HOLE-Y PLANT DATABASES! UNDERSTANDING AND PREVENTING BIASES IN BOTANICAL BIG DATA

Katelin D. Pearson


wikimedia.org

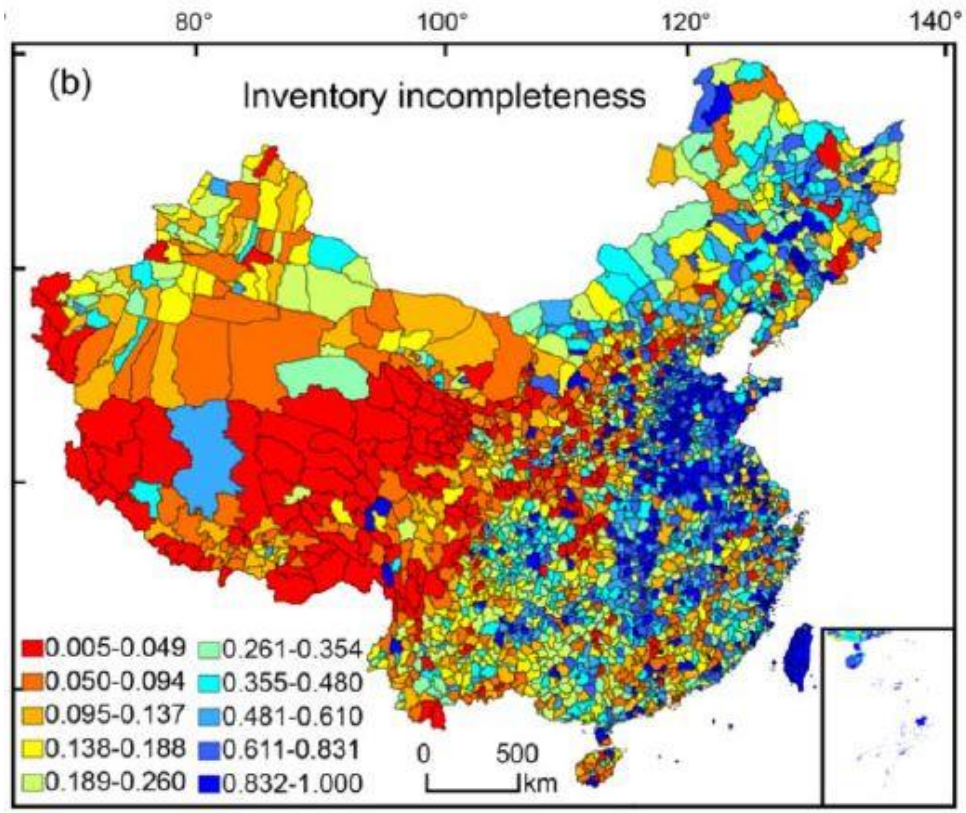
Questions

1. What types of biases exist in herbarium specimen databases?
2. What are the potential effects of biases on analyses using herbarium specimen data?
3. How can we mitigate the effects of biases?

"Scope"

