Improving the Quality Of Citizen Science Data

Carl Lagoze
University of Michigan School of Information
October 9, 2012
Microsoft eScience Workshop

Acknowledgments to:

- Steve Kelling (Cornell Lab of Ornithology)
- Weng-Keen Wong (Oregon State CS)
- Theo Damoulas (Cornell CS)
- The National Science Foundation
- All the eBirders

Citizen Science
(or human computation)
(or participatory scholarship)

Harnessing human intelligence to solve computational problems that are beyond the scope of existing machine intelligence:

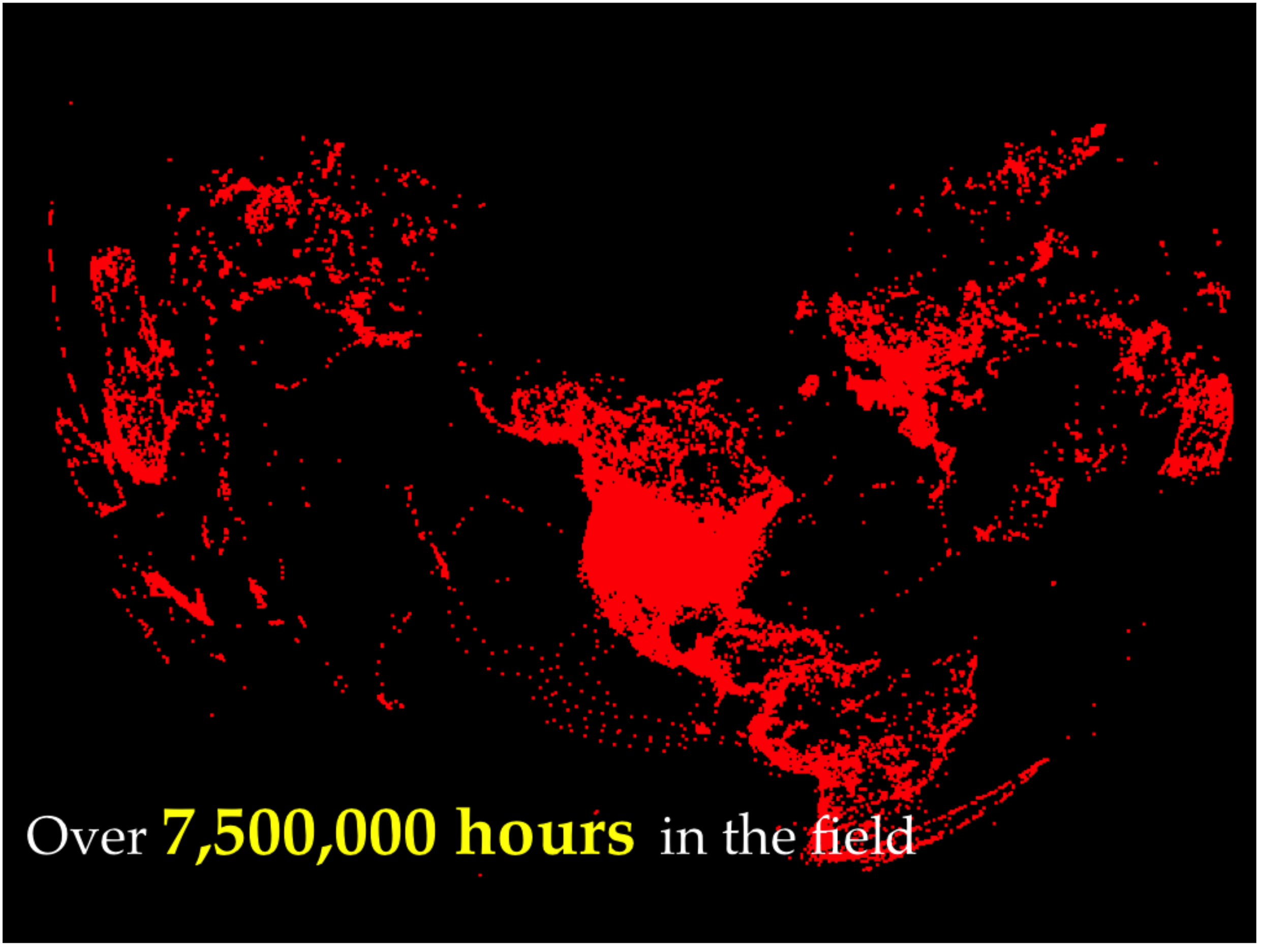
- ESP Game – image labels
- FoldIt – protein structure
- Galaxy Zoo – galaxy classification
- reCAPTCHA – OCR of old print material

eBird

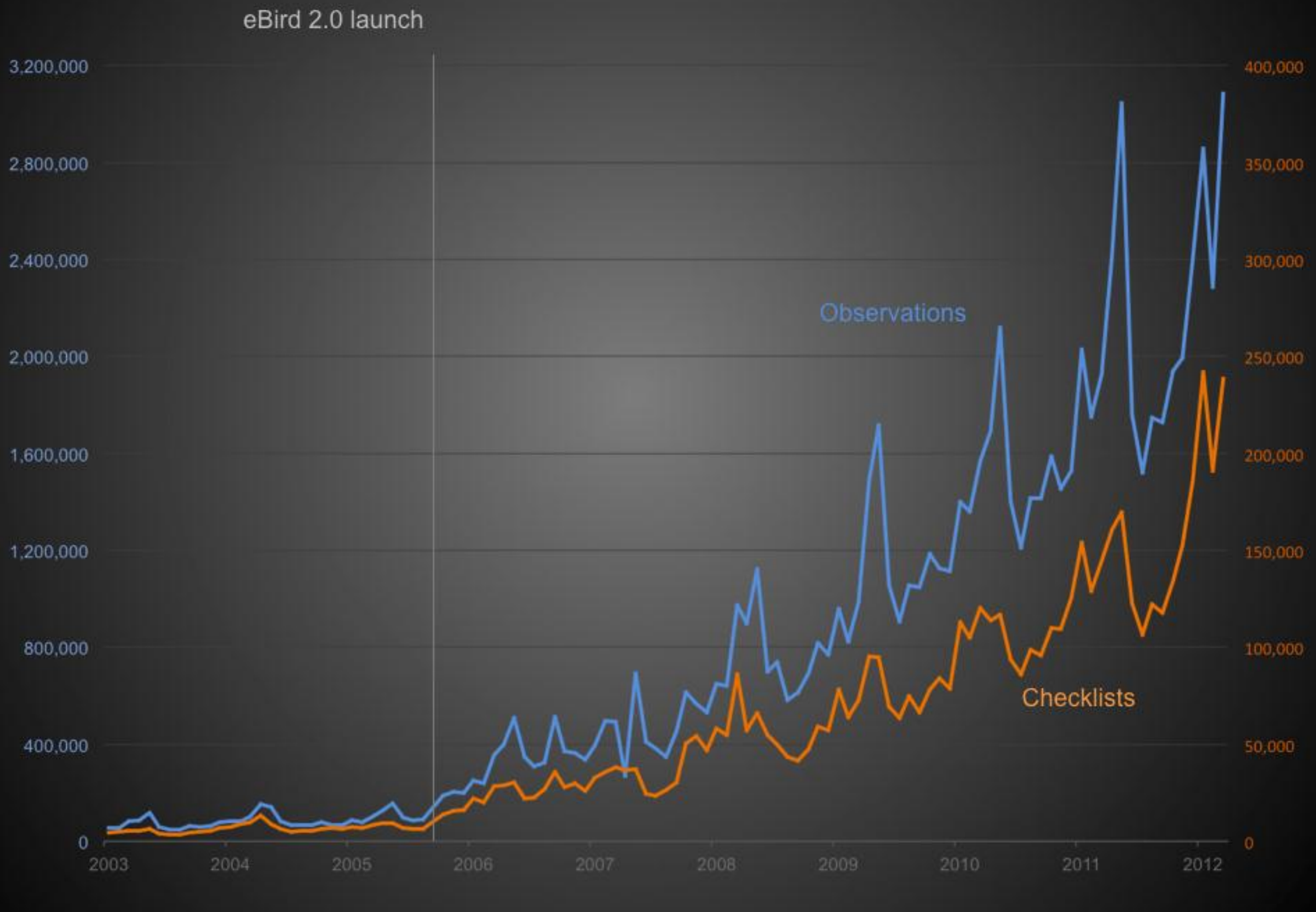
The Cornell Lab of Ornithology

eBird

- More than …
  - **100** million observations submitted
  - 7.5 million checklists entered
  - 85 thousand contributors
  - 9200 species
  - 220 countries


Over 7,500,000 hours in the field

Growth in eBird Observations and Checklists



Birders

Keys to eBird success among birders

- Ease of use
- Tools for improving personal skills and expertise
- Appealing to birders “benevolent competitiveness”

1 2 3 **What did you see or hear?**

Bird Hills Nature Area, Washtenaw County, Michigan, US Change
Wed Oct 03, 2012 7:00 AM Change

Jump to Species...

Add Species
Alphabetic

Show Rarities
Show Subspecies
Group by Most Likely ?

Shortcuts Preferences

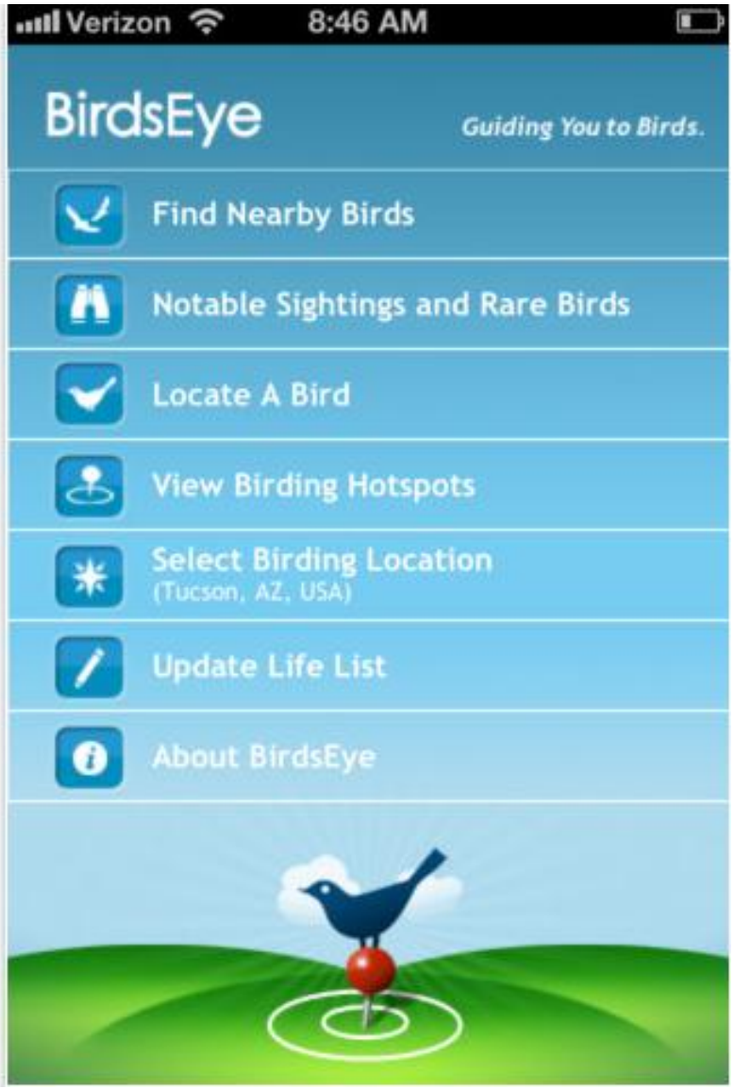
WATERFOWL		Canada Goose
		Cackling/Canada Goose
		Mute Swan
		Trumpeter Swan
		Trumpeter/Tundra Swan
		swan sp.
		Wood Duck
		Gadwall
		American Wigeon
		American Black Duck
		Mallard
		Mallard (Domestic type)
		American Black Duck x Mallard (hybrid)
		Blue-winged Teal
		Northern Shoveler
		Northern Pintail
		Green-winged Teal
		teal sp.
		Canvasback
		Redhead
		Ring-necked Duck
		Greater Scaup
		Lesser Scaup
		Greater/Lesser Scaup
		Common Goldeneye
		Hooded Merganser
		Common Merganser
		Red-breasted Merganser
		Ruddy Duck
		duck sp.
GROUSE, QUAIL AND ALLIES		Ring-necked Pheasant
		Wild Turkey
LOONS AND GREBES		Common Loon
		loon sp.
		Pied-billed Grebe
		Horned Grebe
CORMORANTS, ANHINGAS, AND PELICANS		Double-crested Cormorant

Are you submitting a **complete checklist** of the birds you were able to identify?

