

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

Katelin D. Pearson



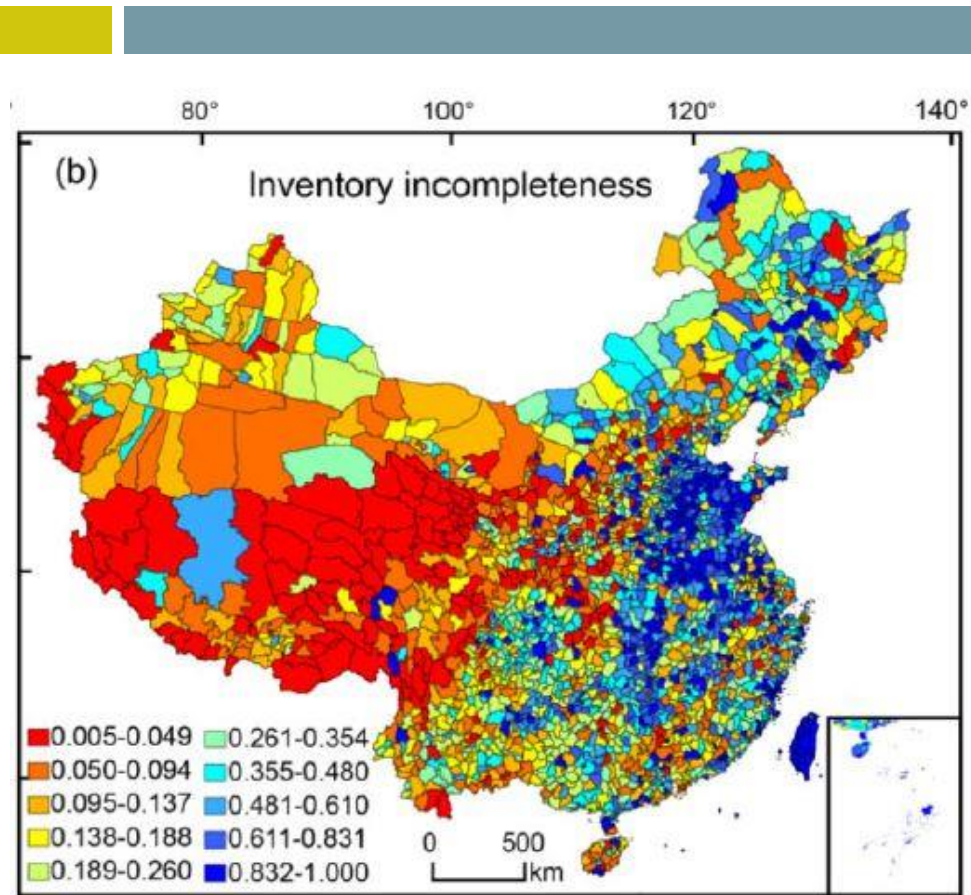
Questions

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”