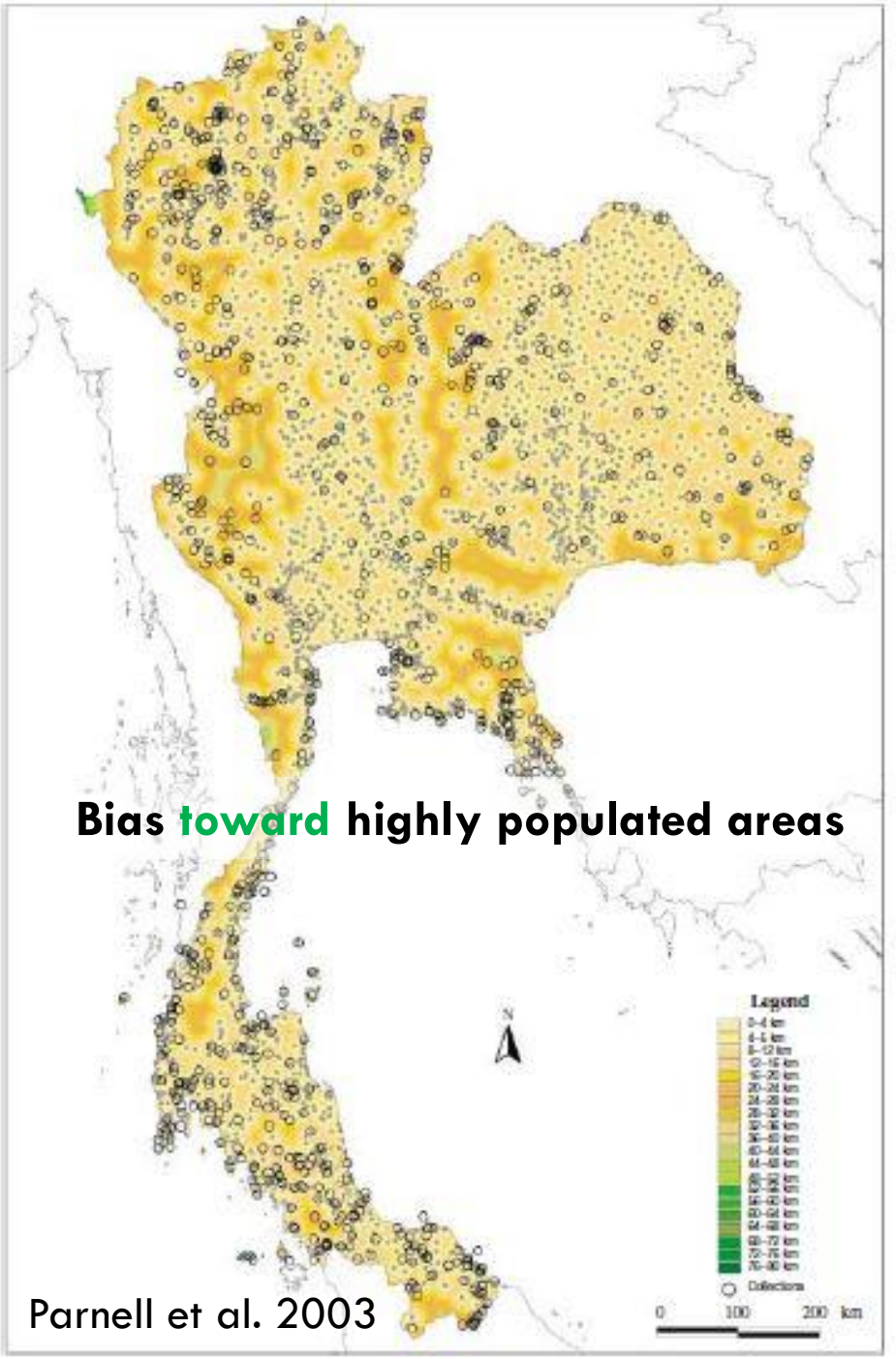
	Regional	Systematic

"Class"