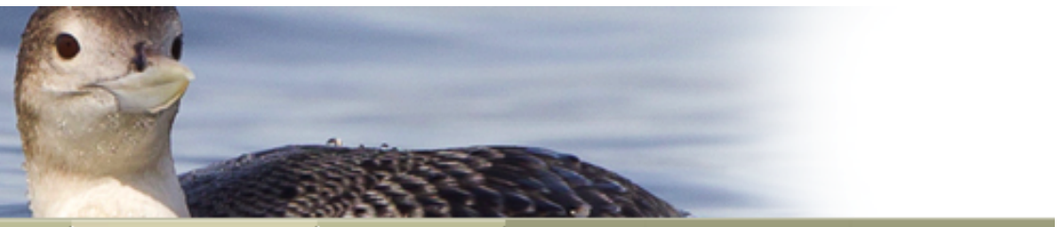
Yes No ?

◀ Submit ▶


Verizon
9:05 AM
Back
Vermilion Flycatcher
© VIREO/ Glenn Bartley


Verizon
8:46 AM
BirdsEye
Guiding You to Birds.
Find Nearby Birds
Notable Sightings and Rare Birds
Locate A Bird
View Birding Hotspots
Select Birding Location
(Tucson, AZ, USA)
Update Life List
About BirdsEye

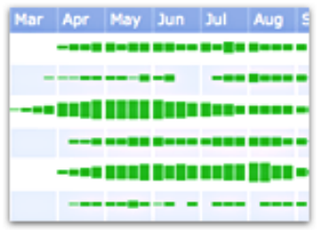
eBird



View and Explore Data

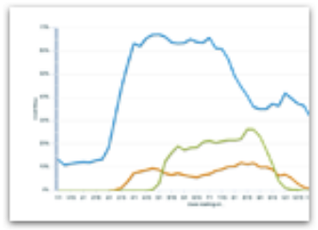
Range and Point Maps

Explore interactive range maps by species or subspecies — zoom in for details



Bar Charts

Find out what birds to expect throughout the year in a region or location



Line Graphs

Explore different metrics of species occurrence in a region or location

Your Totals

Track your totals and compare with other eBirders.

Yard Totals
How many species and checklists have you submitted for your yard?

Patch Totals
How many have you submitted for your favorite birding patches?

Top 100
Compare with the top eBirders in your region.

Arrivals and Departures
Arrivals and departures for a country, state/province, county, or hotspot

All-Time First/Last Records
All-time records for species arrival and departure in a region

High Counts
Species high counts for a region

Alerts
Reports and email alerts for rarities and species you haven't seen

Summary Tables

Observations summarized by week, month, or year

All Observations | My Observations

Bird Observations

Date Range: Change Date

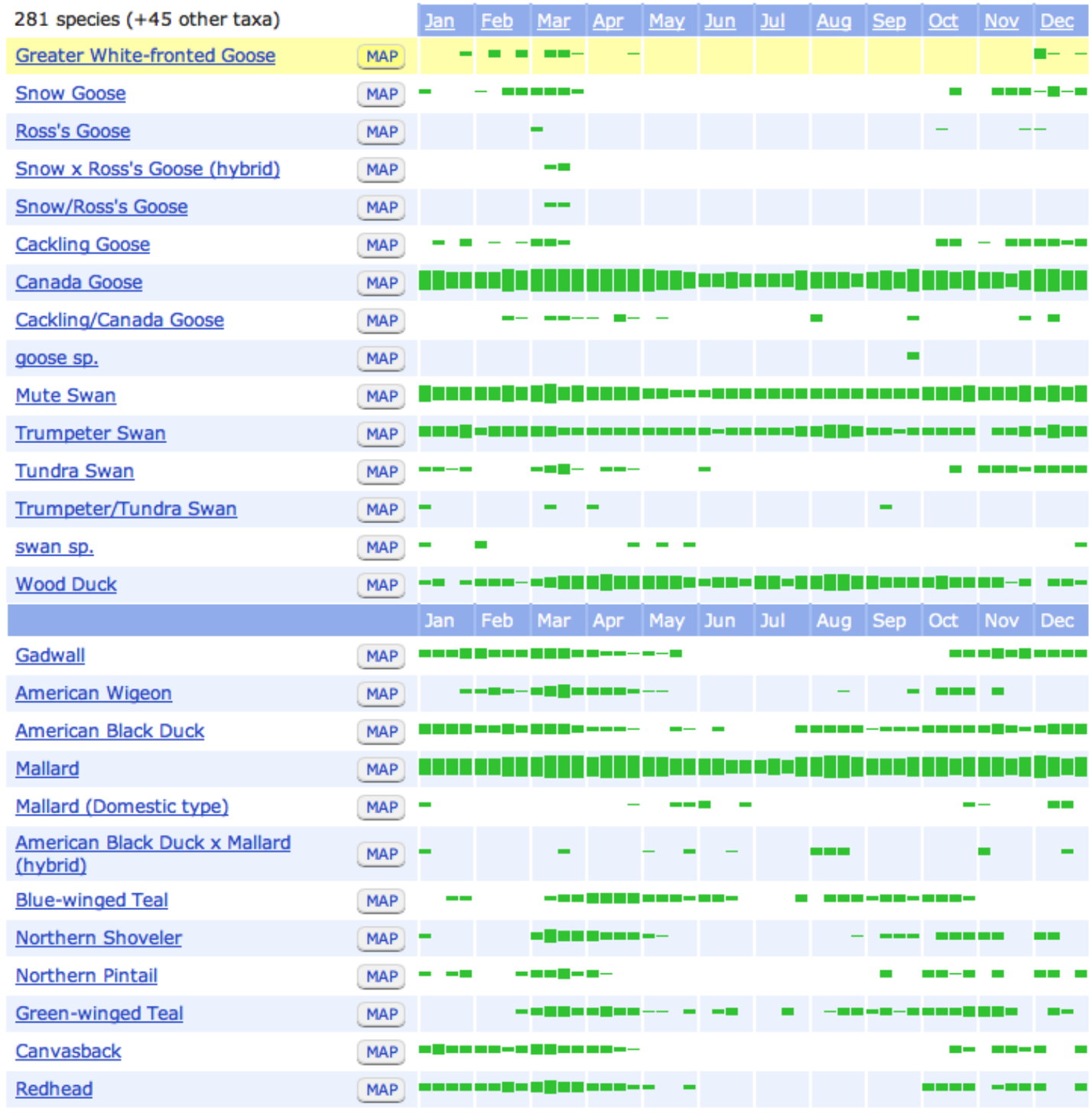
1/1 - 12/31, 1900-2012 **Combine Years**

For Change Location

[Washtenaw]

Last updated ~13 hrs ago.

281 species (+45 other taxa)



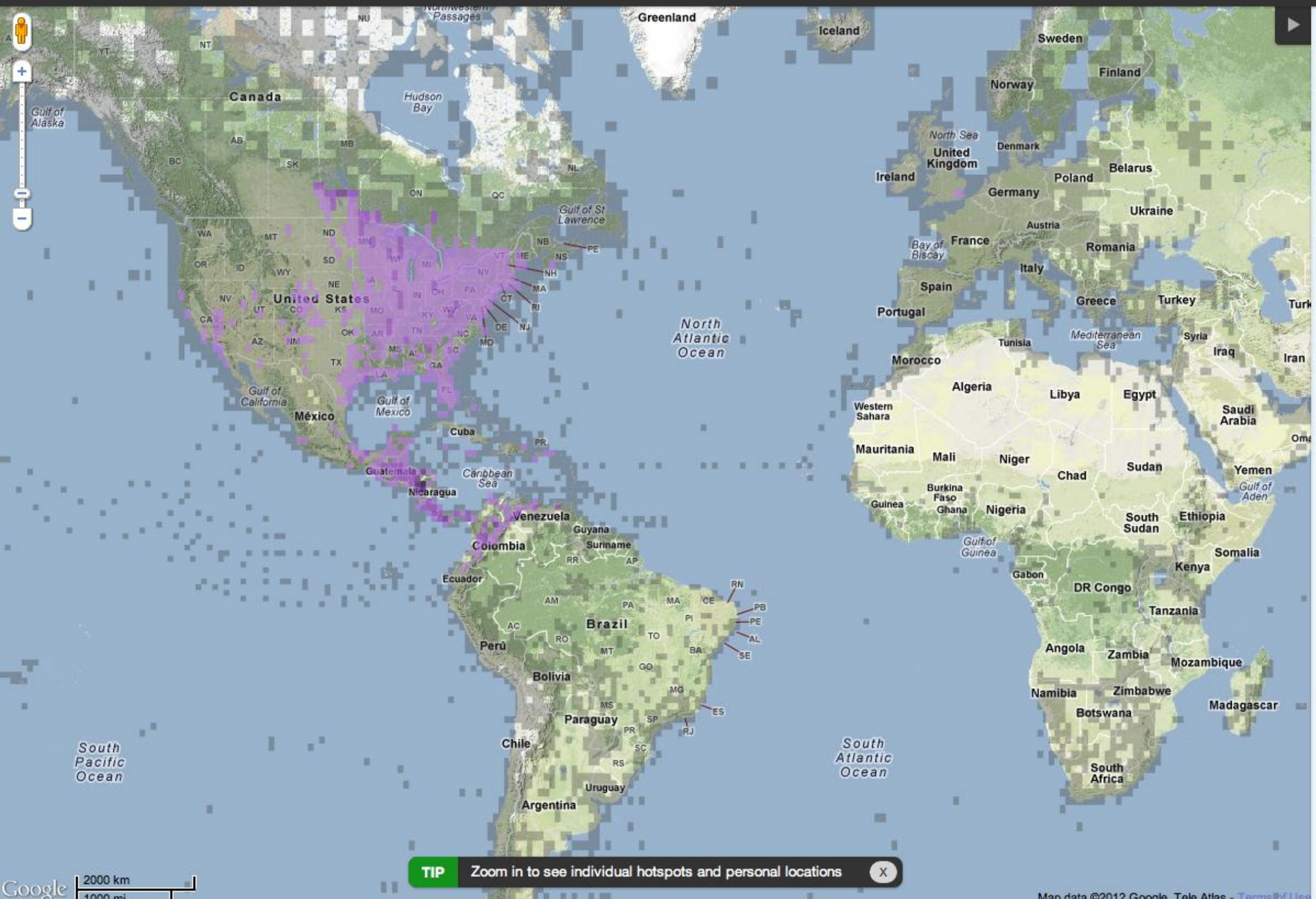
Where can I see a Golden-winged Warbler?

eBird

Year-Round, All Years, Full Range

eBird | Submit Observations | **Explore Data** | My eBird | Chris Wood (zeledonia) | Sign Out | Language | Contact

Species: Golden-winged Warbler (Vermivora chrys | **Date:** Year-Round, All Years | **Location:** Enter place name or address...



TIP Zoom in to see individual hotspots and personal locations X

Zoom Tool

Full Species Range

- Terrain
- Street
- Satellite
- Hybrid

Show Points Sooner

Check this box to show points at broader scales when possible: 2000 points max

FREQUENCY

- 40-100%
- 25-40%
- 10-25%
- 2-10%
- 0-2%
- No Reports

Google 2000 km

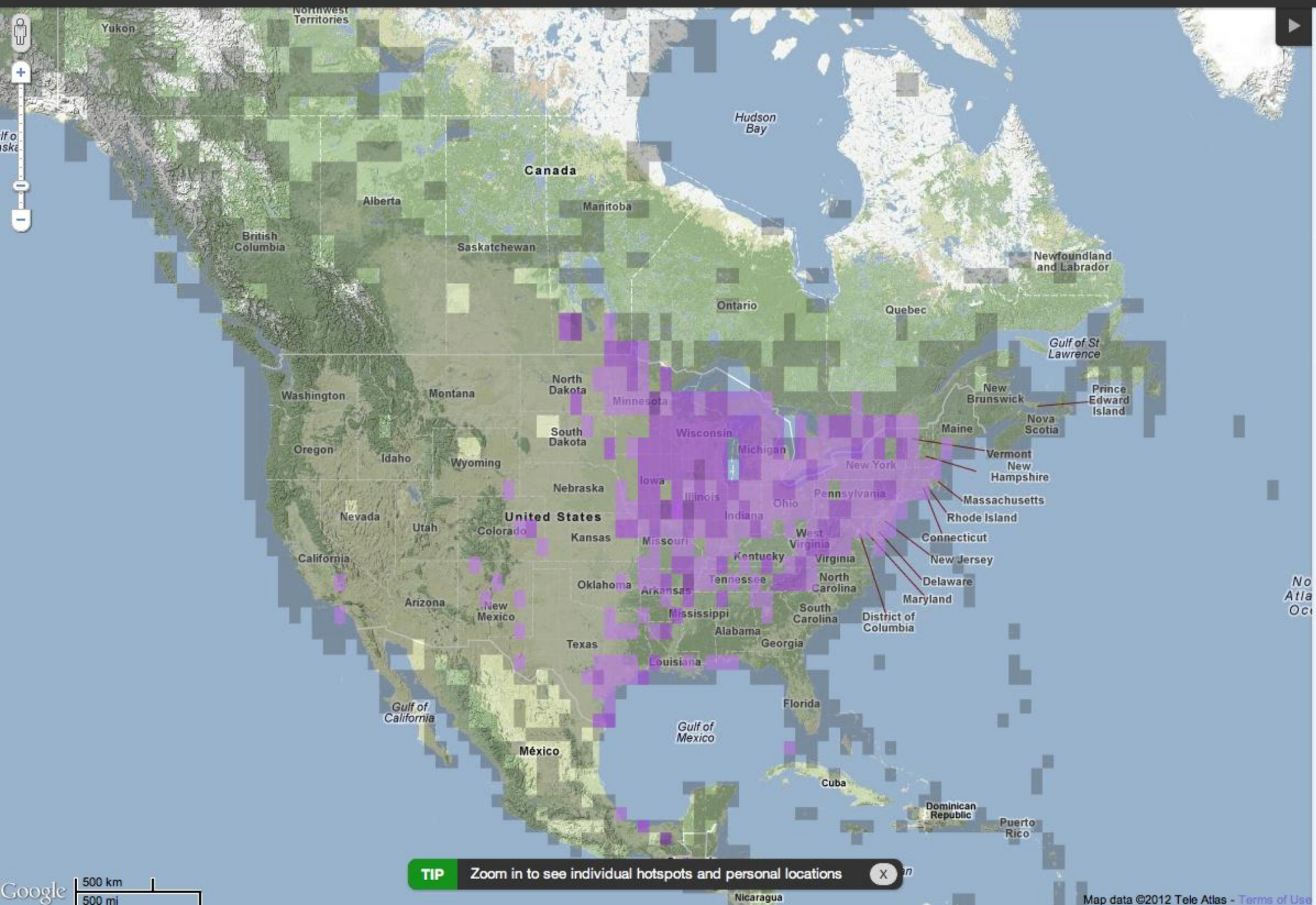
Map data ©2012 Google, Tele Atlas - Terms of Use

