BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

Longitudinal effects on plant species involved in agriculture and pandemic emergence undergoing changes in abiotic stress

Mikaela Cashman[1], Verónica G. Melesse Vergara[2], John Lagergren[2], Matthew Lane[3], Jean Merlet[3], Mikaela Atkinson[3], Jared Streich[2], Christopher Bradburne[4], Raina Plowright[5], Wayne Joubert[2], Daniel Jacobson[2]

OAK RIDGE National Laboratory

THE UNIVERSITY OF TENNESSEE KNOXVILLE

[1]Lawrence Berkeley National Laboratory
[2]Oak Ridge National Laboratory
[3]University of Tennessee
[4]Johns Hopkins University Applied Physics Laboratory
[5]Cornell University

JOHNS HOPKINS APPLIED PHYSICS LABORATORY

Cornell University

June 26th 2023 - PASC'23

Motivation: Environmental Stress

Plant and animal species are under selective pressure

Additionally: Land-use change and a rapidly changing climate adds unprecedented pressure

Identify environmental similar geospatial zones & quantify the change each area is experiencing

Identify potentially suitable areas not currently in use & detect the level of abiotic stress

Method Overview

Detect global regions correlated by environmental features from longitudinal and agglomerative perspectives

- We leverage and enhance a high-performance computing methodology[1]

  - improve computational efficiency
  - remove bias against extreme climates
  - scale from 500,000 to 8.8 million dry land points

- We demonstrate the applicability on species of interest in:
  - agriculture (e.g., coffee, wine, chocolate)
  - bioenergy (e.g., poplar, switchgrass, pennycress)
  - zoonotic spillover (e.g., eucalyptus, flying foxes)

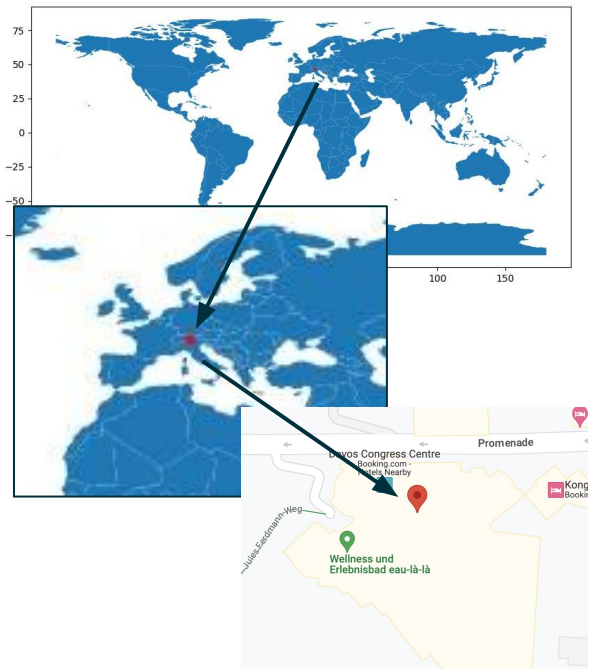
[1] John Lagergren, Mikaela Cashman, Melesse Vergara Veronica G., P. R. Eller, J. Gabriel Felipe Machado Gazolla, Hari B. Chhetri, Jared Streich, S. Climer, Peter Thornton, Wayne Joubert, and Dan Jacobson. 2022. Climatic clustering and longitudinal analysis with impacts on food, bioenergy, and pandemics. Phytobiomes (Sep 2022)

Methods

Climatic clustering and analysis over multiple geological and longitudinal perspectives