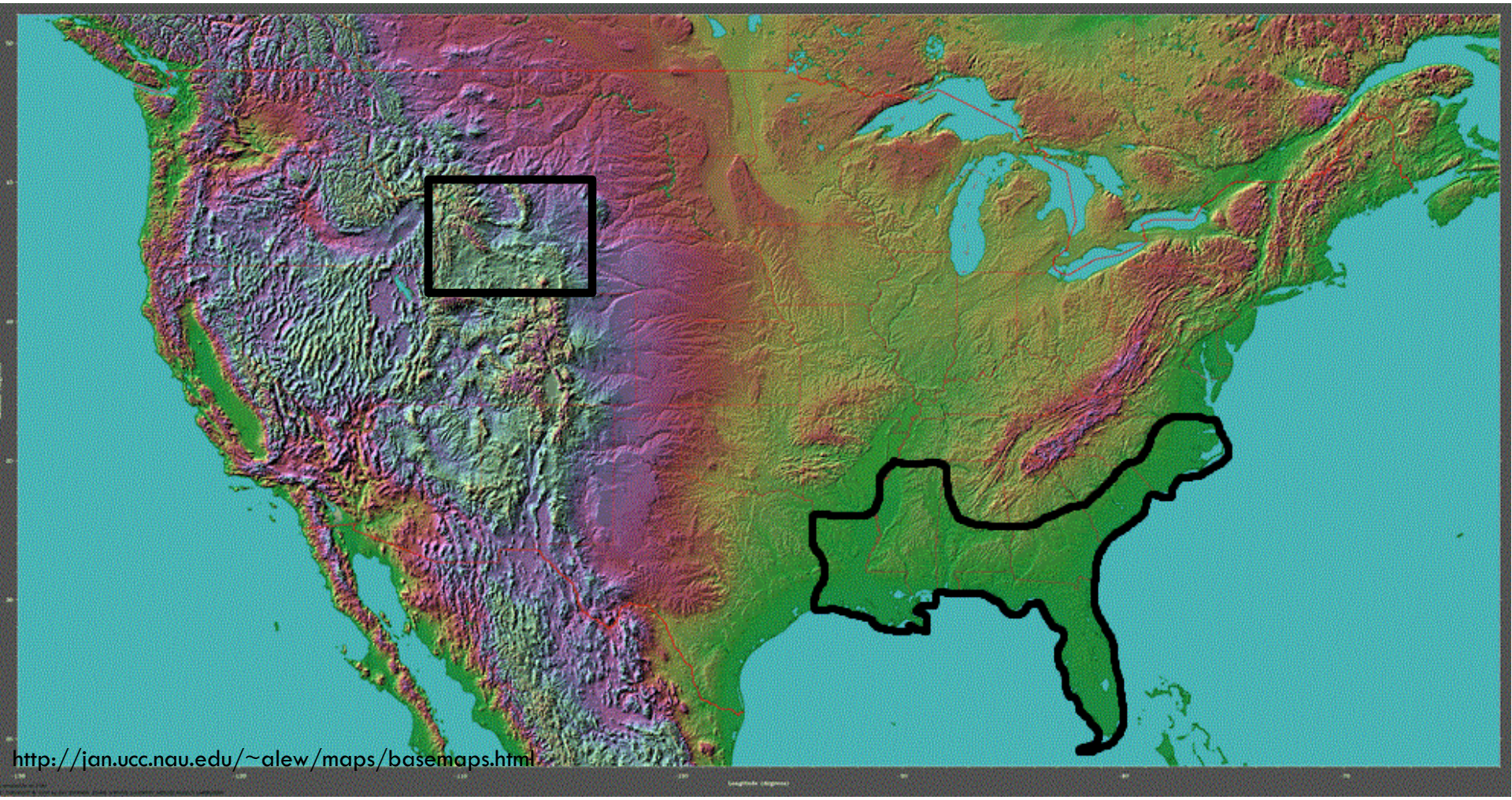
Bias against highly populated areas

Yang et al. 2014

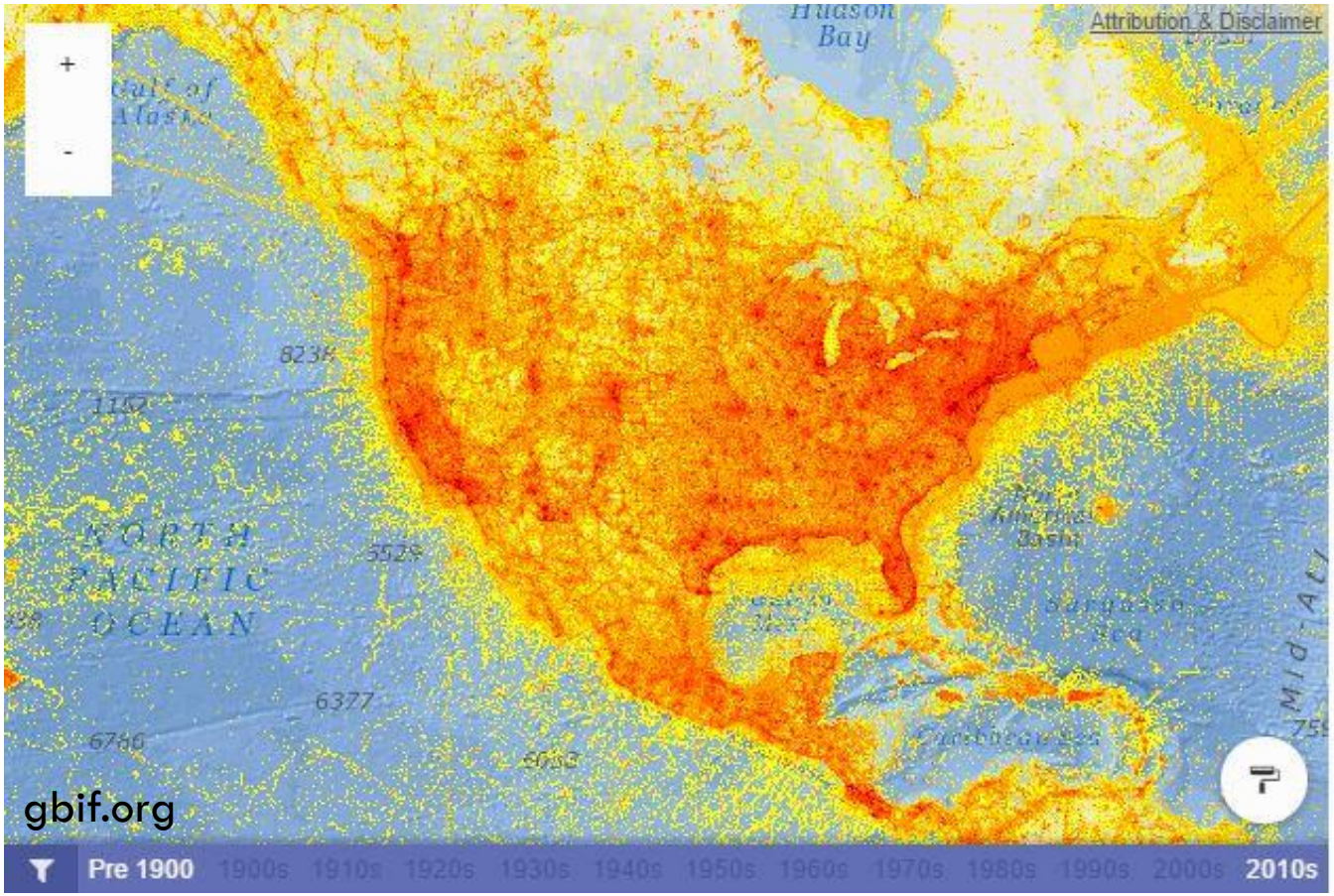


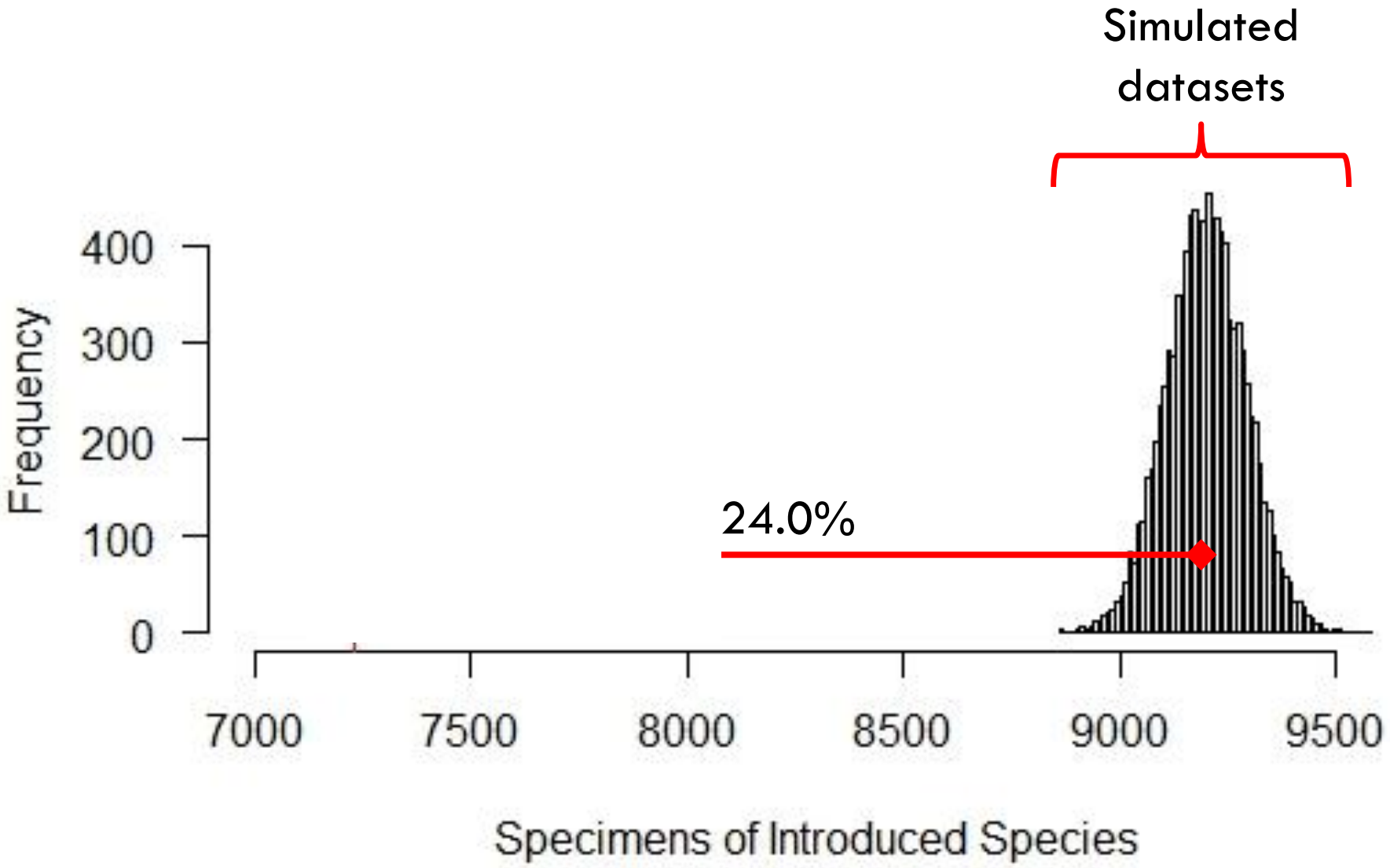
Bias toward highly populated areas

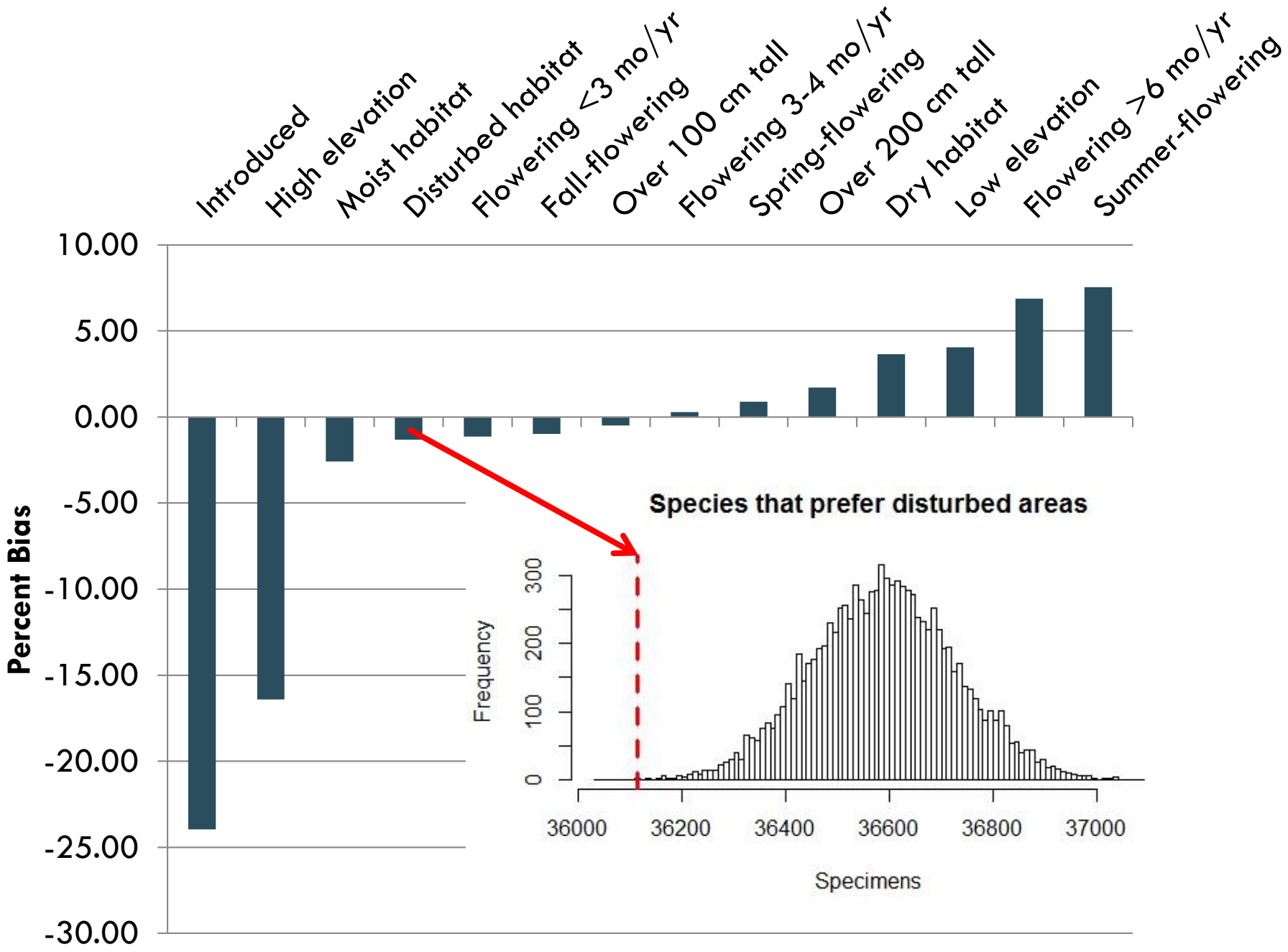
Parnell et al. 2003

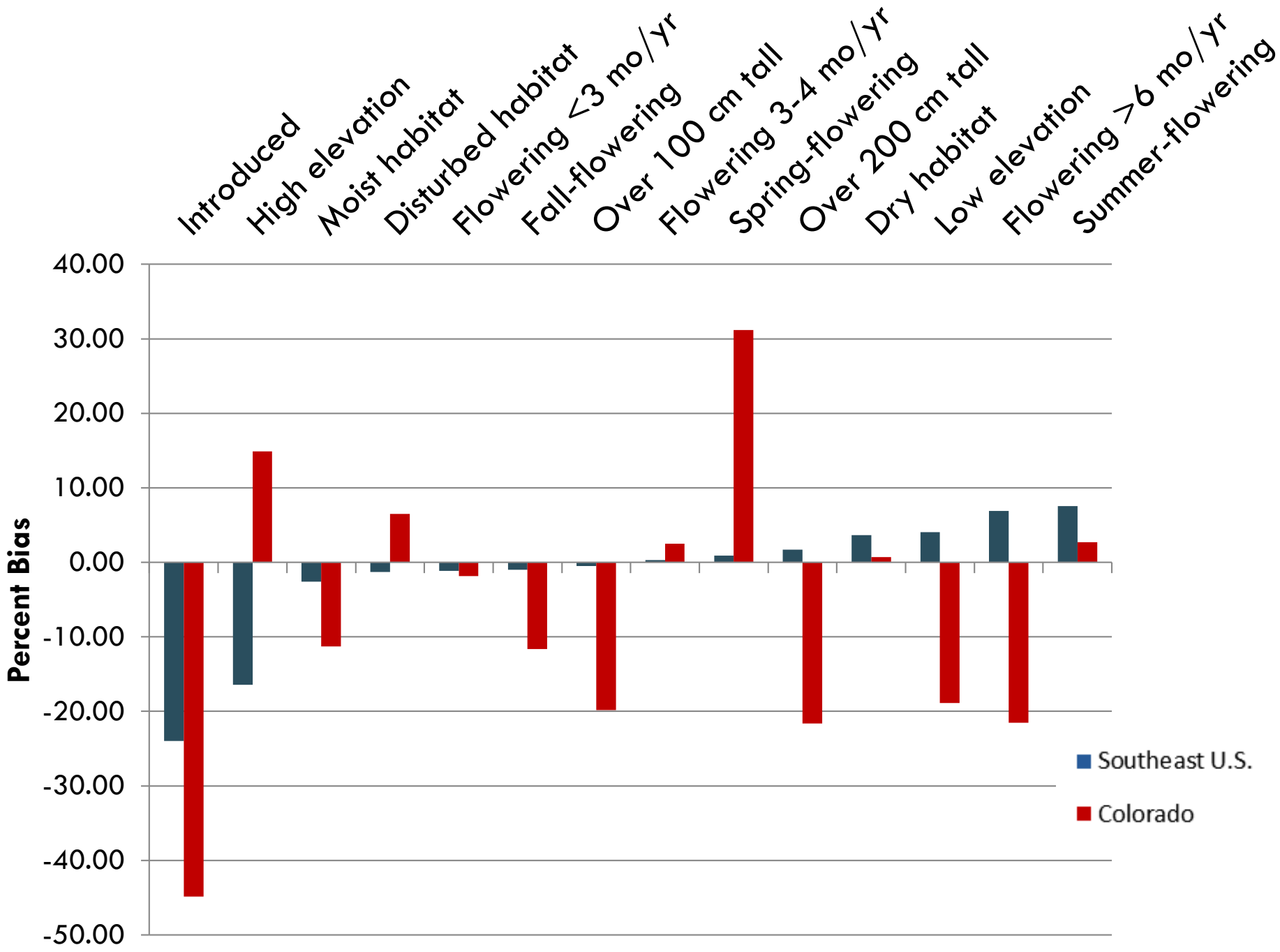

http://jan.ucc.nau.edu/~alew/maps/basemaps.html

- Simulated 10,000 datasets of “random collections” accounting for geographic collecting bias




Simulated datasets
24.0%
Frequency
400
300
200
100
0
7000
7500
8000
8500
9000
9500
Specimens of Introduced Species


Introduced
High elevation
Moist habitat
Disturbed habitat
Flowering <3 mo/yr
Fall-flowering
Over 100 cm tall
Flowering 3-4 mo/yr
Spring-flowering
Over 200 cm tall
Dry habitat
Low elevation
Flowering >6 mo/yr
Summer-flowering
10.00
5.00
0.00
-5.00
-10.00
-15.00
-20.00
-25.00
-30.00
Percent Bias
Species that prefer disturbed areas
Frequency
0
100
200
300
36000
36200
36400
36600
36800
37000
Specimens


Introduced
High elevation
Moist habitat
Disturbed habitat
Flowering <3 mo/yr
Fall-flowering
Over 100 cm tall
Flowering 3-4 mo/yr
Spring-flowering
Over 200 cm tall
Dry habitat
Low elevation
Flowering >6 mo/yr
Summer-flowering
40.00
30.00
20.00
10.00
0.00
-10.00
-20.00
-30.00
-40.00
-50.00
Percent Bias
Southeast U.S.
Colorado

How could biases affect analyses?