eBird

May, All Years, Full Range

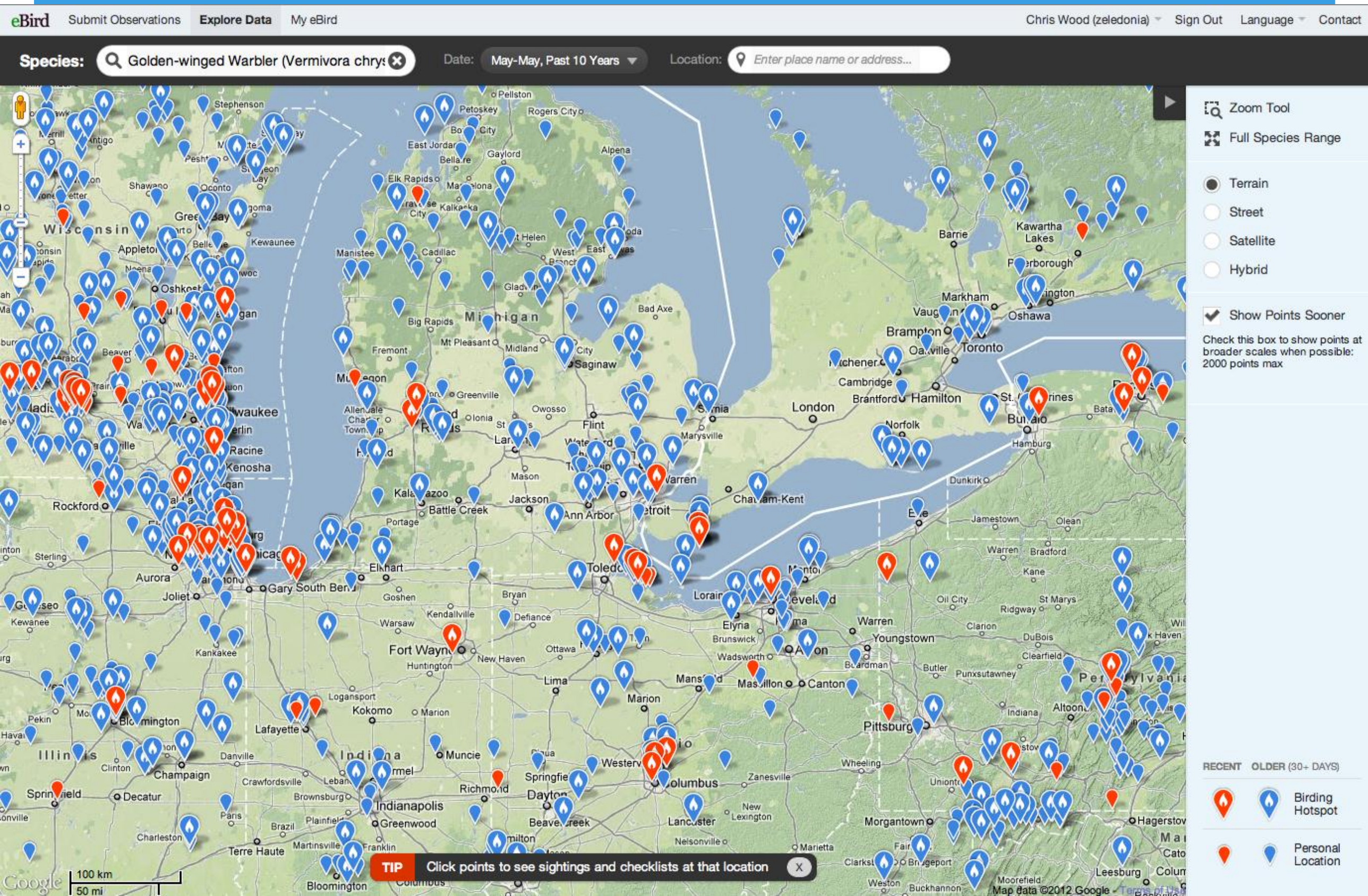
eBird | Submit Observations | **Explore Data** | My eBird | Chris Wood (zeledonia) | Sign Out | Language | Contact

Species: Golden-winged Warbler (Vermivora chrys | Date: May-May, Past 10 Years | Location: Enter place name or address...



eBird

May, All Years, Zoomed



eBird — All sightings from Magee Boardwalk

eBird | Submit Observations | **Explore Data** | My eBird | Chris Wood (zeledonia) | Sign Out | Language | Contact

Species: Golden-winged Warbler (Vermivora chrys | Date: May-May, Past 10 Years | Location: Enter place name or address...

Magee Marsh Wildlife Area--Boardwalk

Lucas, US-OH

DATE	#	OBSERVER	
2012-05-02	1	Ben Warner	Checklist
2011-05-13	1	Ben Winger	Checklist
2011-05-13	1	Sarah Fowler	Checklist
2011-05-13	1	Sarah Fowler	Checklist
2011-05-13	1	Henry Trimpe	Checklist
2011-05-12	1	Jay Bolden	Checklist
2011-05-11	1	Roberta K Allen	Checklist
2011-05-11	1	Christina Harrington	Checklist
2011-05-11	1	Leslie Houser	Checklist
2011-05-07	X	Jen Cawley	Checklist
2011-05-07	1	Patrick Johnson	Checklist
2011-05-06	1	Kenn Kaufman	Checklist
2011-05-06	1	James Fox	Checklist
2011-05-06	2	Annie Crary	Checklist
2011-05-06	2	Patrick Johnson	Checklist
2011-05-06	1	Ken Ostermiller	Checklist
2011-05-06	X	Adam Zorn	Checklist
2011-05-06	X	Clair Cogar	Checklist
2011-05-06	X	Mark Rozmarynowycz	Checklist
2011-05-06	1	Richard Naber	Checklist
2011-05-06	1	Sue Tackett	Checklist
2011-05-06	1	Thomas Dorazio	Checklist

Zoom Tool
Full Species Range

Terrain
Street
Satellite
Hybrid

Show Points Sooner

Check this box to show points at broader scales when possible: 2000 points max

RECENT OLDER (30+ DAYS)

Birding Hotspot
Personal Location

TIP Click points to see sightings and checklists at that location X

Map data ©2012 Google

Kenn Kaufman's 6 May 2011 list

eBird | Submit Observations | Explore Data | My eBird

Checklist S8190527

Location Magee Marsh Wildlife Area--Boardwalk, Lucas County, Ohio, US (Map)

Date and Effort Fri May 06, 2011 5:15 PM

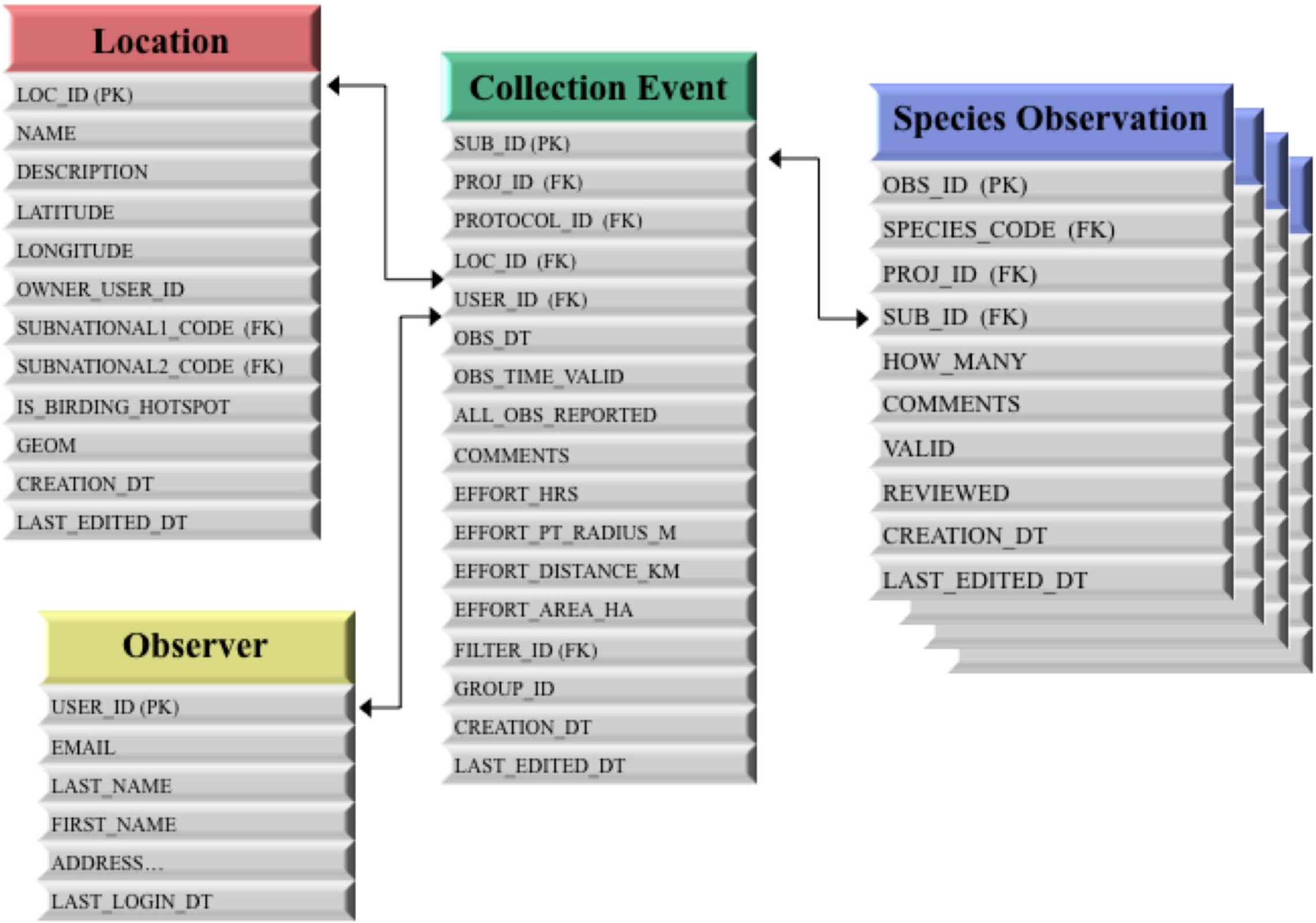
Protocol: Traveling
Party Size: 1
Duration: 1 hour(s), 25 minute(s)
Distance: 0.75 mile(s)
Observers: Kenn Kaufman