Methods | Overview

(1) Vector Generation



latitude: 46.801

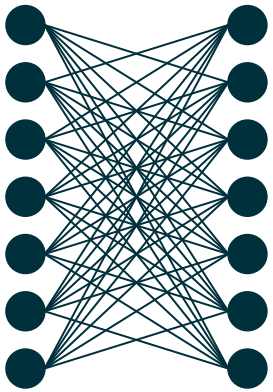
longitude: 9.831

<feature1, feature2, …>

(2) Correlation Computation

<feature1, feature2, …>

<feature1, feature2, …>

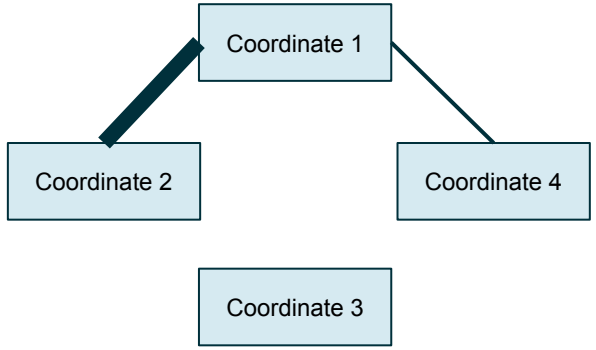


(3) Correlation Output

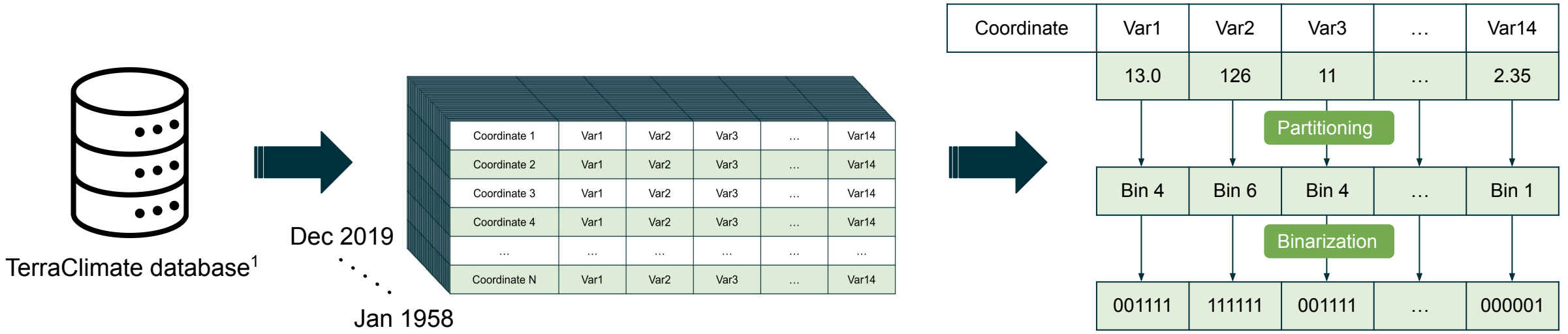
Coordinate 1	Coordinate 2	**0.894**
Coordinate 1	Coordinate 3	**0.012**
Coordinate 1	Coordinate 4	**0.543**
.....		
Coordinate N	Coordinate N-3	**0.999**
Coordinate N	Coordinate N-2	**0.924**
Coordinate N	Coordinate N-1	**0.003**