- It depends!
  - On the analysis
  - On the biases in the dataset

How can we mitigate biases?

- Increase sample size
  - Combine data sources
- Rarefy data
- Standardize according to a well-collected species or group of species
- Use bias-insensitive tests
- Use modeling/statistical approaches that correct for biases

Iverson and Prasad 1998
Abbott et al. 1999
Bergamini et al. 2009
Schmidt et al. 2005
Jacome et al. 2007
Davis et al. 2015
Velland et al. 2015
Solow and Roberts 2006
Burley et al. 2012
Magwe-Tindo et al. 2016

Hedenas et al. 2001
Delisle et al. 2003
Hofmann et al. 2007
Case et al. 2007
Aikio et al. 2010
Follak et al. 2015
Murray-Smith et al. 2009
Raes et al. 2009

Wolf et al. 2011
MacPherson et al. 2014

Recommendations

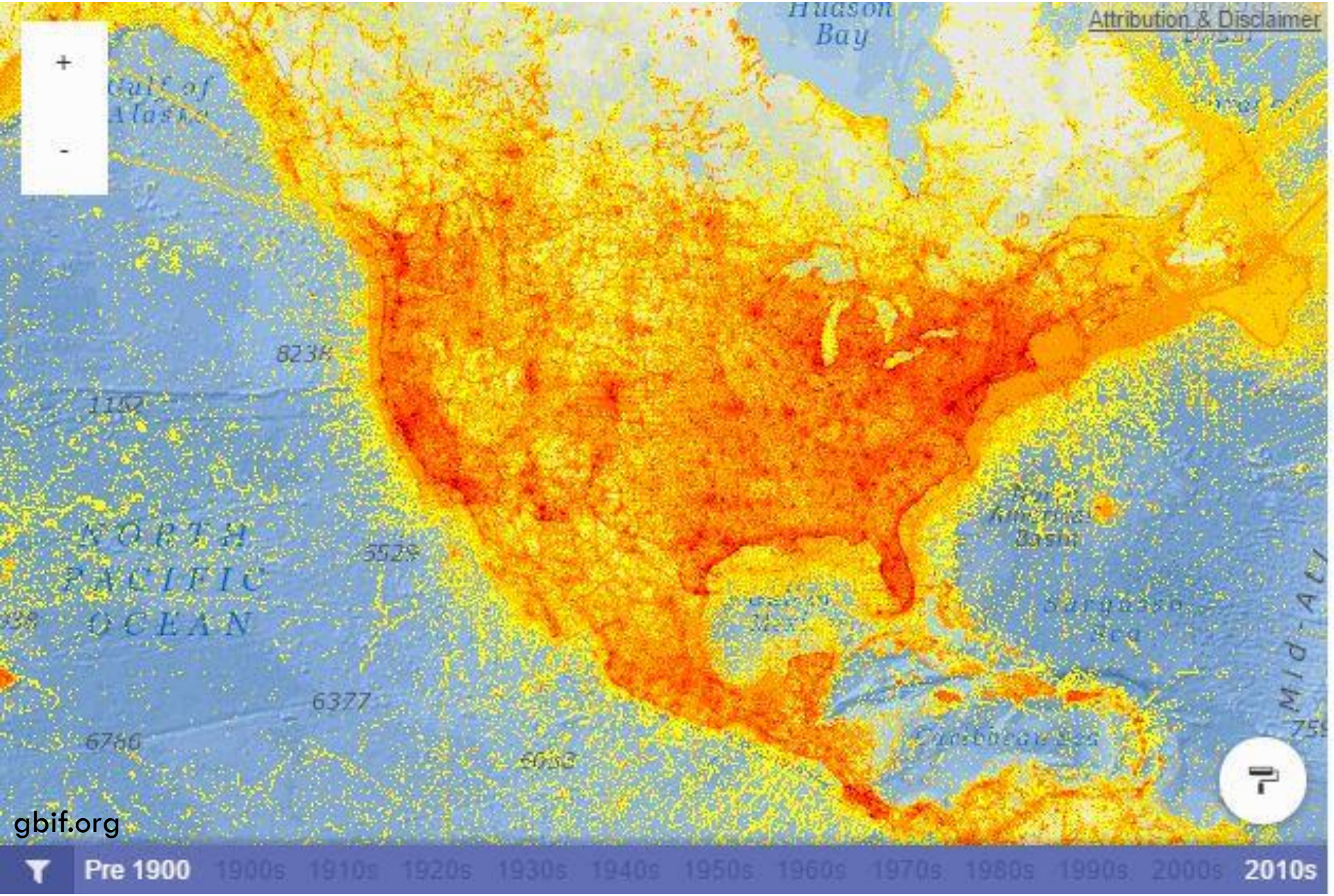
- For:
  - Data users
  - Data managers/providers
  - Data creators (i.e. collectors)

Recommendations

- Data users:
  1. Understand the history of your dataset

Recommendations

- Data users:
  1. Understand the history of your dataset
  2. Always account for geographic bias


Hudson Bay
Attribution & Disclaimer
+
-
NORTH PACIFIC OCEAN
6377
gbif.org
Pre 1900
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s