Species 80 species total

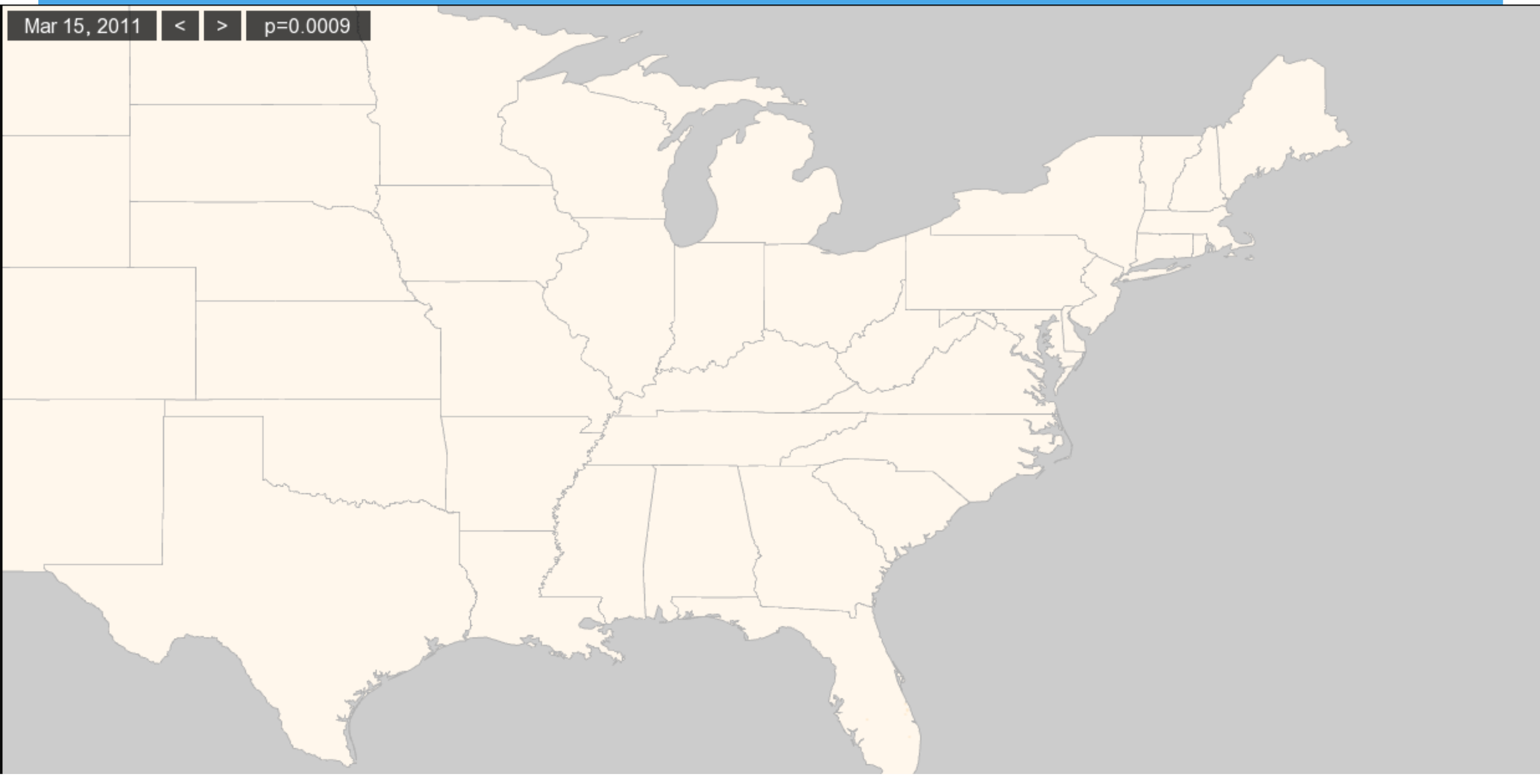
- 15 Canada Goose
- 6 Wood Duck
- 1 Mallard
- 8 Double-crested Cormorant
- 5 Great Blue Heron
- 2 Great Egret
- 1 Osprey
- 1 Bald Eagle
- 1 Sora
- 5 American Coot
- 1 Killdeer
- 1 Ovenbird
- 4 Northern Waterthrush
- 1 Blue-winged Warbler
- 1 Golden-winged Warbler
- 15 Black-and-white Warbler
- 3 Tennessee Warbler
- 15 Nashville Warbler
- 3 Common Yellowthroat
- 10 American Redstart
- 5 Cape May Warbler
- 20 Northern Parula
- 4 Magnolia Warbler
- 1 Bay-breasted Warbler
- 6 Blackburnian Warbler
- 24 Yellow Warbler
- 8 Chestnut-sided Warbler
- 7 Black-throated Blue Warbler
- 22 Palm Warbler