(4) Analysis

Graph representation



Methods | Vector Generation



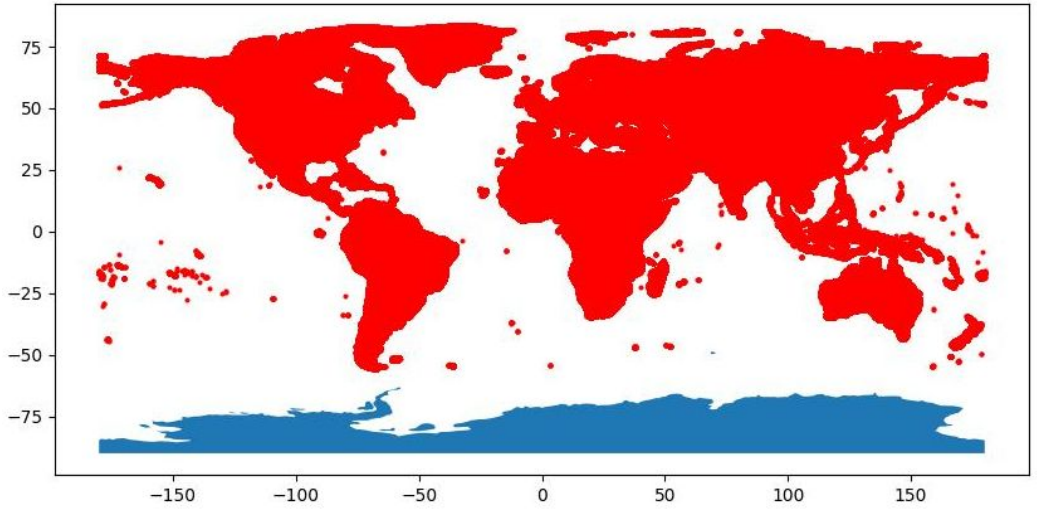
Climatic variables

minimum temperature
temperature range
vapor pressure
precipitation accumulation
downward surface shortwave radiation
wind-speed
reference evapotranspiration
runoff
actual evapotranspiration
climate water deficit
soil moisture
snow water equivalent
palmer drought severity index
vapor pressure deficit

[1] John T Abatzoglou, Solomon Z Dobrowski, Sean A Parks, and Katherine C Hegewisch. 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Scientific data 5, 1 (2018), 1–12.

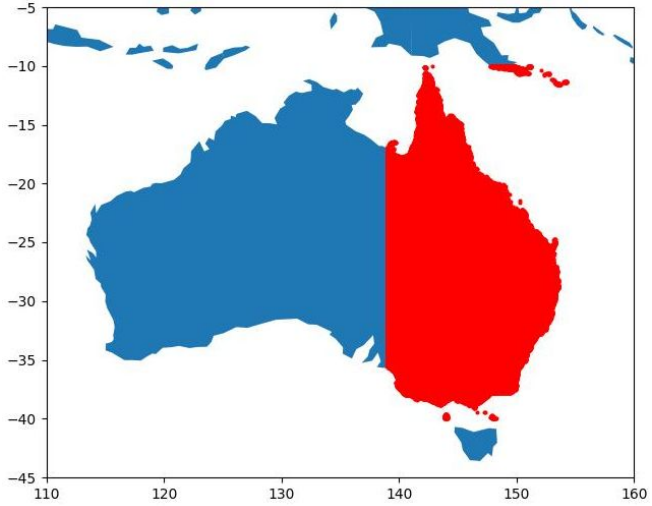
Methods | Coordinate Perspectives

(1) Global View – 8,834,910 dry-land geolocations



*Selected coordinates in red

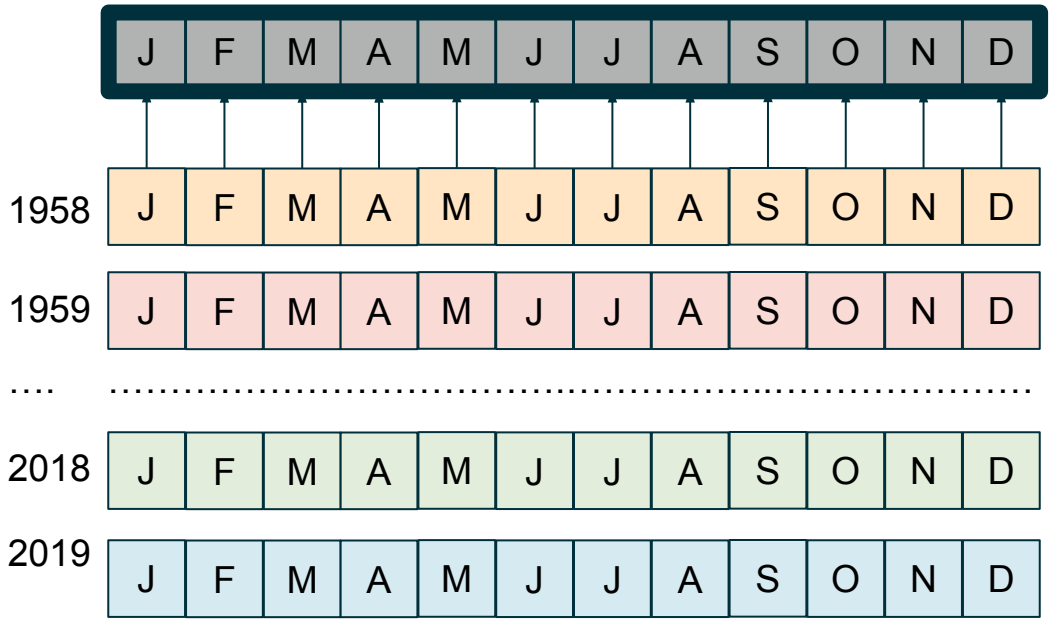
(2) Regional view of Eastern Australia – 153,149 dry-land geolocations



