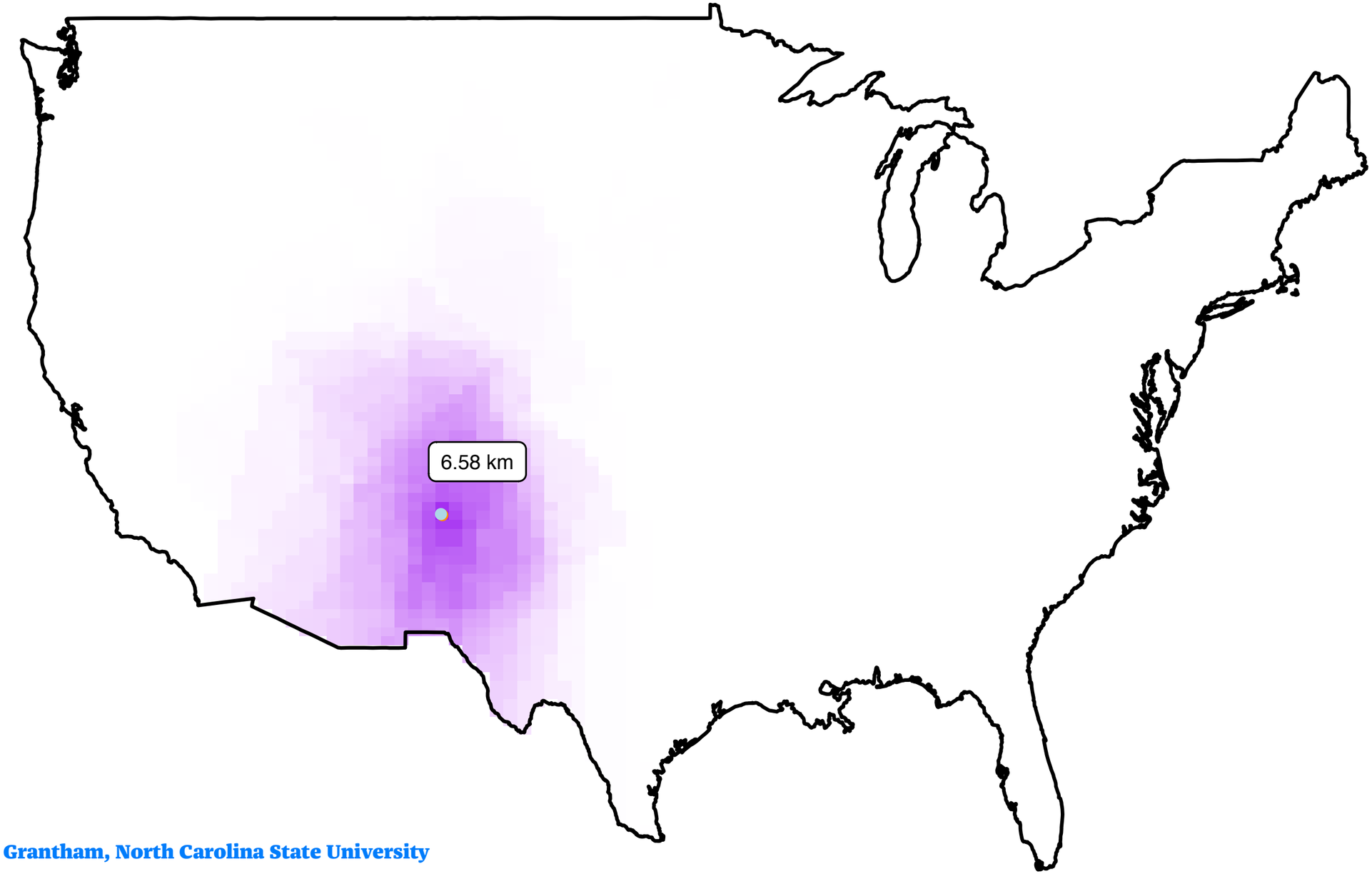


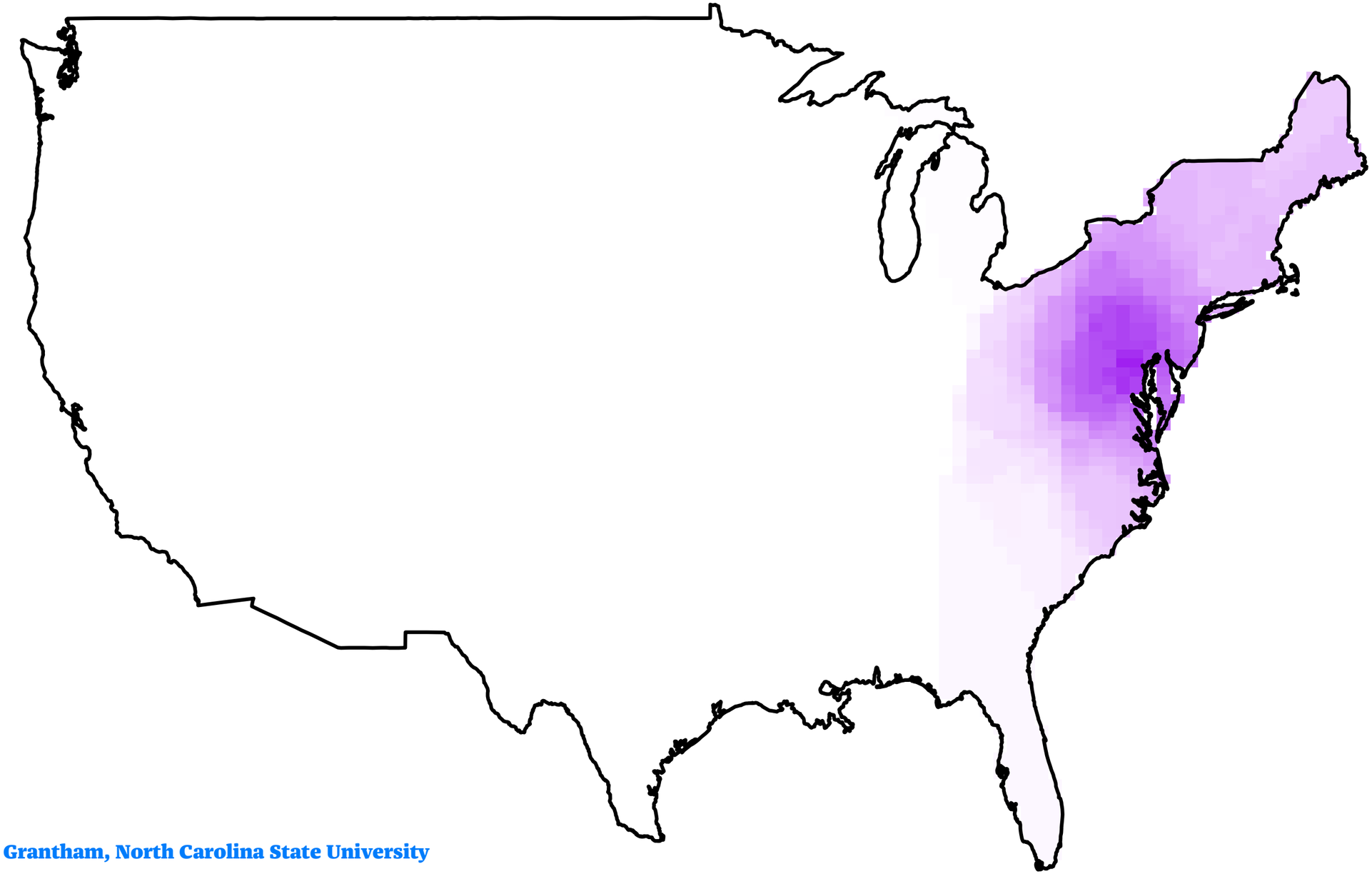


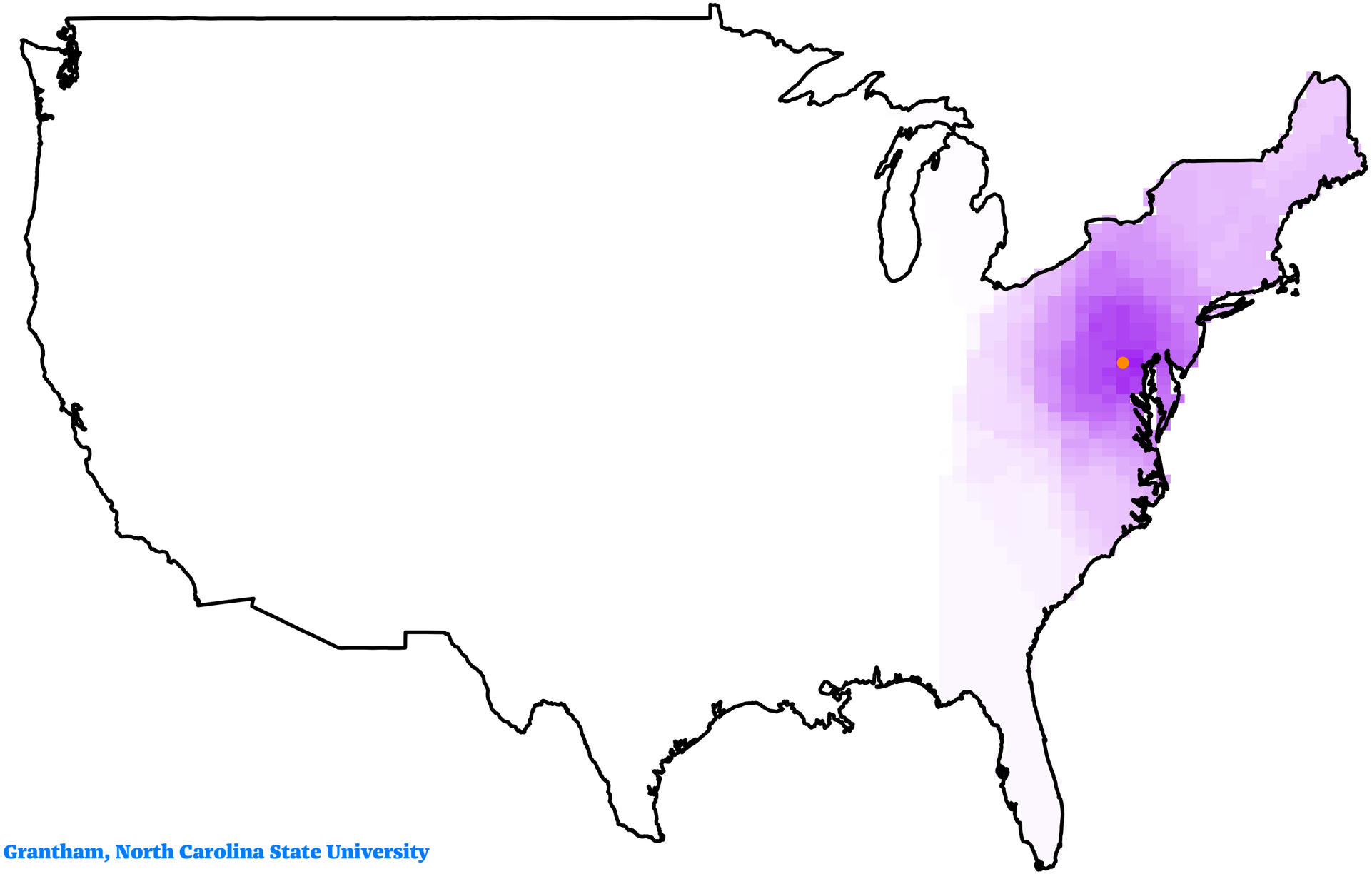

53.41 km

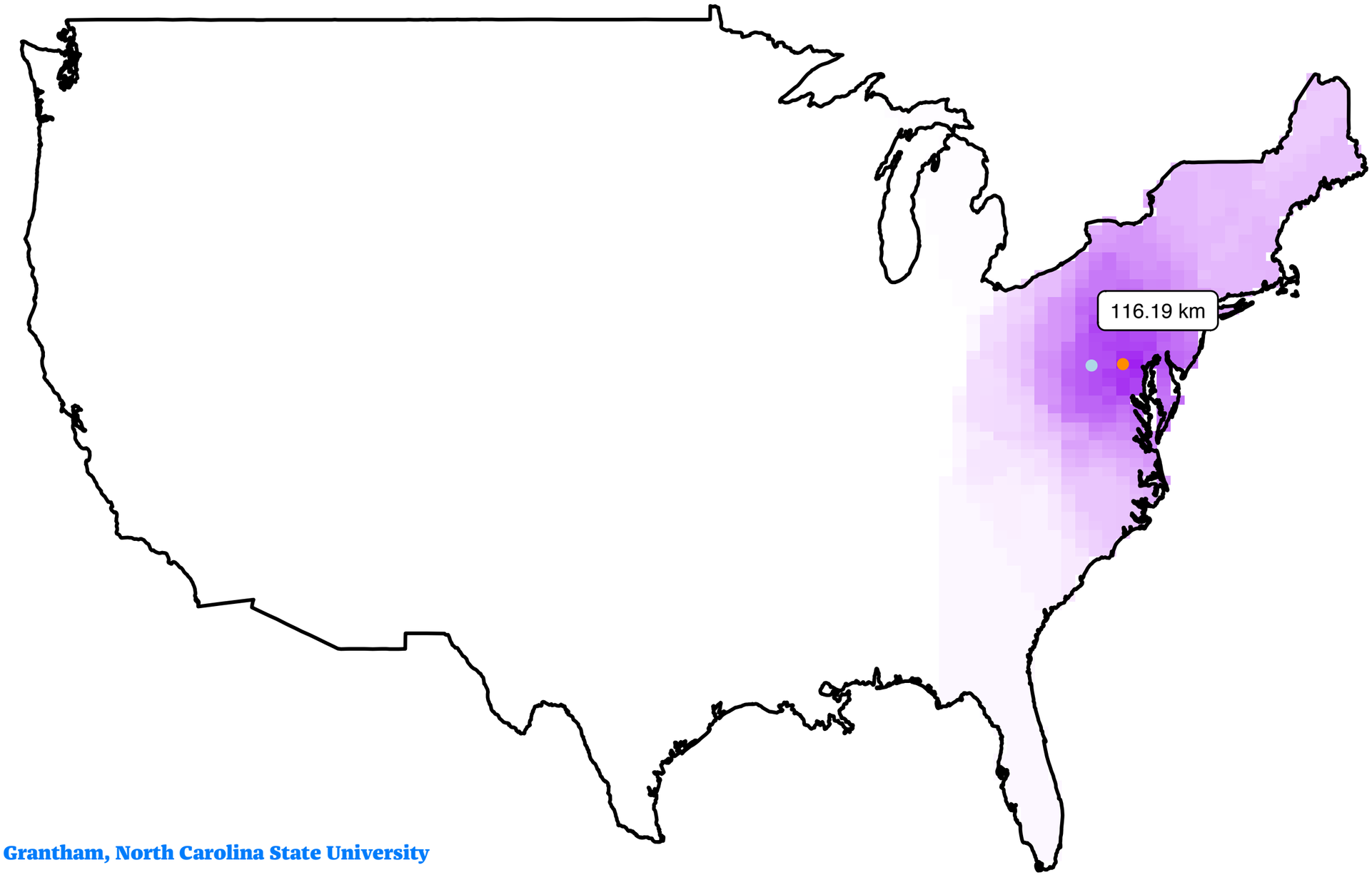




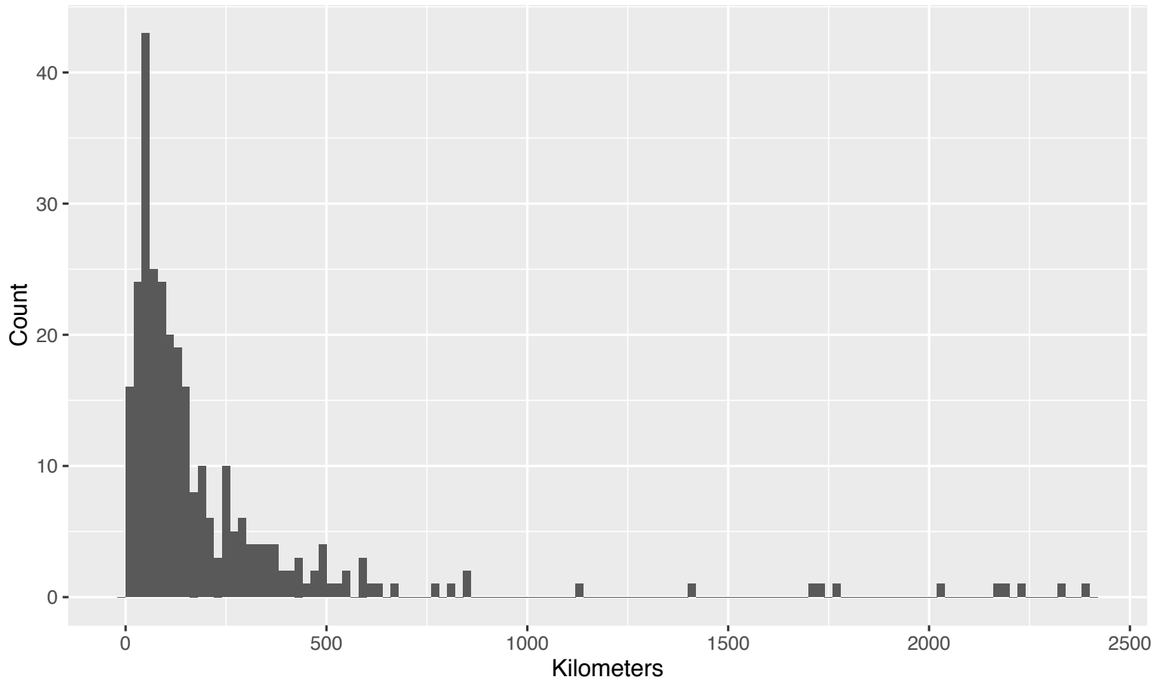

6.58 km






116.19 km

Prediction errors




```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
  1.751   53.990  114.700  230.000  245.500 2394.000 
```

Further work

Develop hypothesis testing framework to test if a sample originates from a particular county, state, etc.

Which fungi are most endemic to different biogeographies? Inference in deep learning is an active area of research.

New applications? The algorithm is not restricted to these data.

Python module `deepspace` to be made available at github.com/nsgrantham/deepspace upon publication.

Thank You! Questions?

Slides available at nsgrantham.github.io/documents/jsm-2016.pdf

Contact me:

☞ Website: nsgrantham.github.io

☞ Github: github.com/nsgrantham

☞ Twitter: twitter.com/nsgrantham