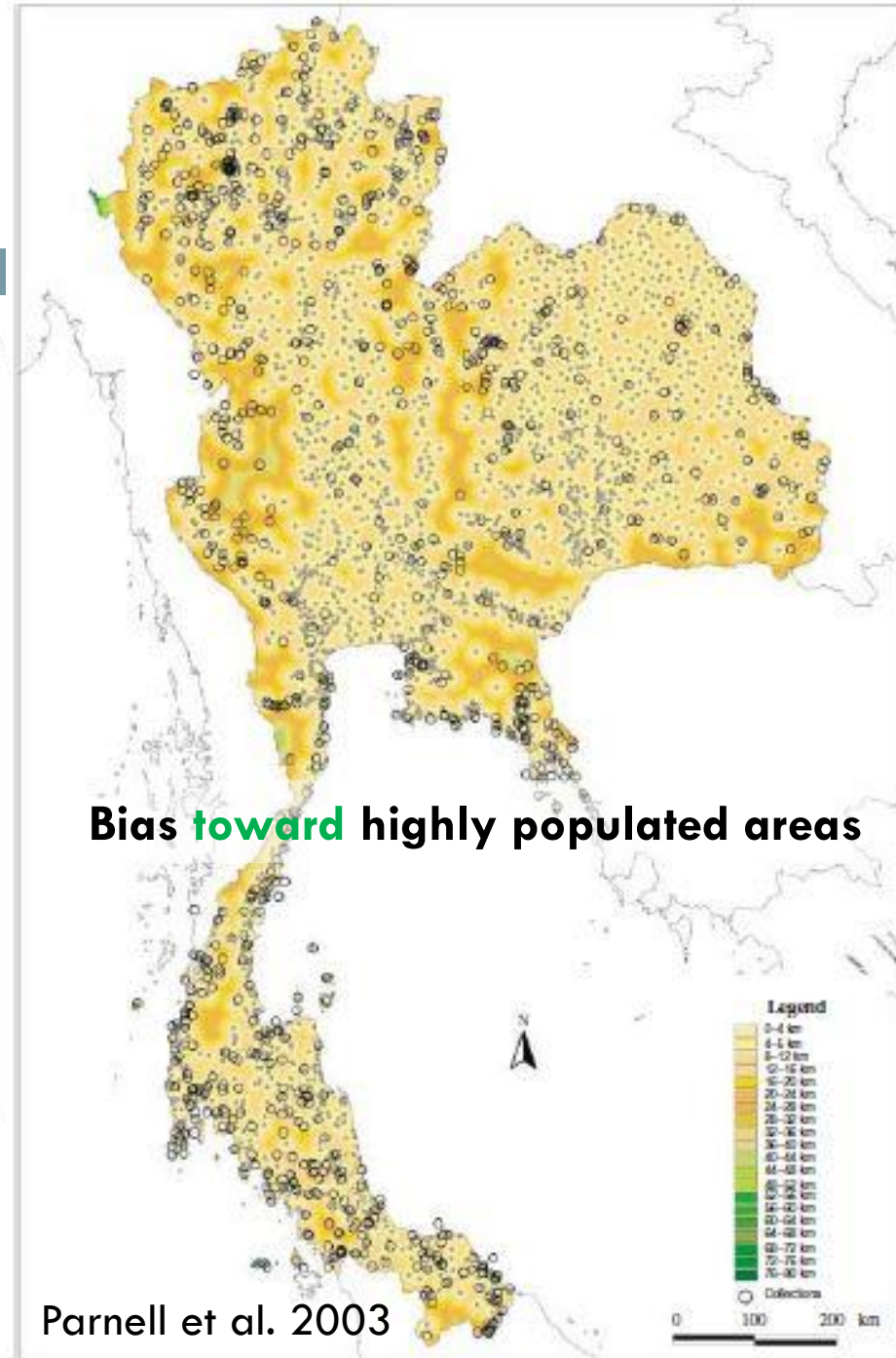
“Class”

| | | Regional | Systematic |
|---------|------------|--|---|
| “Class” | Geographic | - Toward locations that have been locally inventoried | - Toward roadsides, areas near universities or herbaria, natural areas, parks |
| | Temporal | - Toward times during the production of e.g., regional floras | - Against times of war (e.g., World War I and II) |
| | Species | - Toward species of specific interest to local researcher(s) or agencies | - Toward rare species |
| | Individual | - Toward first-flowering individuals in a population | - Toward plants in reproductive condition |