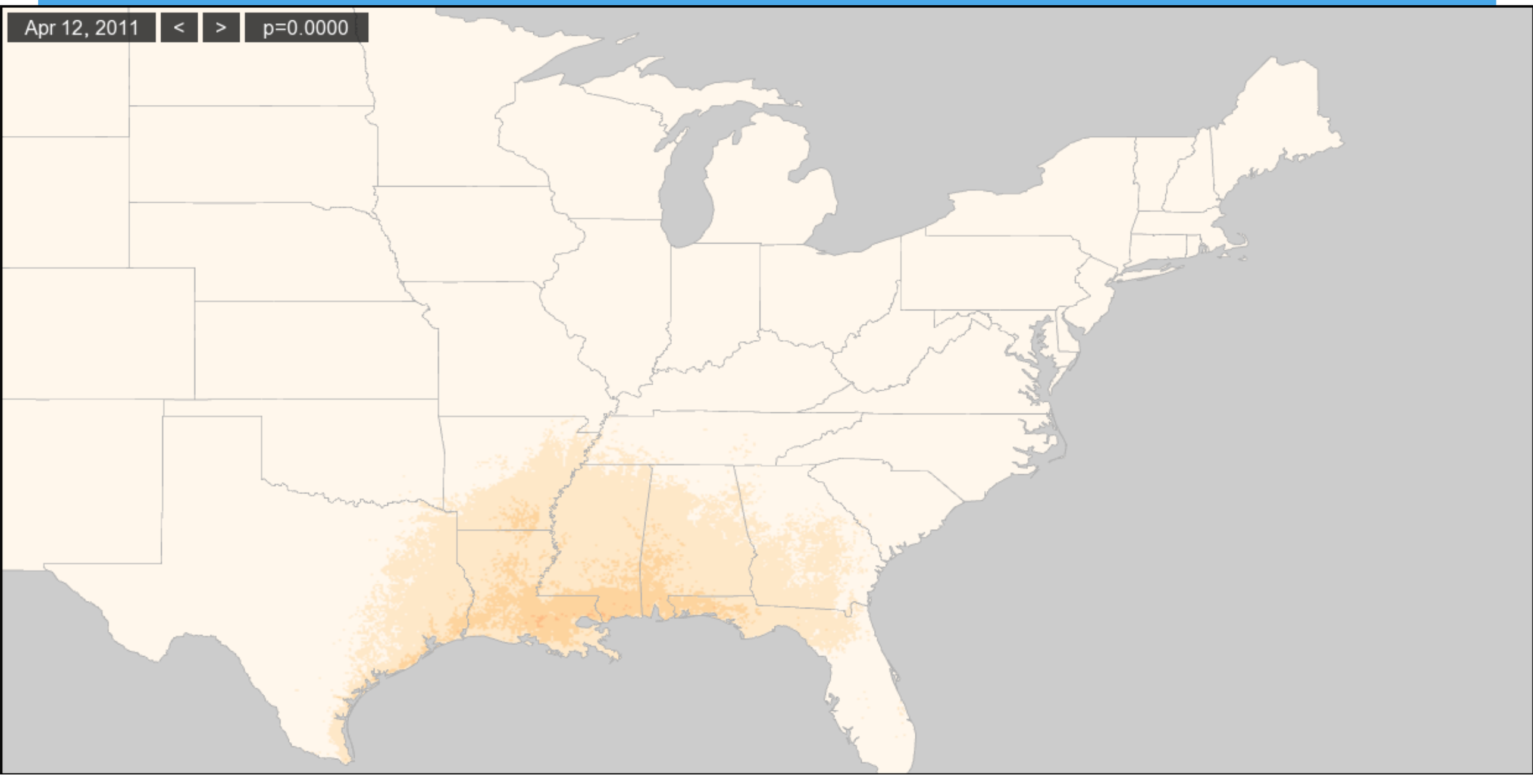
Science

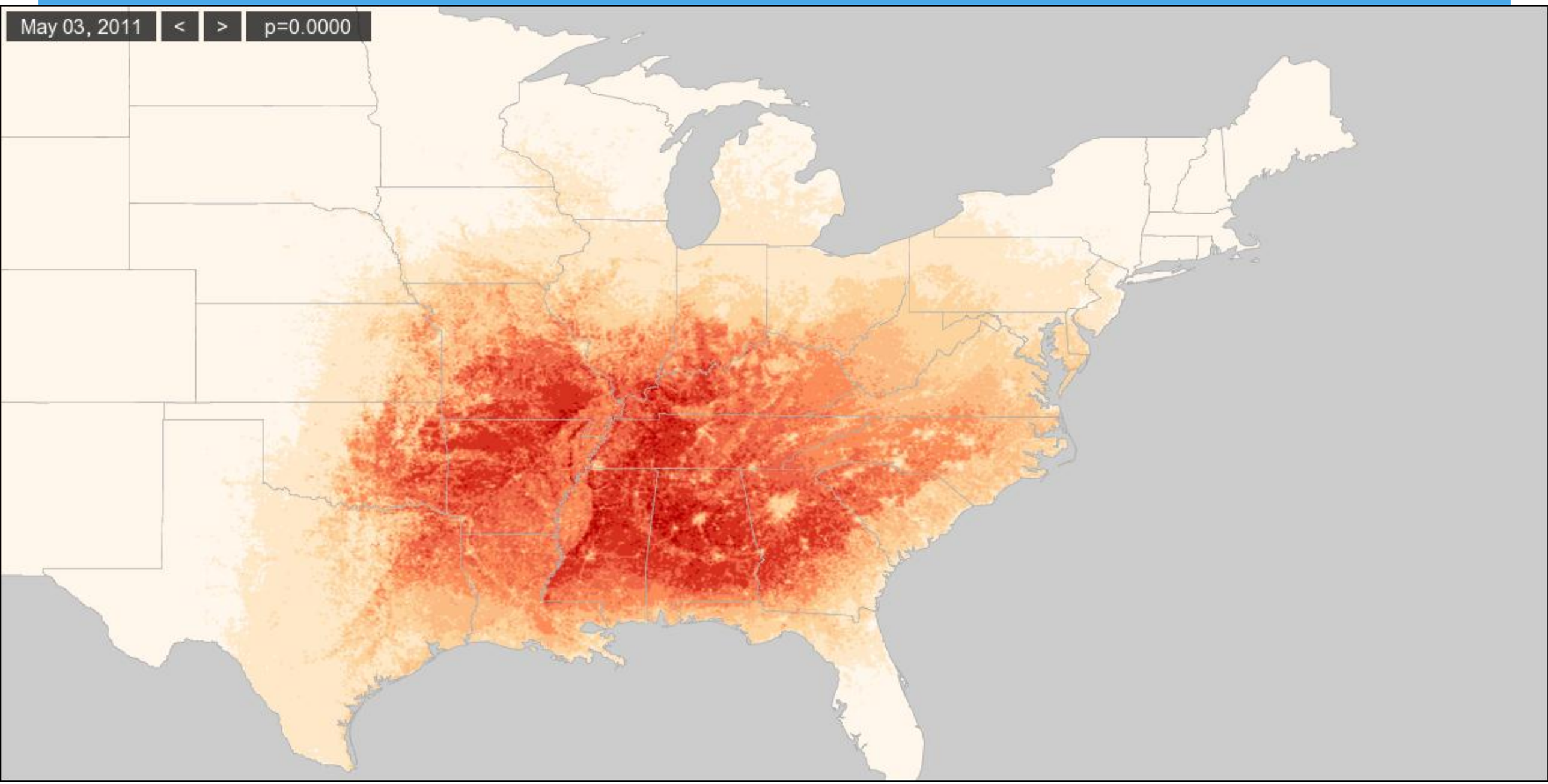
Observational Data Model

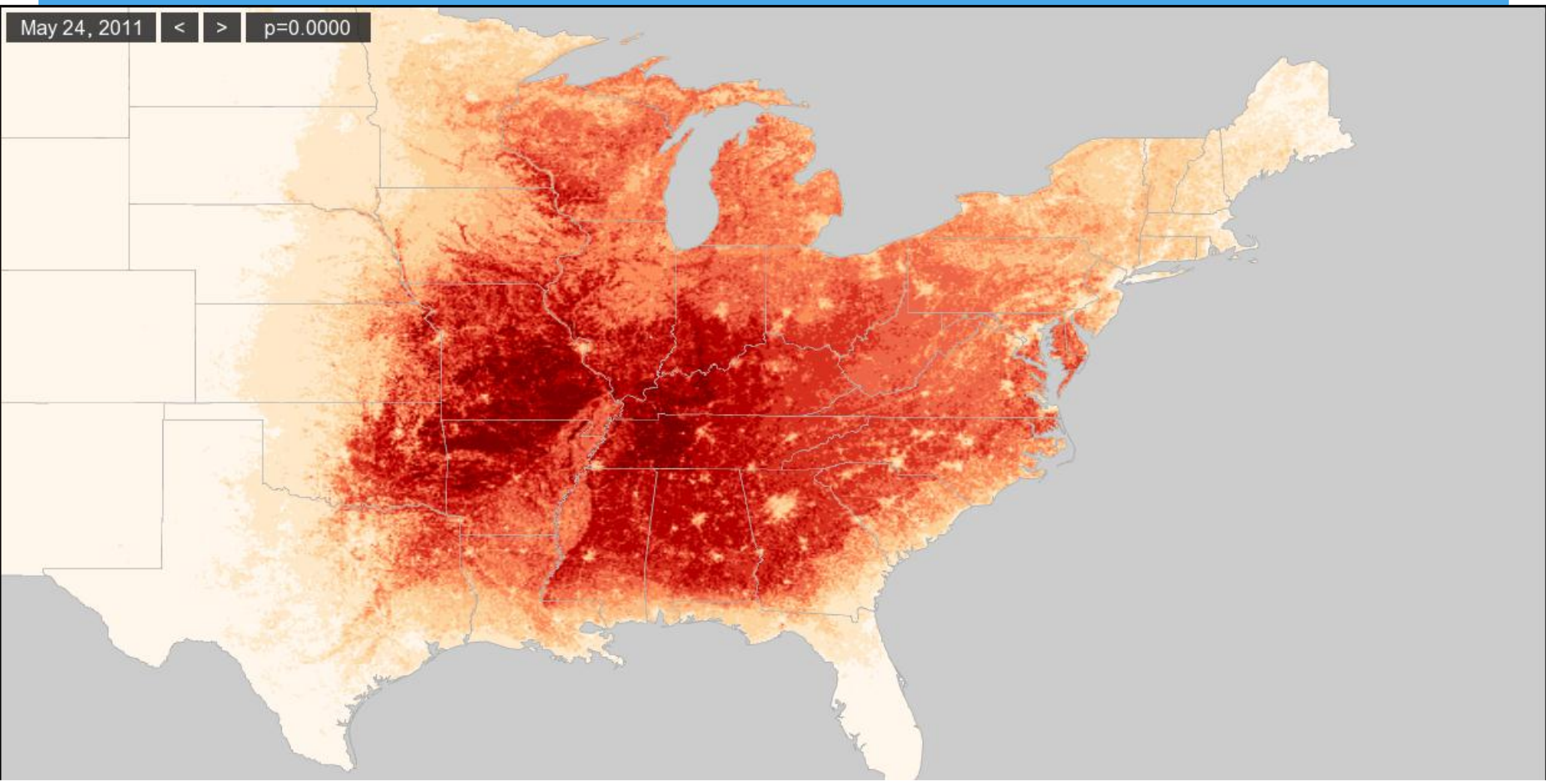


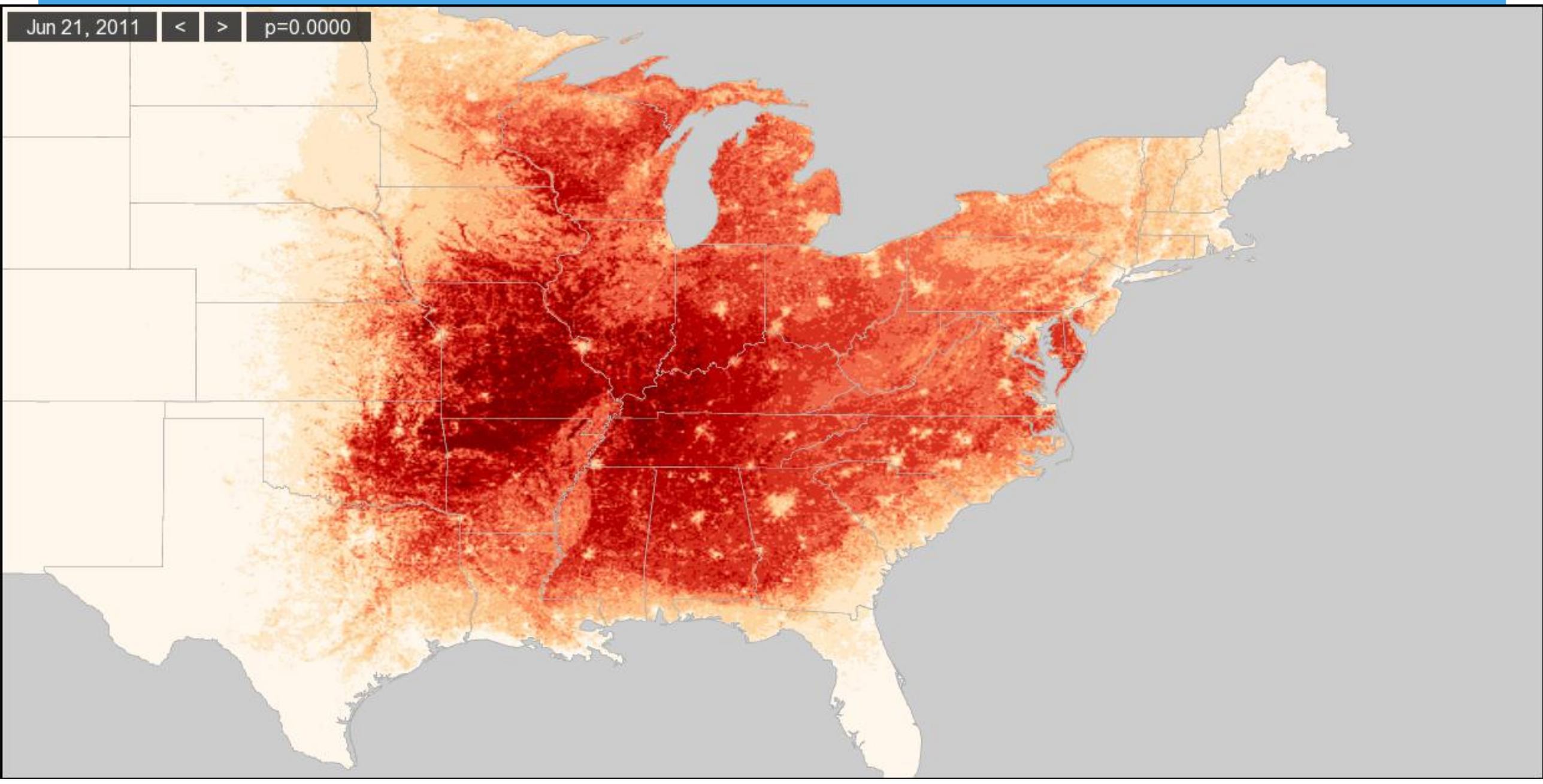
Seasonal Patterns of Occurrence in the Indigo Bunting During 2011