Recommendations

- Data users:
  1. Understand the history of your dataset
  2. Always account for geographic bias
  3. Consider integrating multiple sources of data

Recommendations

- Data users:
  1. Understand the history of your dataset
  2. Always account for geographic bias
  3. Consider integrating multiple sources of data
- Data managers/providers:
  1. Make digitization metadata available: describe your methods

Recommendations

- Data users:
  1. Understand the history of your dataset
  2. Always account for geographic bias
  3. Consider integrating multiple sources of data
- Data managers/providers:
  1. Make digitization metadata available: describe your methods
- Data creators (i.e. collectors):
  1. Collect in under-collected regions

Recommendations

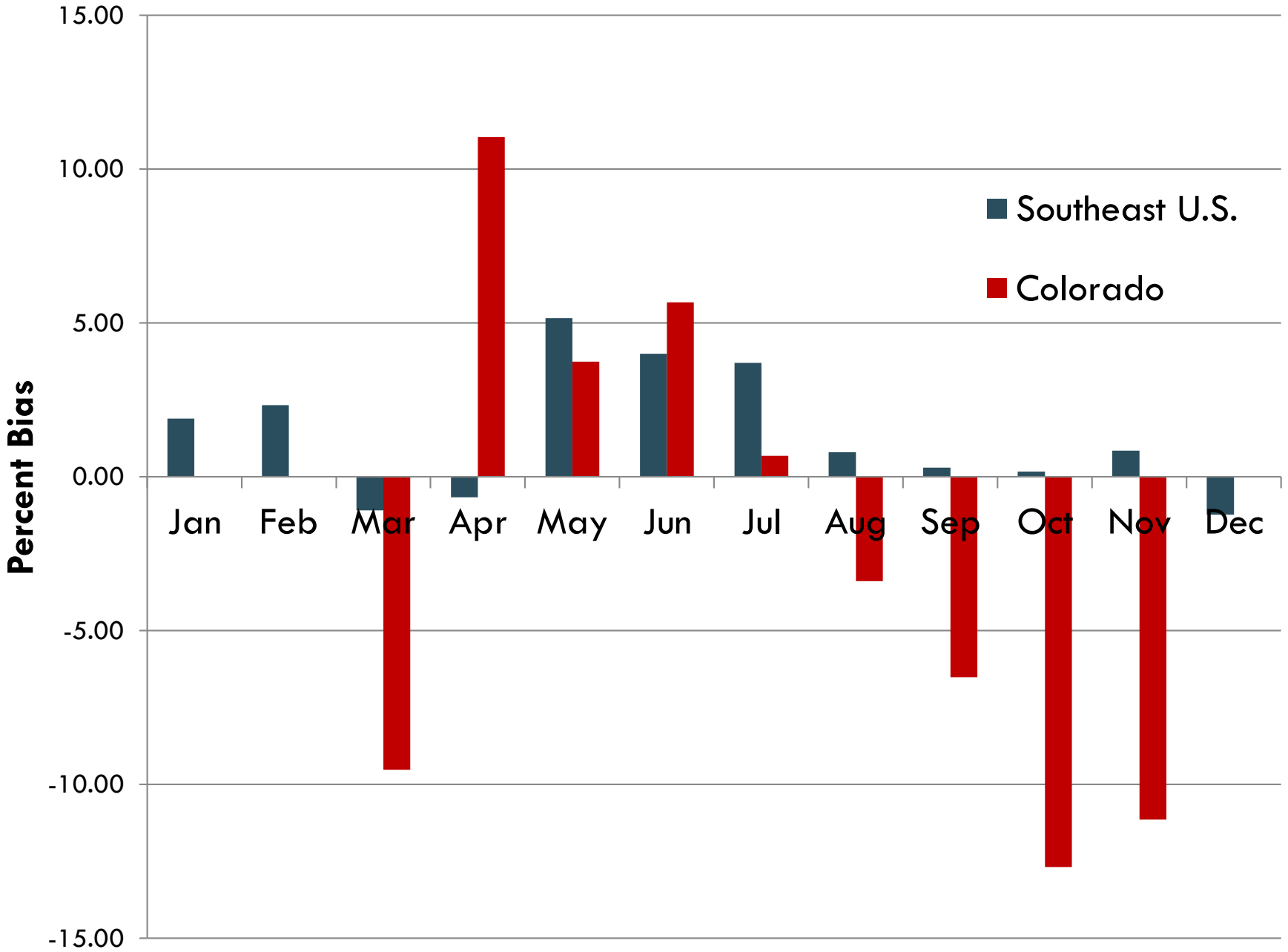
- Data users:
  1. Understand the history of your dataset
  2. Always account for geographic bias
  3. Consider integrating multiple sources of data
- Data managers/providers:
  1. Make digitization metadata available: describe your methods
- Data creators (i.e. collectors):
  1. Collect in under-collected regions
  2. Collect introduced species

Recommendations

- Data users:
  1. Understand the history of your dataset
  2. Always account for geographic bias
  3. Consider integrating multiple sources of data
- Data managers/providers:
  1. Make digitization metadata available: describe your methods
- Data creators (i.e. collectors):
  1. Collect in under-collected regions
  2. Collect introduced species
  3. Consider expanding data collection strategies (e.g., observational data)

Acknowledgements: Thank You!