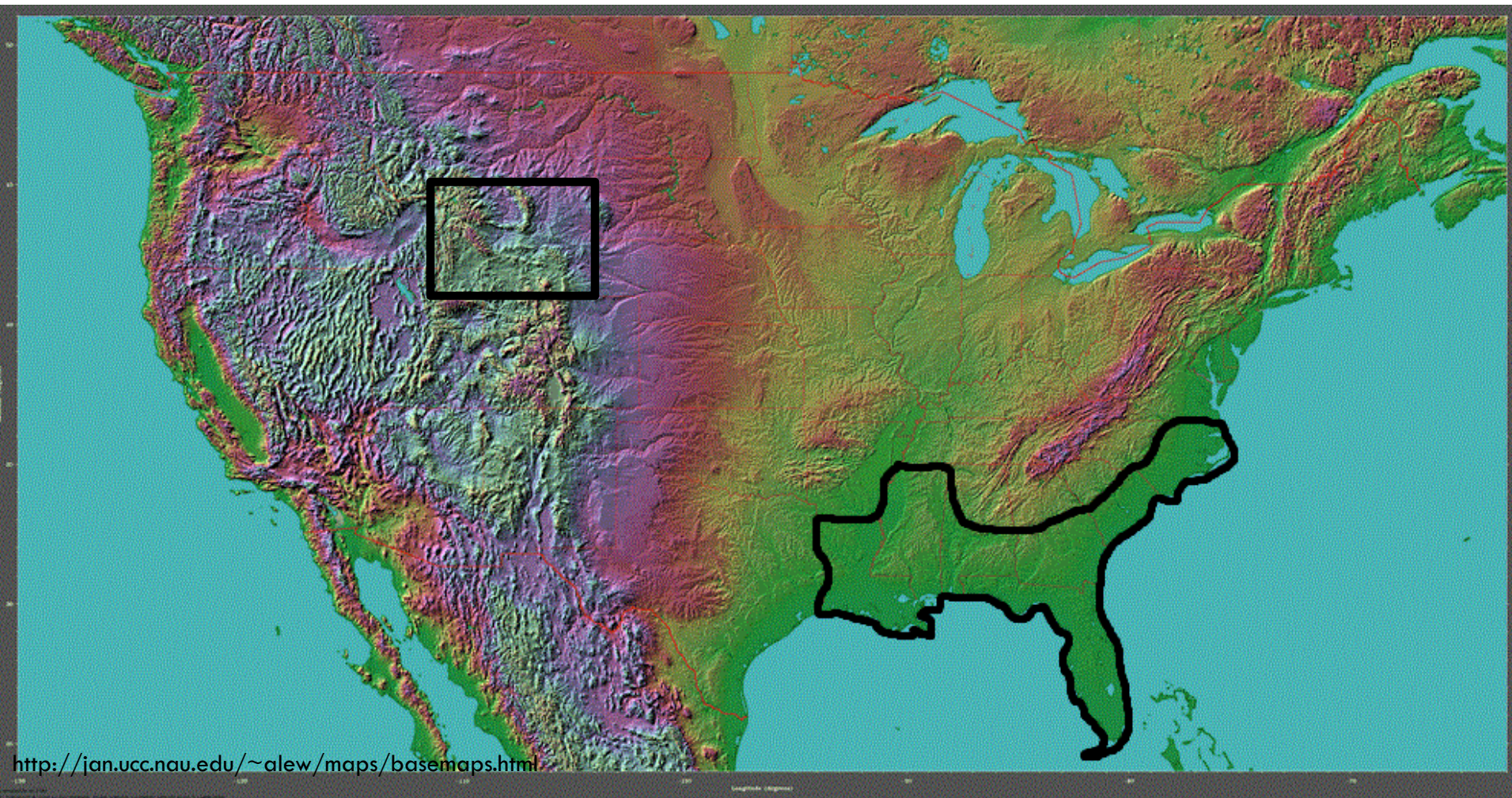


Bias against highly populated areas