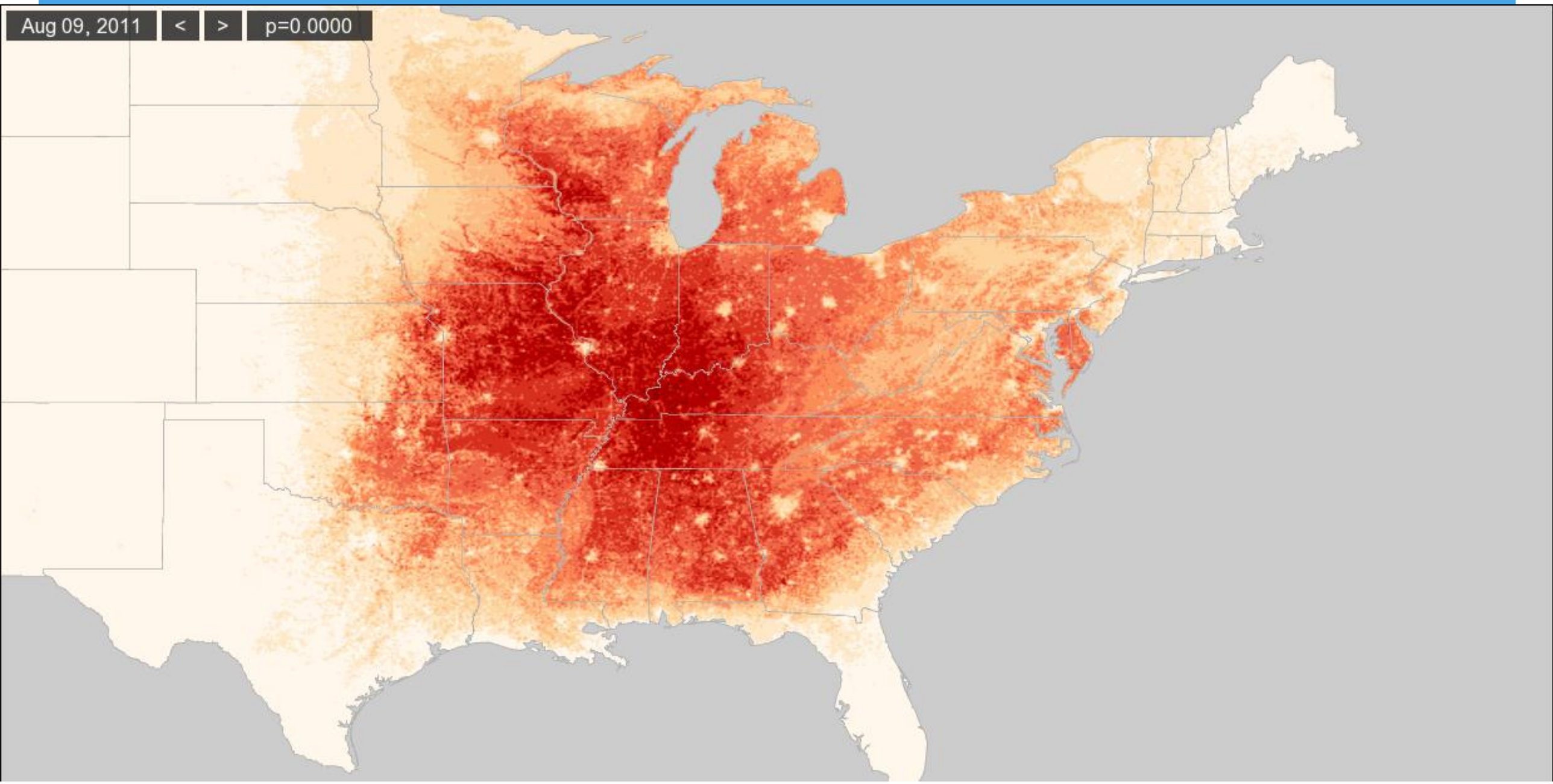


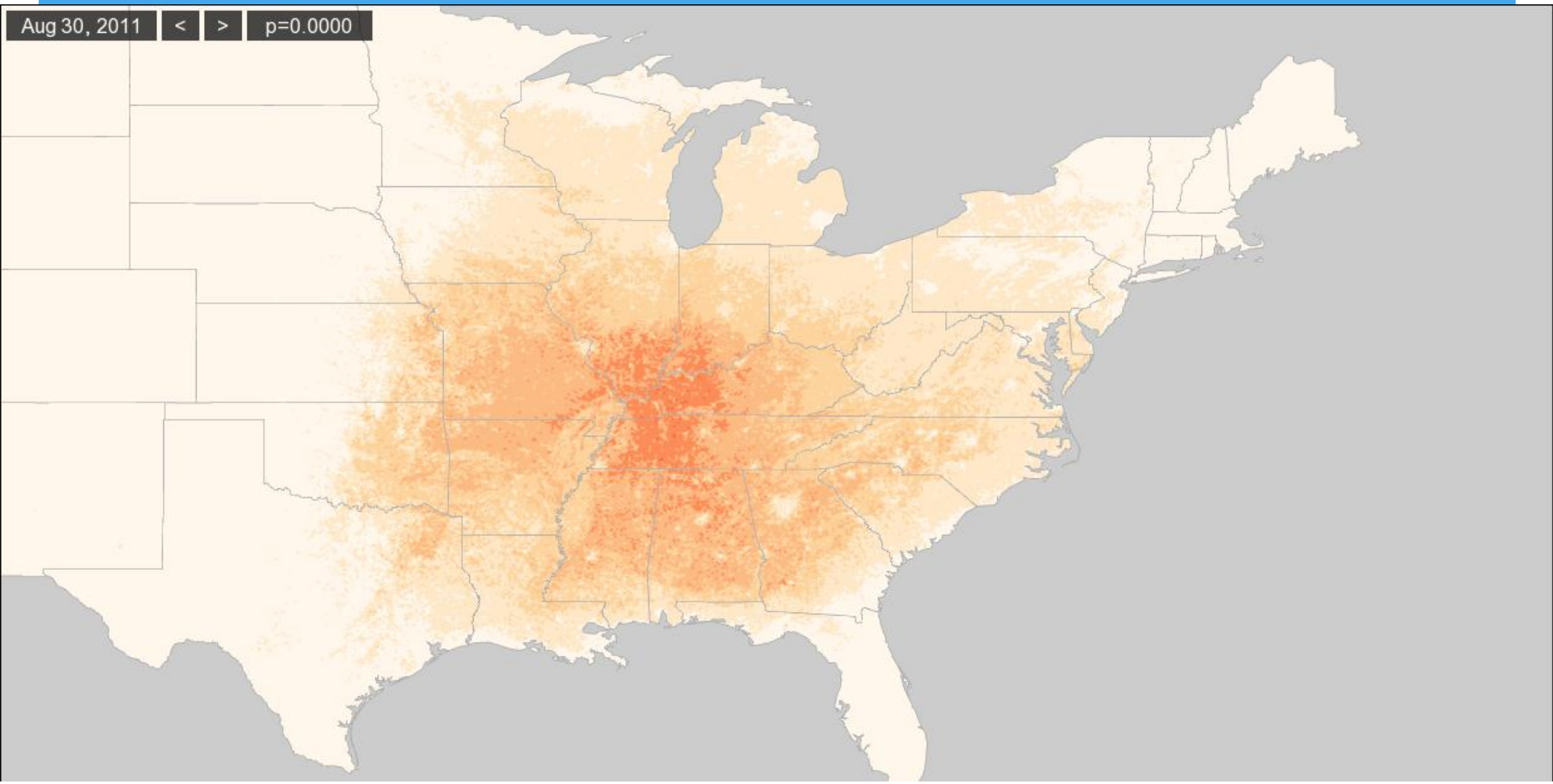

Apr 12, 2011
<
>
p=0.0000

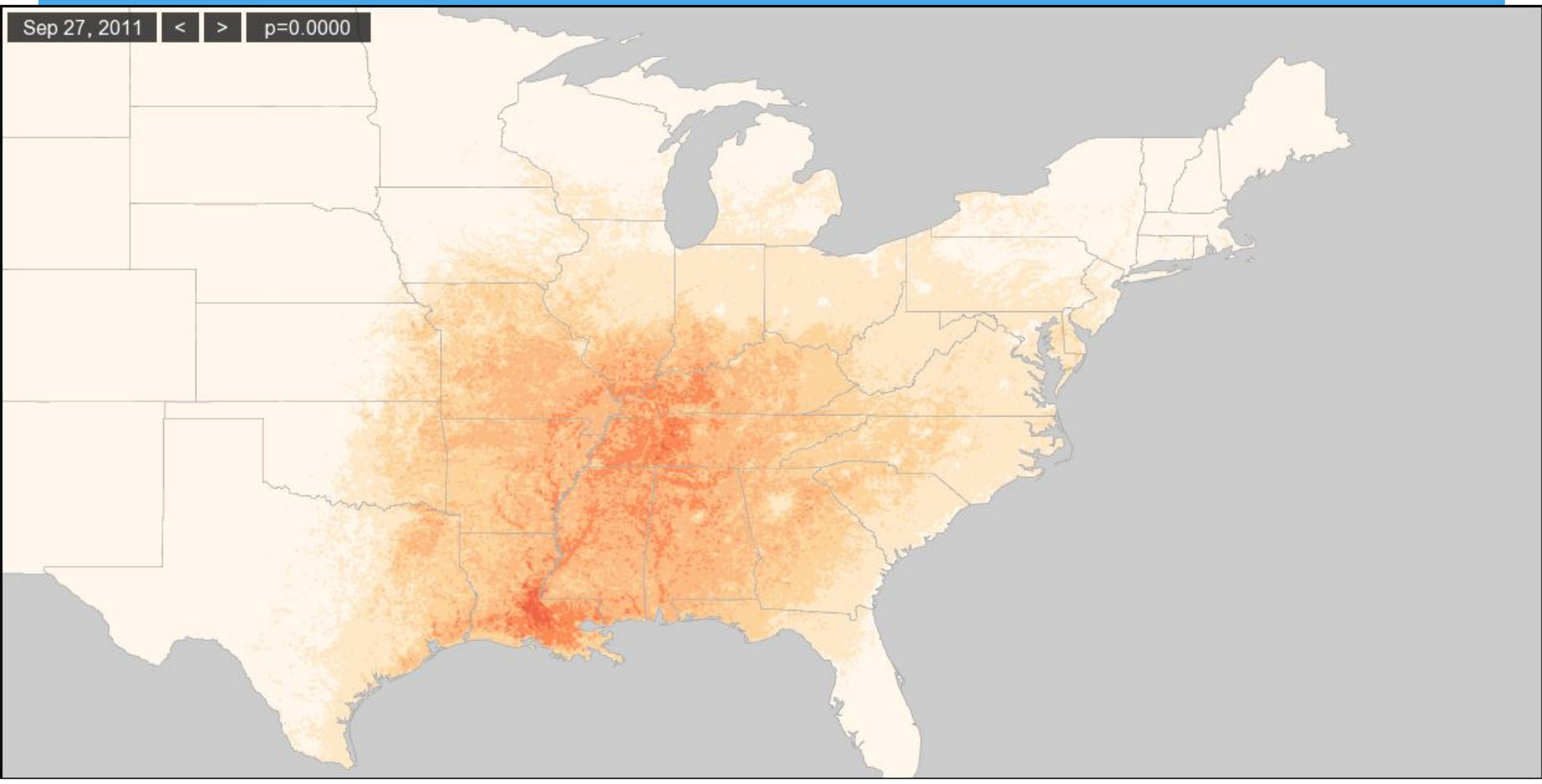

May 03, 2011
<
>
p=0.0000

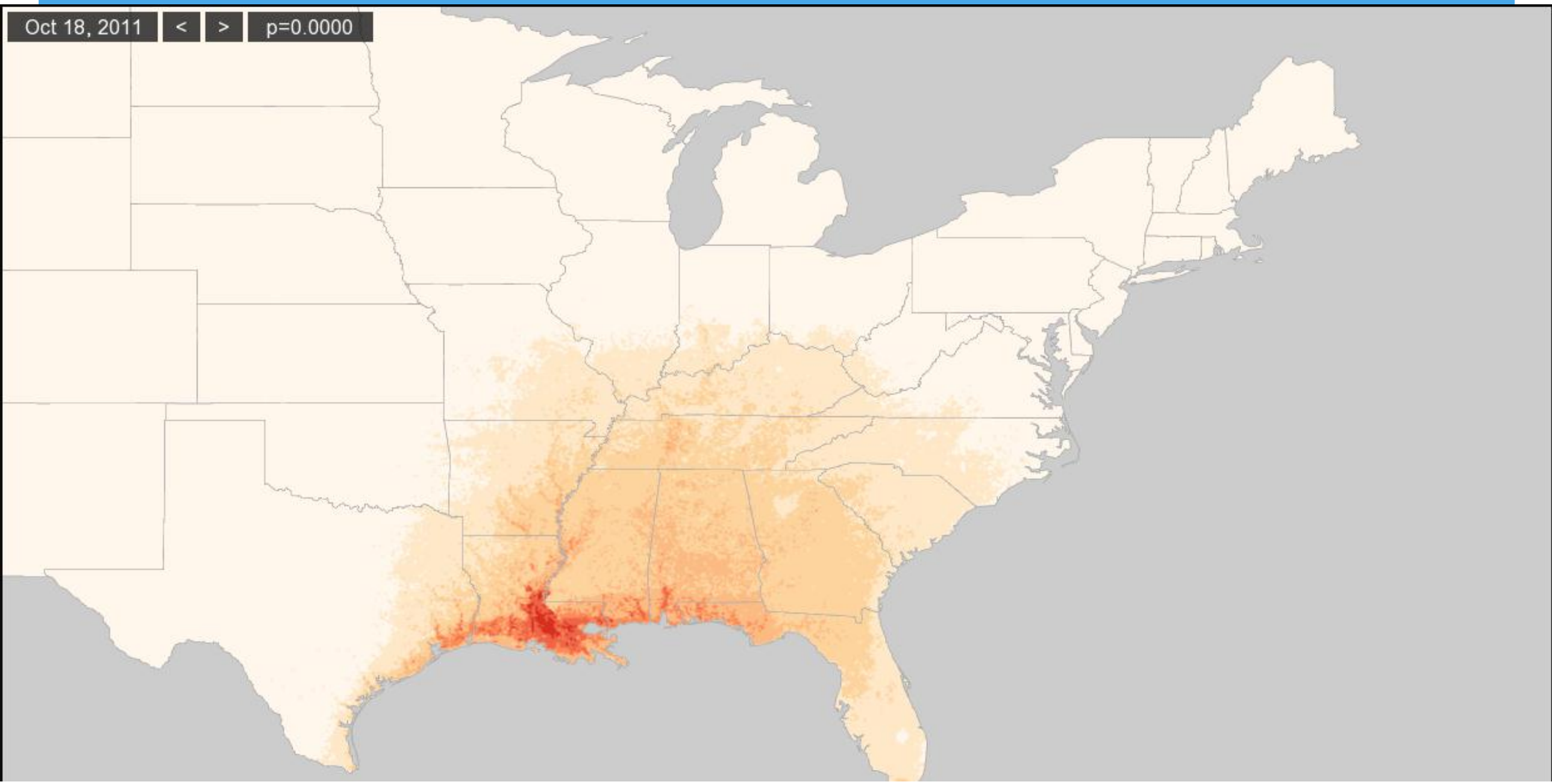

May 24, 2011
<
>
p=0.0000

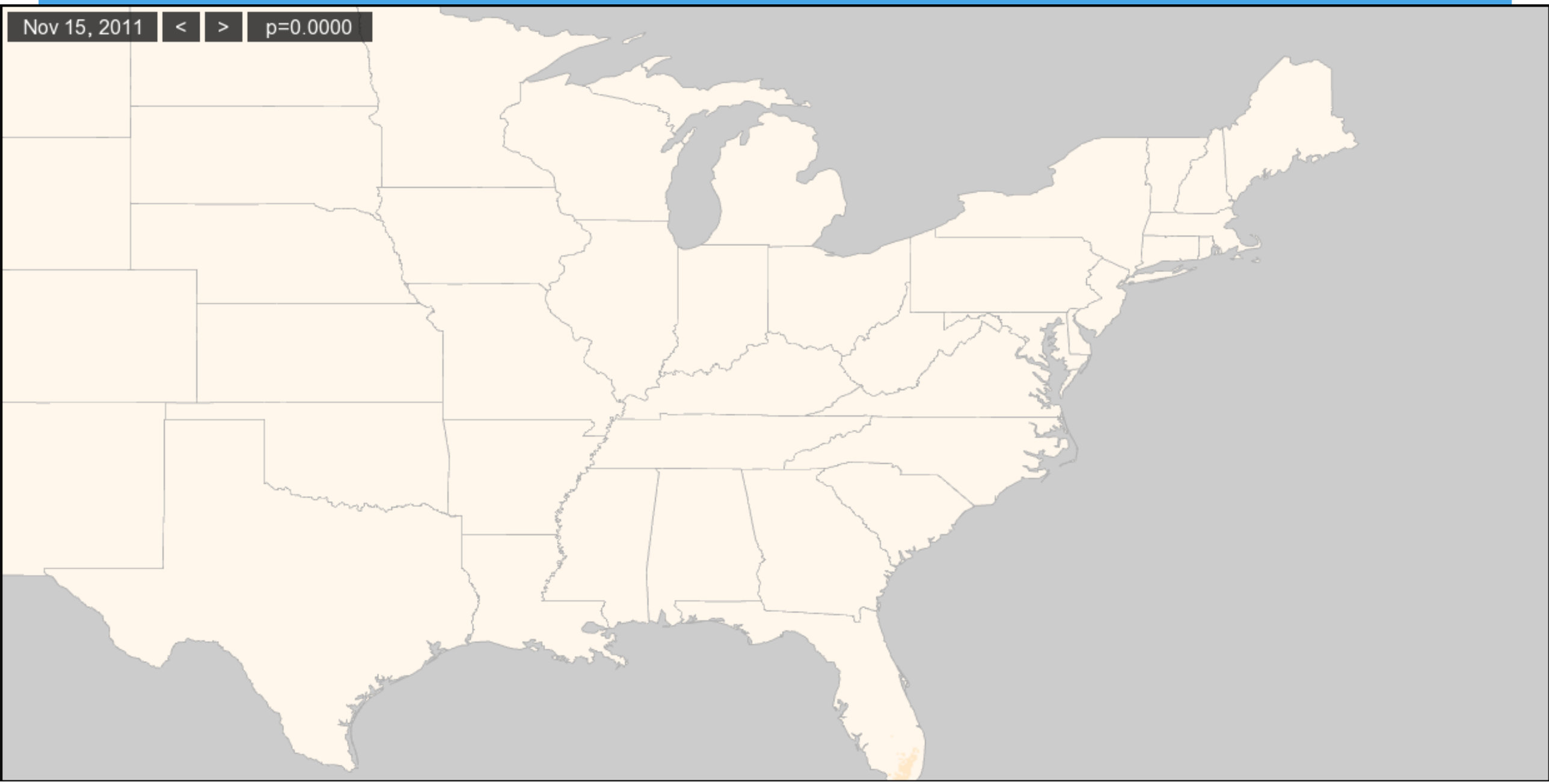

Jun 21, 2011
<
>
p=0.0000


Aug 09, 2011
<
>
p=0.0000


Aug 30, 2011
<
>
p=0.0000


Sep 27, 2011
<
>
p=0.0000


Oct 18, 2011
<
>
p=0.0000


Nov 15, 2011
<
>
p=0.0000

Conservation

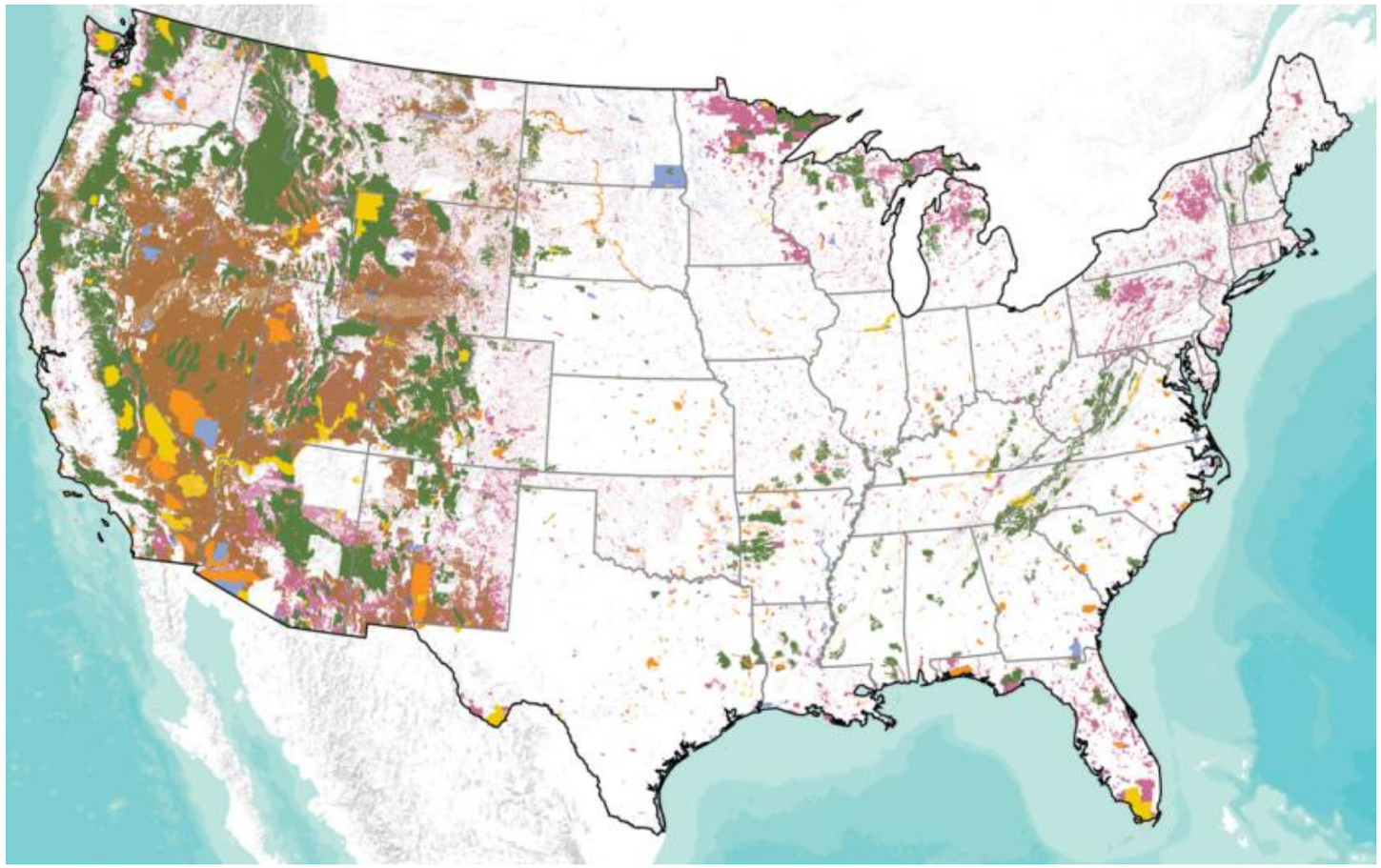
The State of the Birds 2011

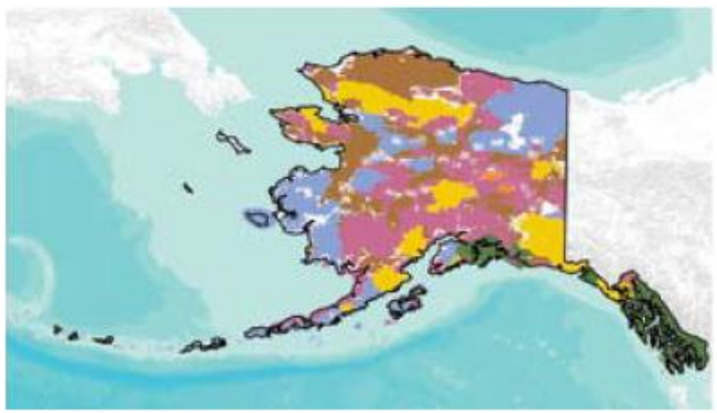
Report on Public Lands and Waters

United States of America

U.S. Public Lands

- Protected Areas Database for the U.S. (PAD-US 1.1; USGS-GAP)