Methods | Longitudinal Perspectives

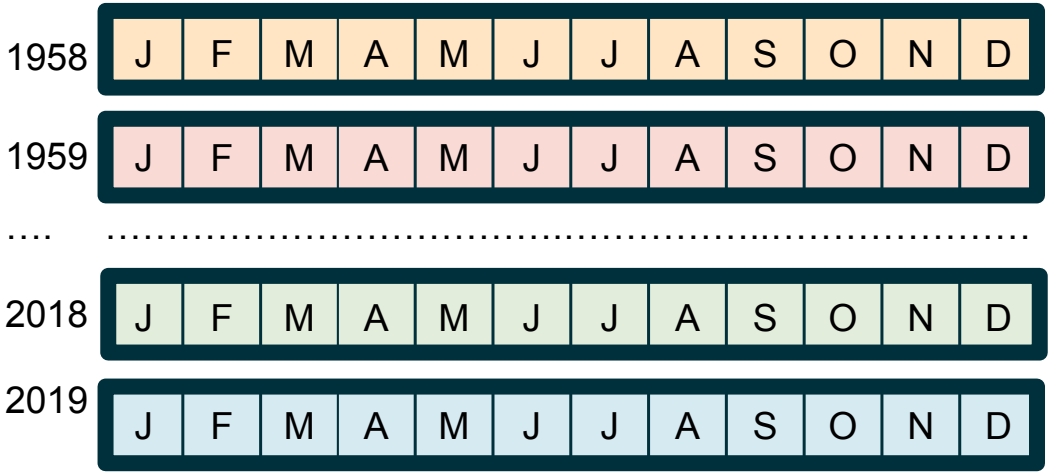
Agglomerative Perspective (Monthly Means)

(x1) 12 months of 62 years averaged of size 8,404
(14 variables * 12 months *50 bits)



Longitudinal Perspective (Yearly)

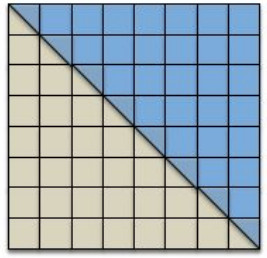
(x62) 62 years split into 62 vectors each of size 8,404
(14 variables * 12 months *50 bits)



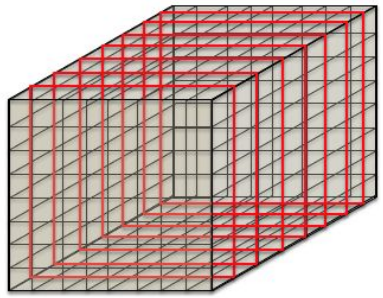
Methods | Correlation Computation

CoMet

- **Co**mbinatorial **Met**rics (CoMet) library[1,2,3]
  - Exhaustively compute similarity metrics
  - ultra-low precision mathematics using binary data
  - 1-bit general matrix-matrix multiplications



2-way comparisons



3-way comparisons

summit[3]

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

	Node hours	Storage
Global Agglomerative	1,973	15.6 TB
Global Yearly	113,865	1.3 PB

[1] Wayne Joubert, James Nance, Sharlee Climer, Deborah Weighill, and Daniel Jacobson. 2019. Parallel accelerated Custom Correlation Coefficient calculations for genomics applications. Parallel Comput. 84 (may 2019), 15–23.