- Austin Mast (advisor)
- Committee
  - Gil Nelson
  - Scott Burgess
- Alexander Schmidt-Lebuhn
- UROP students
  - Jordan Williams
  - Maddie Funaro

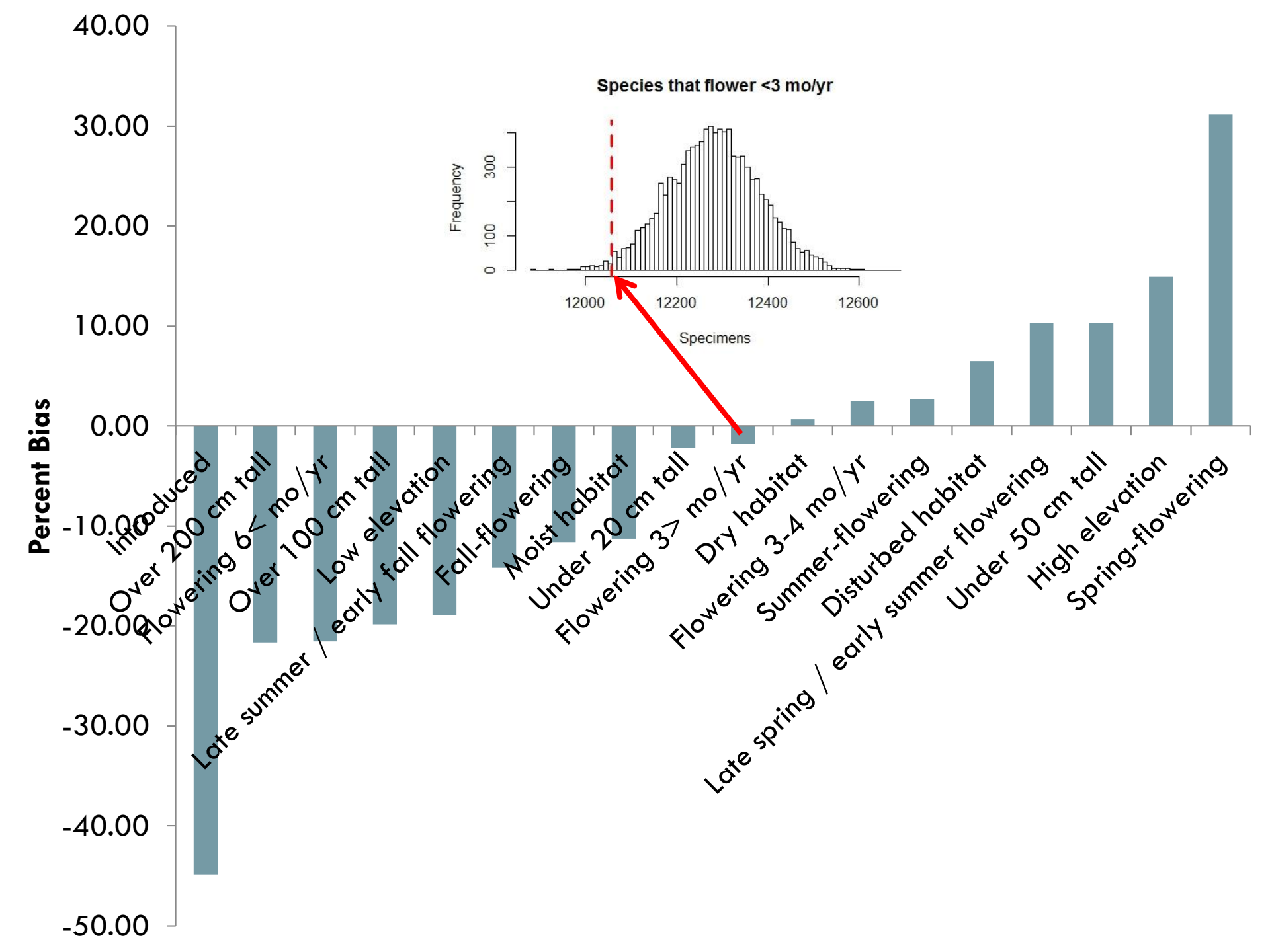

15.00
10.00
5.00
0.00
-5.00
-10.00
-15.00
Percent Bias
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Southeast U.S.
Colorado

Works cited

Parnell JAN, Simpson DA, Moat J, Kirkup DW, Chantaranothai P, Boyce PC, Bygrave P, Dransfield S, Jebb MHP, Macklin J, Meade C, Middleton DJ, Muasya AM, Prajaksood A, Pendry CA, Pooma R, Suddee S, Wilkin P. 2003. Plant collecting spread and densities: their potential impact on biogeographical studies in Thailand. Journal of Biogeography. 30:193-209.

Schmidt-Lebuhn AN, Knerr NJ, Kessler M. 2013. Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). Biodiversity Conservation. 22:905-919.

Yang W, Ma K, Kreft H. 2014. Environmental and socio-economic factors shaping the geography of floristic collections in China. Global Ecology and Biogeography. 23:1284-1292.


Percent Bias
40.00
30.00
20.00
10.00
0.00
-10.00
-20.00
-30.00
-40.00
-50.00
Species that flower <3 mo/yr
Frequency
0
100
300
12000
12200
12400
12600
Specimens
Introduced
Over 200 cm tall
Flowering 6< mo/yr
Over 100 cm tall
Low elevation
Late summer / early fall flowering
Fall-flowering
Moist habitat
Under 20 cm tall
Flowering 3> mo/yr
Dry habitat
Flowering 3-4 mo/yr
Summer-flowering
Disturbed habitat
Late spring / early summer flowering
Under 50 cm tall
High elevation
Spring-flowering