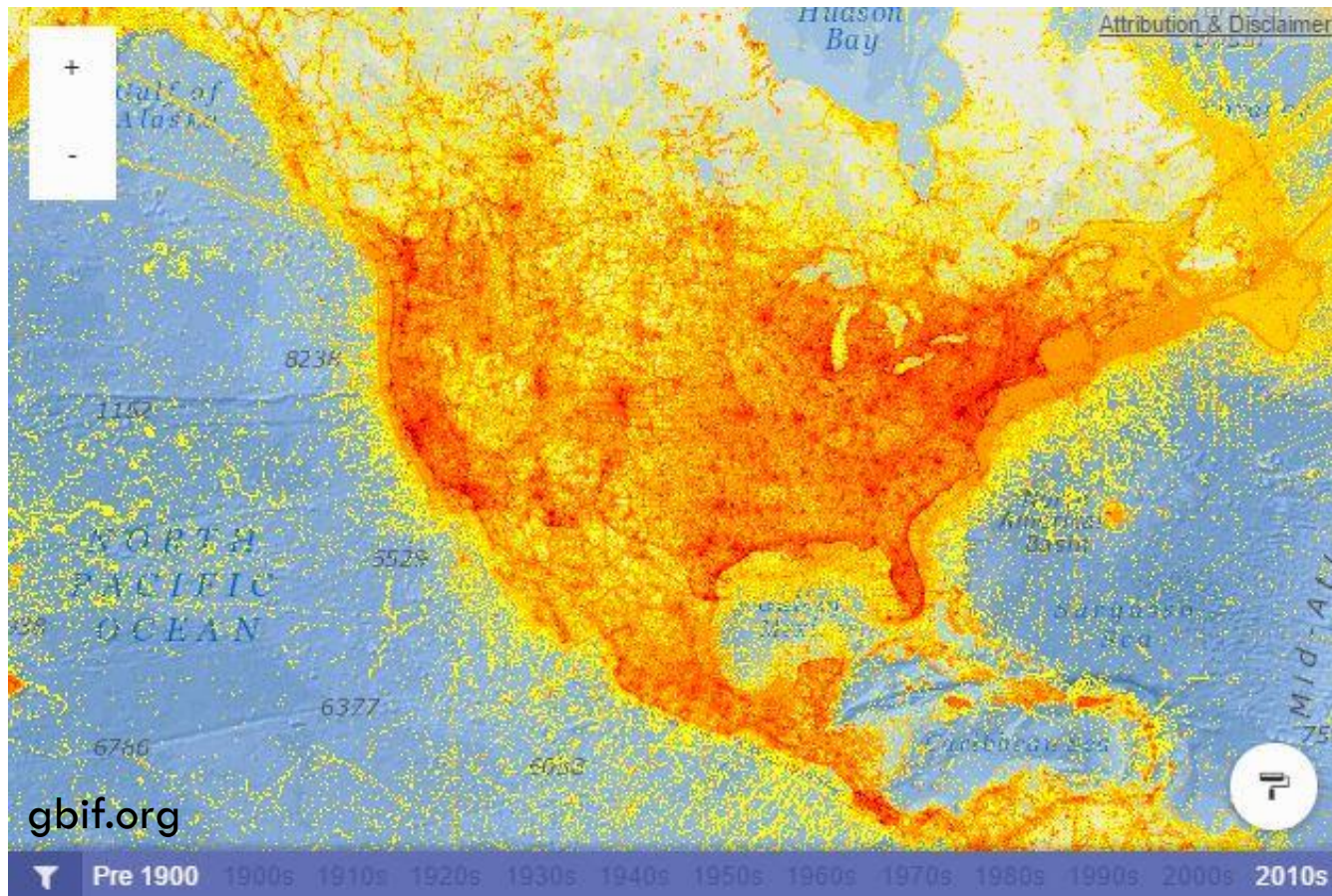
Yang et al. 2014

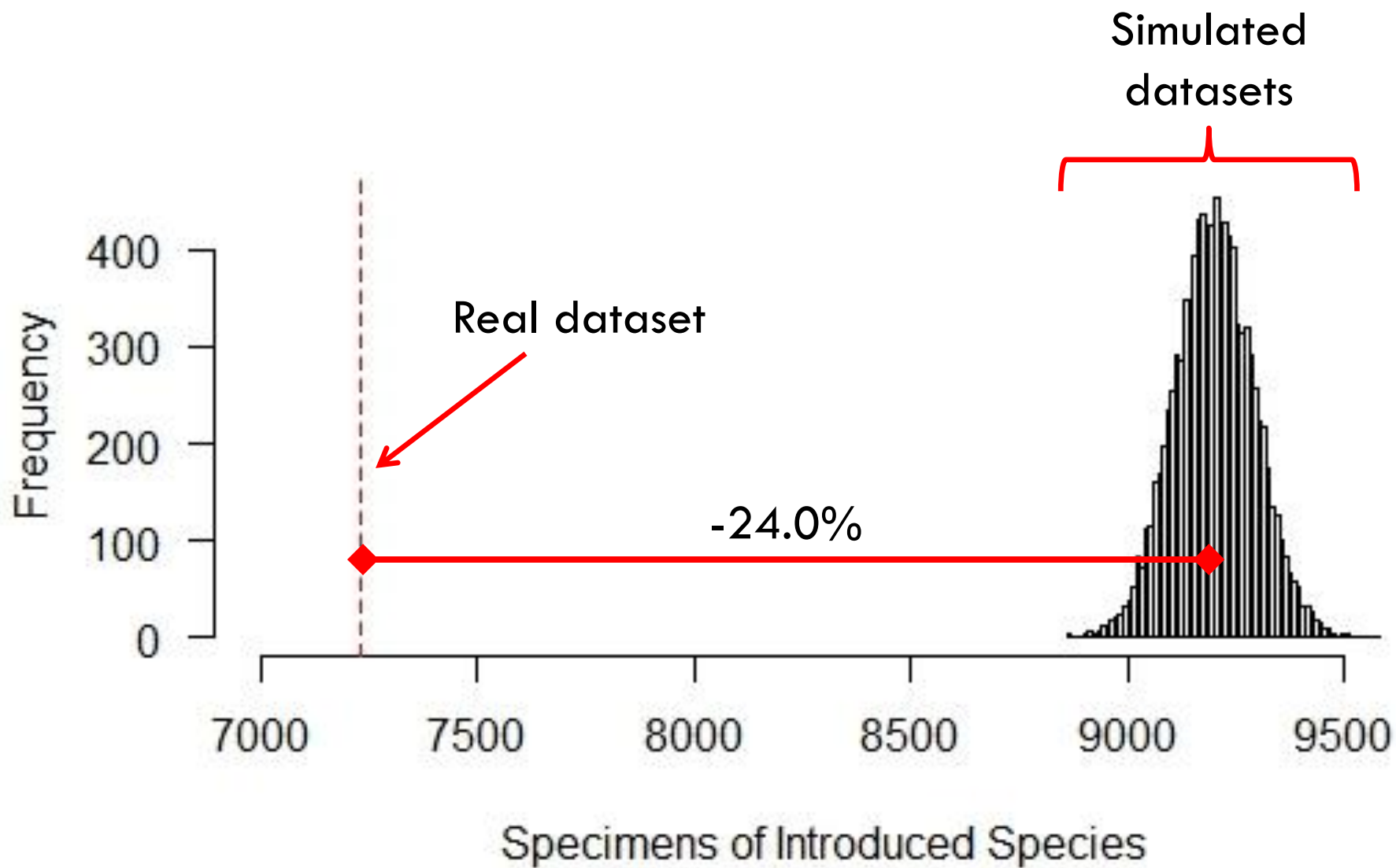