- Includes Federal and State agencies
- Biodiversity protection status
- Bird distributions overlaid on public land ownership map





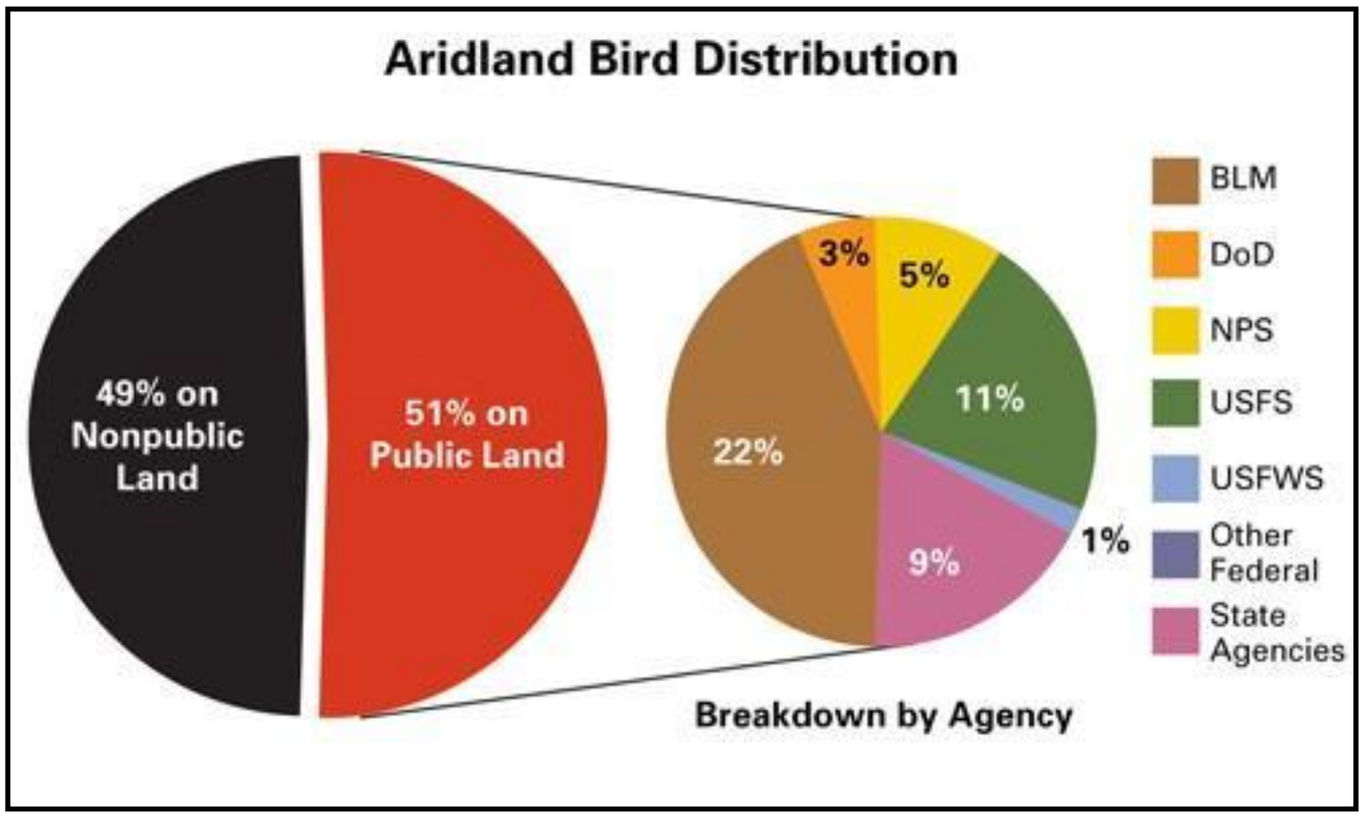
Public Lands

- Bureau of Land Management (BLM)
- Department of Defense (DoD)
- National Park Service (NPS)
- USDA Forest Service (USFS)
- U.S. Fish and Wildlife Service (USFWS)
- State lands

NOAA lands are included on both maps but are too small to detect.

Maps produced by National Gap Analysis Program

ARIDLANDS and Bureau of Land Management

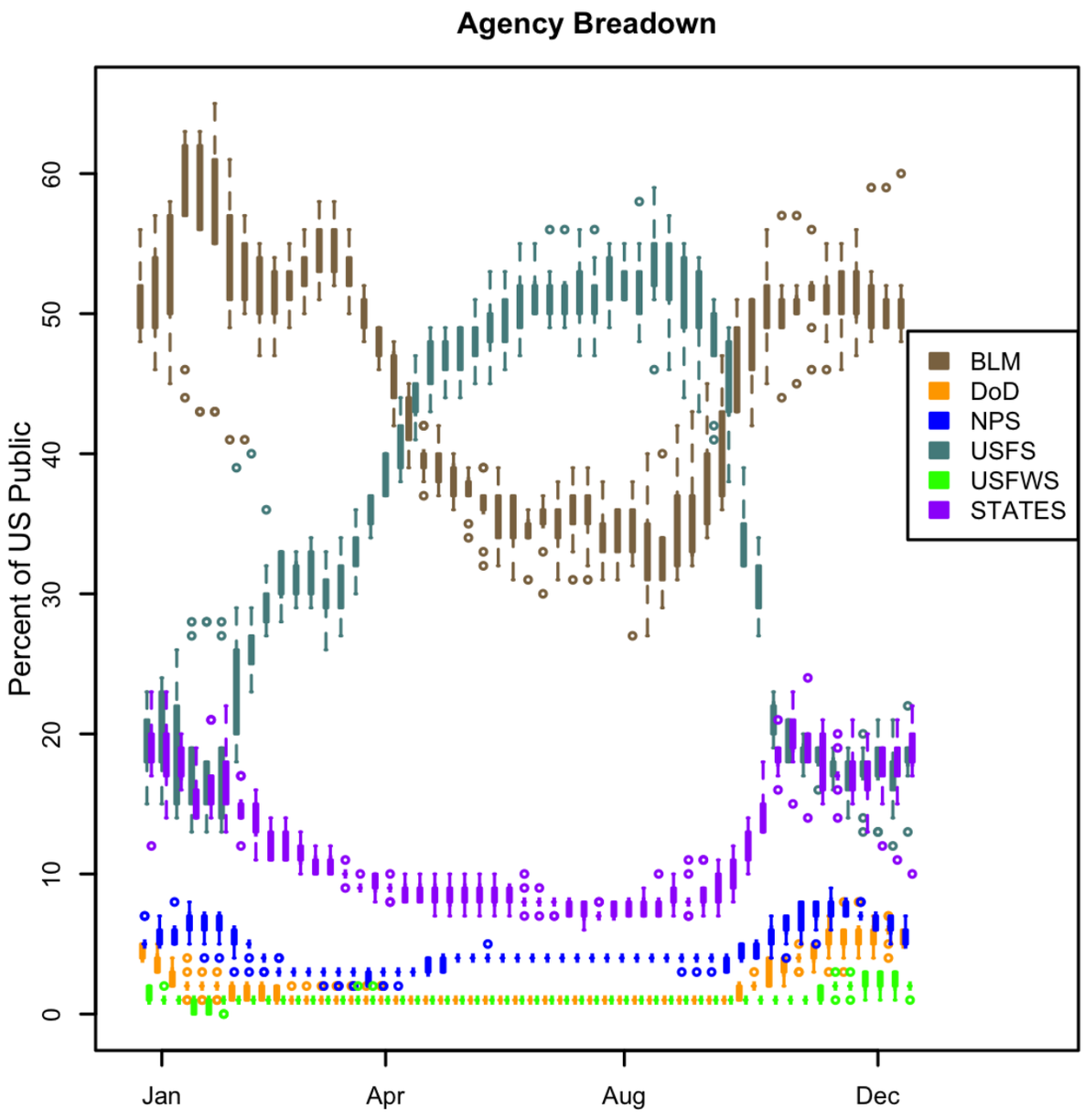


- Public lands support more than half the distribution of aridland birds (36 obligate species)
- Declining aridland birds affected by grazing, invasive species, energy development