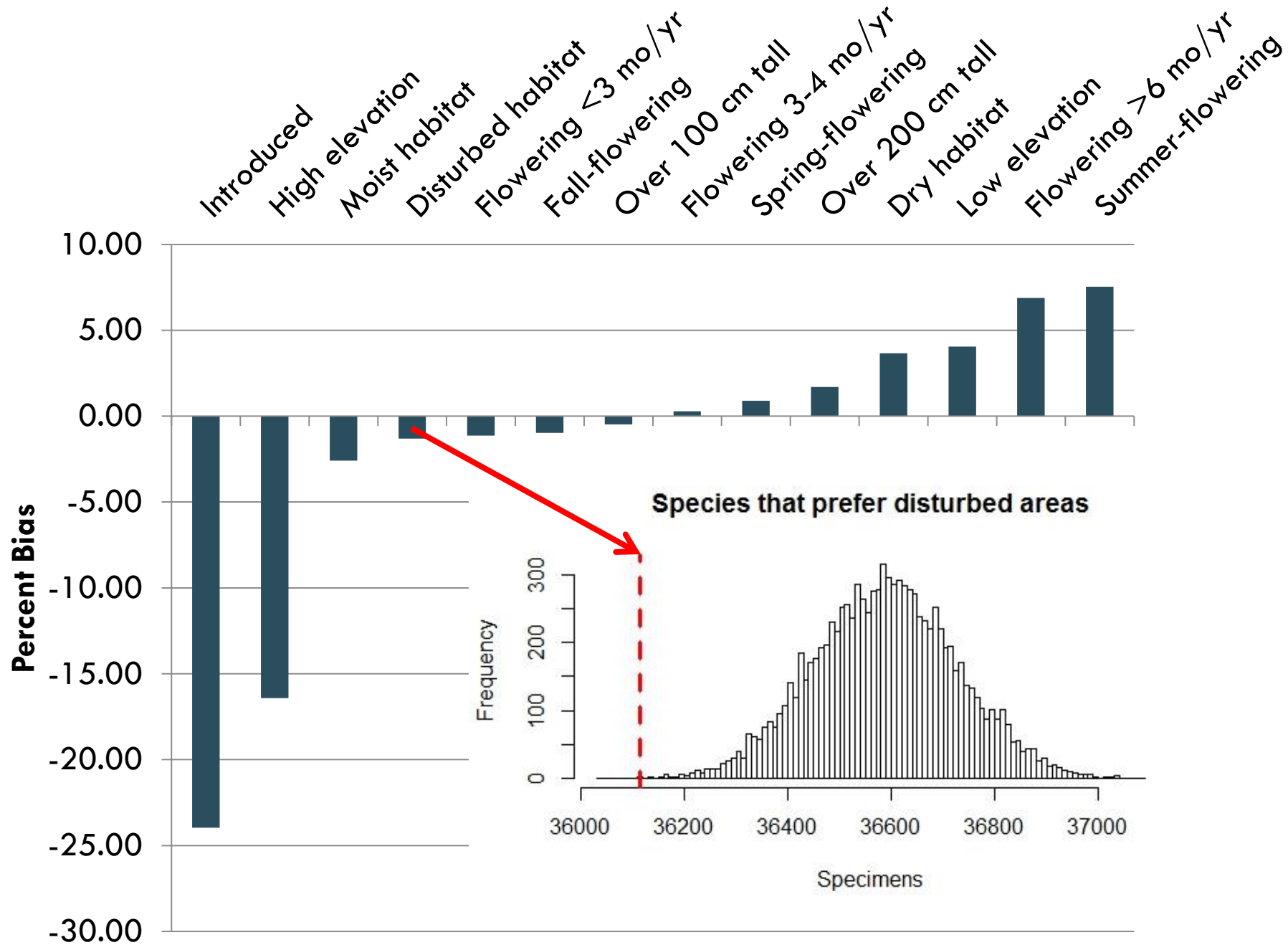
Parnell et al. 2003

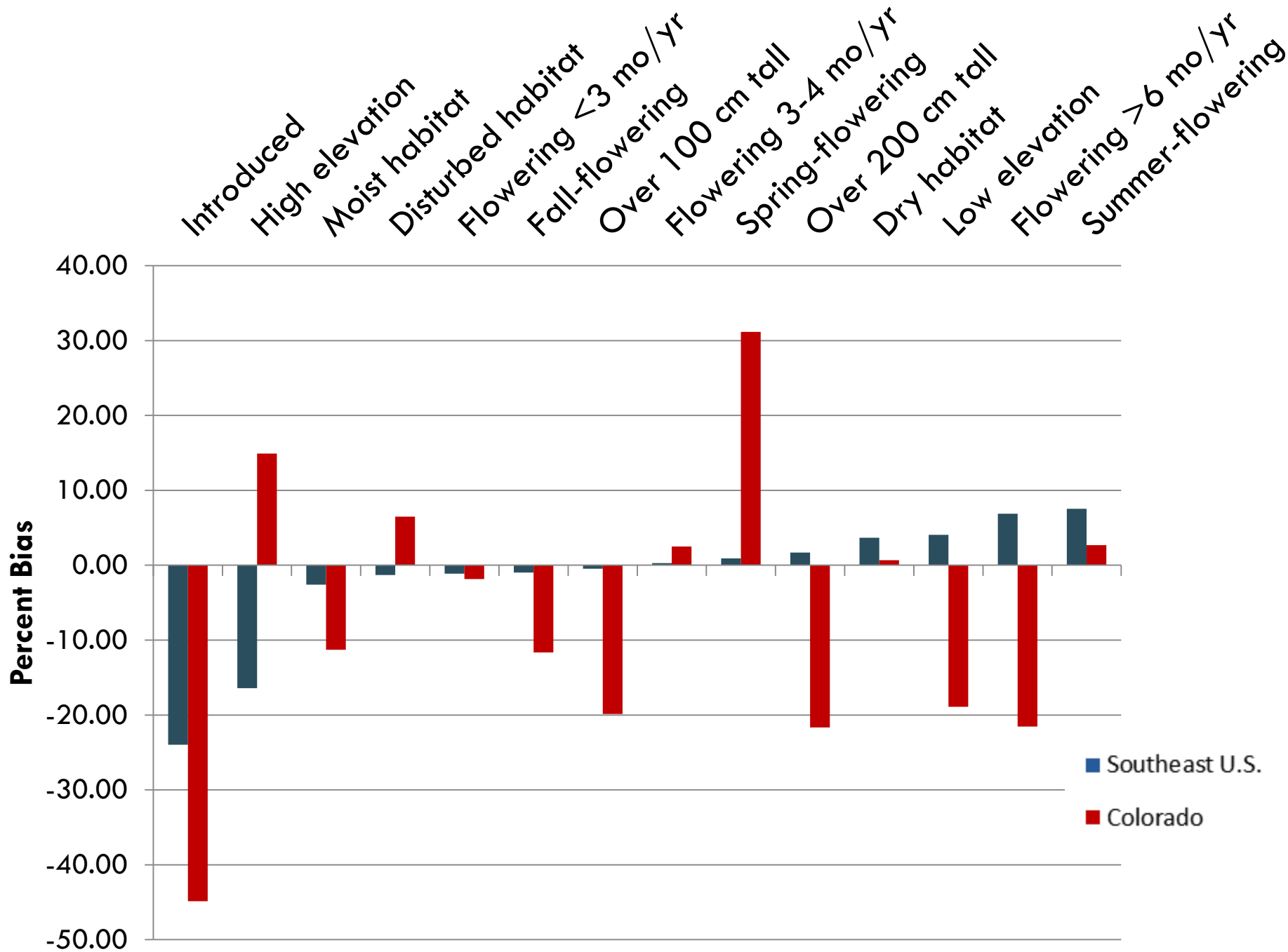