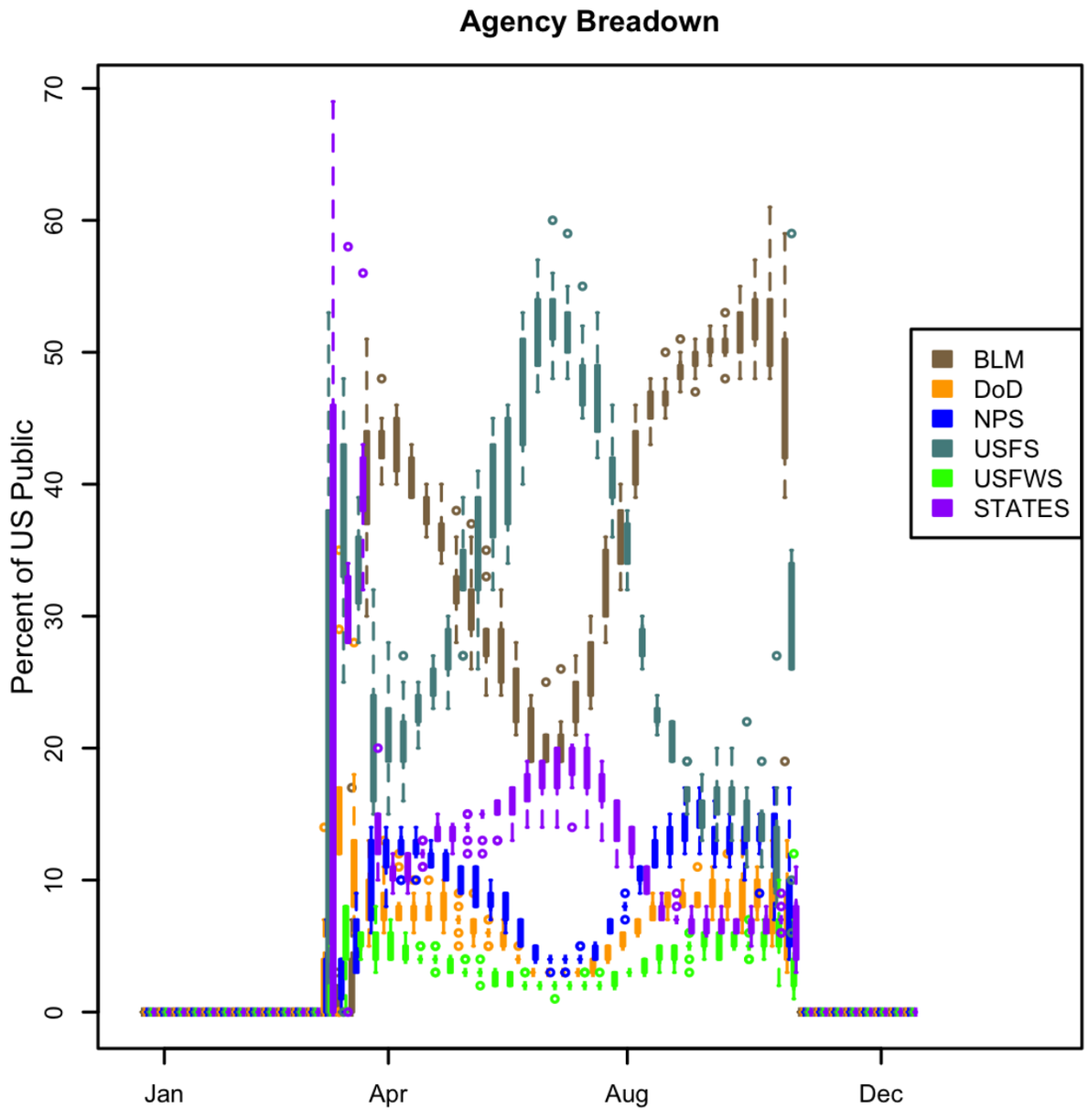
- 79% of Gunnison Sage-Grouse distribution on public lands
- BLM manages 245 million acres, including more aridland habitat than any other agency

Stewardship Responsibilities for Birds

Mountain Bluebird



Willow Flycatcher



Mapping Migratory Bird Stopover Habitat to Inform Wind Energy Siting

Ron Rohrbaugh, Frank La Sorte, Daniel Fink,
Andrew Farnsworth, and Steve Kelling

New York Natural Heritage Program

The Cornell Lab of Ornithology

Observer/Data Quality

About eBird

eBird Data Quality

Did you know that every record submitted to eBird goes through the eBird data verification process? Using a combination of automated data filters and a network of local experts, eBird tackles the issue of data quality in Citizen-Science. In order for us to maintain the integrity of the database, and for it to be used fully by the science and conservation community, we as observers must fully understand and strive to reach the highest level of data quality. Therefore, we've developed procedures to facilitate communication between eBird observers and scientists, including some new and improved review tools for our editors. Through our combined effort to maintain high data quality, eBird will take its place among the most valuable large-scale data sets on bird distribution and abundance in the world. Read more about our data verification process....

Taiga Flycatcher, Putah Creek, California, 25 October 2006. Photo by John Sterling.

What do we do when you report birds as rare or far out of range as this Siberian Taiga Flycatcher at Putah Creek, California, found by eBirder John Sterling, or the Ivory Gull found near New York City? Even more complex is the issue of how to deal with records of early/late migrants, or out of season records of lingering birds at local scales. Issues surrounding how to verify data like these certainly come into play when maintaining a database of records that is meant to become part of the scientific record. As with any large-scale citizen-science project open to the public, there is the possibility that erroneous data will be submitted. At eBird we consider data quality to be paramount, and we're taking every step possible to ensure that our data are the best they can be. Using advanced data vetting technology, we've developed a combination of automated filters and a network of regional editors that work together to verify eBird data. Each eBird submission, regardless of observer or location, is checked for data quality in exactly the same way.

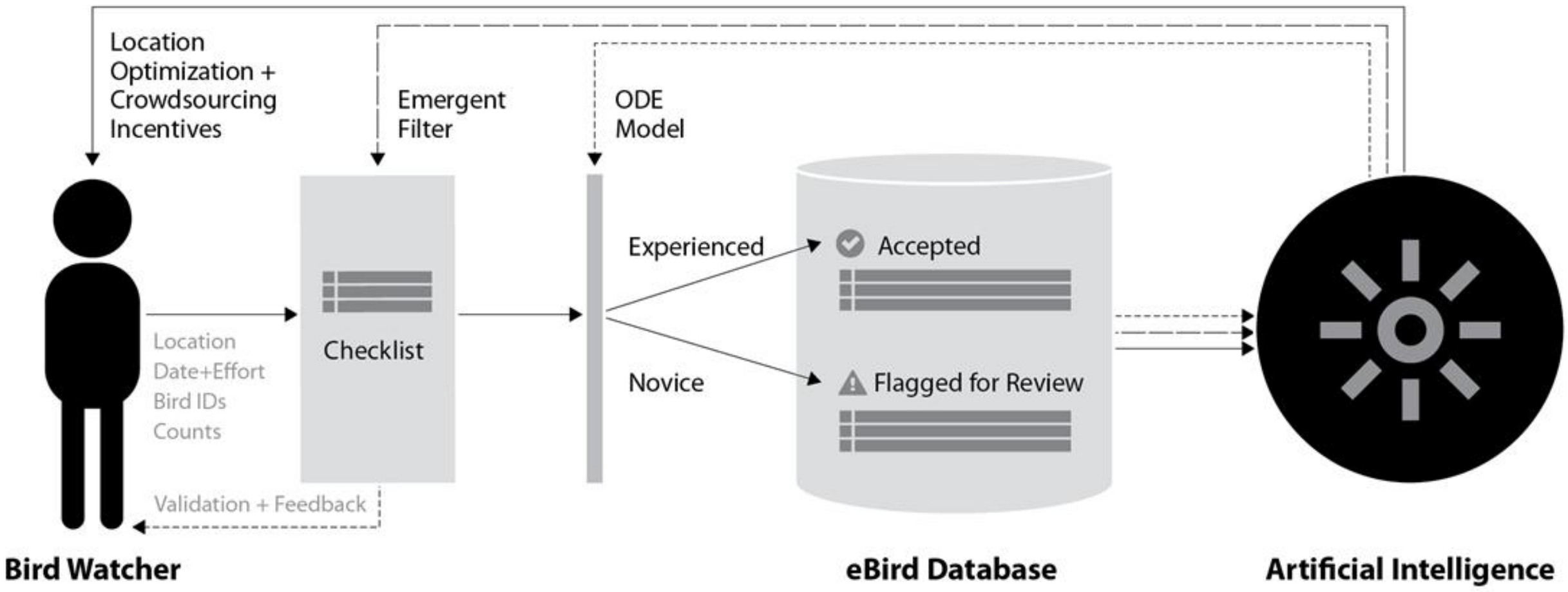
Why worry about data quality?

A database is only as good as its weakest record. If even a few records can be deemed questionable, then the entire data set can be labeled as such. With that in mind, we should all strive to keep the eBird data as clean as possible. You can do your part by being conservative in the field and meticulous with your data entry, and we can do ours by building better connections between the eBird community and scientists.

Enhance the effectiveness of citizen science

- build accurate data input filters to limit erroneous data submission
- identify variation in observer ability
- address the spatial bias of where observers make their observations

Human/Computer Learning Network
Active Learning



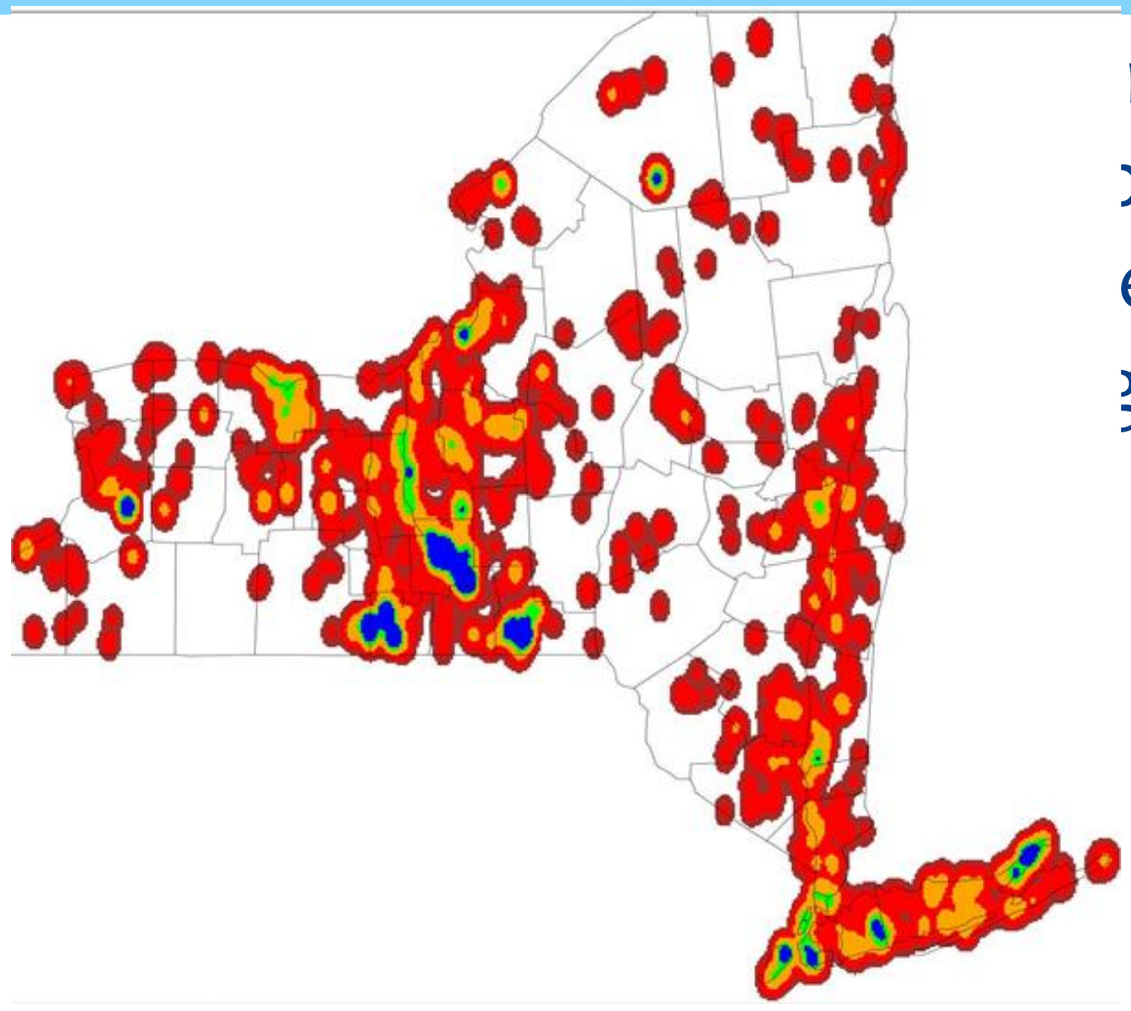
Emergent filters to limit erroneous observations

- Replace unscalable system of volunteer regional experts
- Apply frequency of occurrence filters, generated from eBird data, to delineate when a species can be reported in a region
- Automatically flag unusual observations

Addressing observer variability by modeling observer ability

- Can't establish "ground truth" through multiple observers
- Occupancy-Detection-Experience Model (ODE)
  - probability of the occupancy of the site by a particular species given a set of environmental covariates describing that site
  - Factor in "ease of detection" covariates
  - Quantify individual observer expertise variation based on covariates such as number of checklists submitted, number of records flagged by experts, number of species reported, etc.
  - Improved expertise by suggesting species they "should" see

Improving spatial coverage in citizen science



mprove the predictive
performance of the machine
earning algorithms by
guiding the sampling process

Measuring HCLN impact on citizen science participants

- Does the system and expertise ranking provide incentives for greater participation?
- Does the system encourage greater geographic distribution of observations?
- Is the system “educating” the observers?
- What is the nature of the detection curve for various types of observers and how does it change over time?

Questions?