- Simulated 10,000 datasets of “random collections” accounting for geographic collecting bias









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

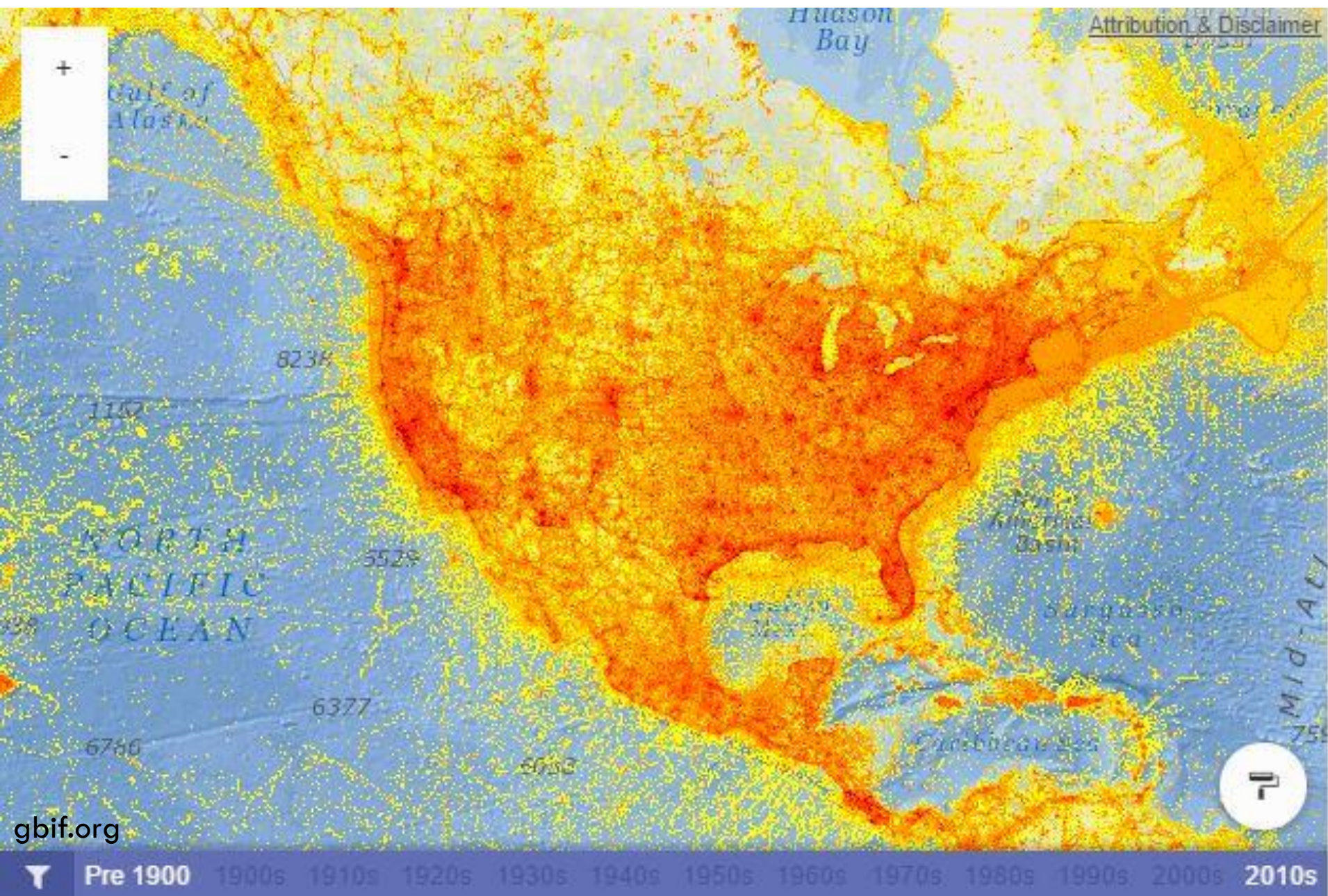
- Data users:

1. Understand the history of your dataset

Recommendations

□ Data users:

1. Understand the history of your dataset
2. Always account for geographic bias



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

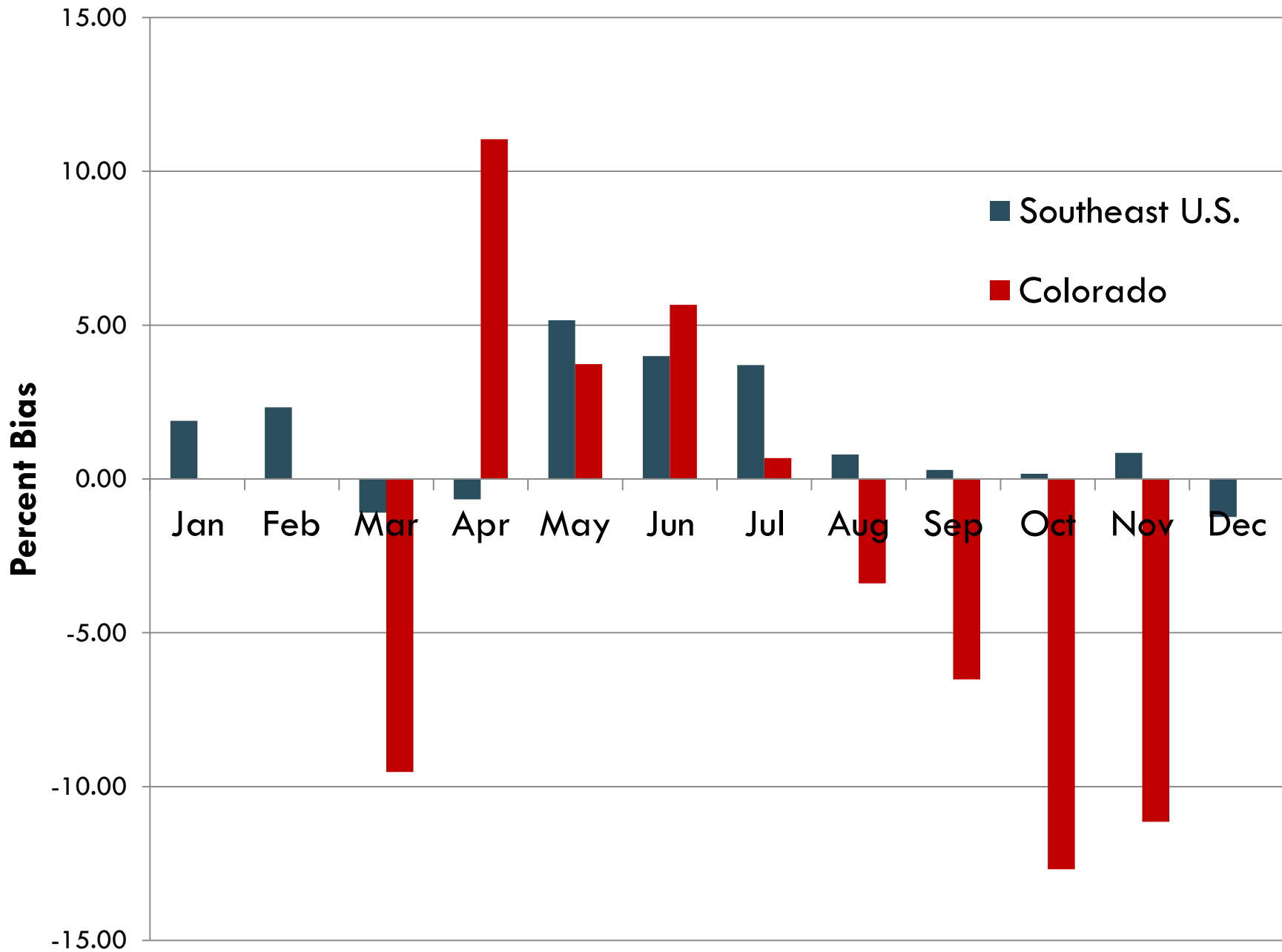
□ Data creators (i.e. collectors):

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





Works cited

- Parnell JAN, Simpson DA, Moat J, Kirkup DW, Chantaranonthai 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.