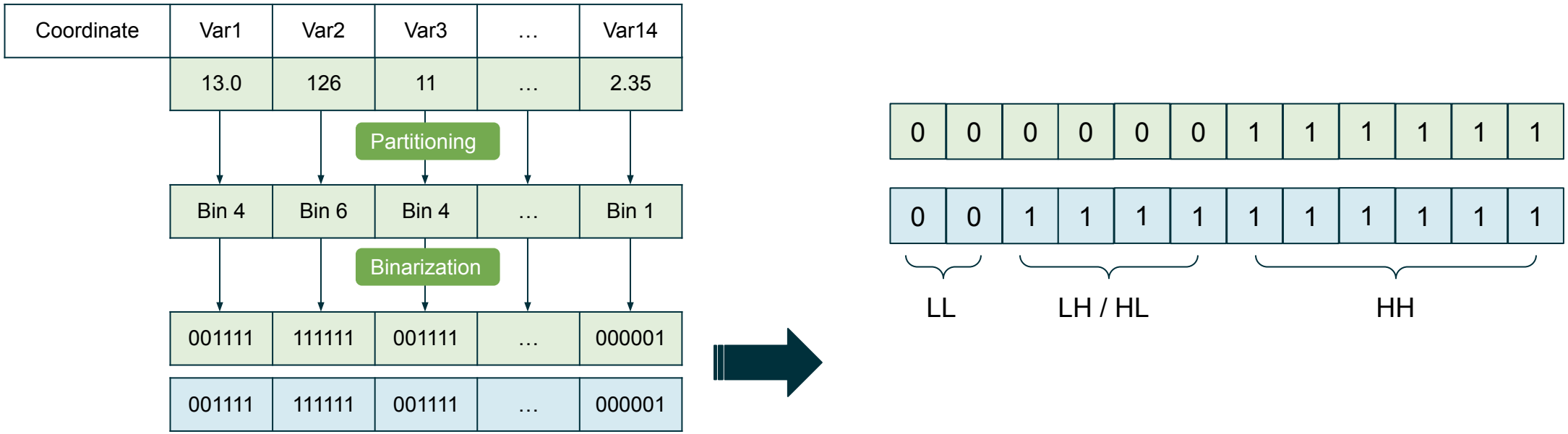
[2] Wayne Joubert, James Nance, Deborah Weighill, and Daniel Jacobson. 2018. Parallel accelerated vector similarity calculations for genomics applications. Parallel Comput. 75 (2018), 130–145.

[3] Wayne Joubert, Deborah Weighill, David Kainer, Sharlee Climer, Amy Justice, Kjiersten Fagnan, and Daniel Jacobson. 2018. Attacking the Opioid Epidemic: Determining the Epistatic and Pleiotropic Genetic Architectures for Chronic Pain and Opioid Addiction. In Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '18). IEEE Press, Piscataway, NJ, USA, Article 57.

[3] Summit: Scale new heights. Discover new Solutions. https://www.olcf.ornl.gov/summit/

Methods | Correlation Metric

- Binary formatting of feature information translates into a High (1) feature value or a Low (0) feature value
  - → results in four categories of relationships: High-High (1, 1), High-Low (1, 0), Low-High (0, 1), and Low-Low (0, 0)



- 3-way: results in eight possible categories of relationships: HHH, HHL, LHH, HLH, HLL, LLH, LHL, and LLL

Methods | Correlation Metric

The Duo metric between two vectors i and j is defined as:

$$\mathrm{Duo}_{i,j}(r) = 4D_{i,j}(r)\left(1 - \frac{f_i(r)}{q}\right)\left(1 - \frac{f_j(r)}{q}\right)$$

frequency terms

relationship type (e.g. HL)

proportion of vectors with the given relationship

scaling factor

$$D_{i,j}(r) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{i_n=r_1, j_n=r_2\}}$$

relationship type

vector position

indicator function

$$f_i(r) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{i_n=r_1\}}$$

$$f_j(r) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{j_n=r_2\}}$$

relationship type

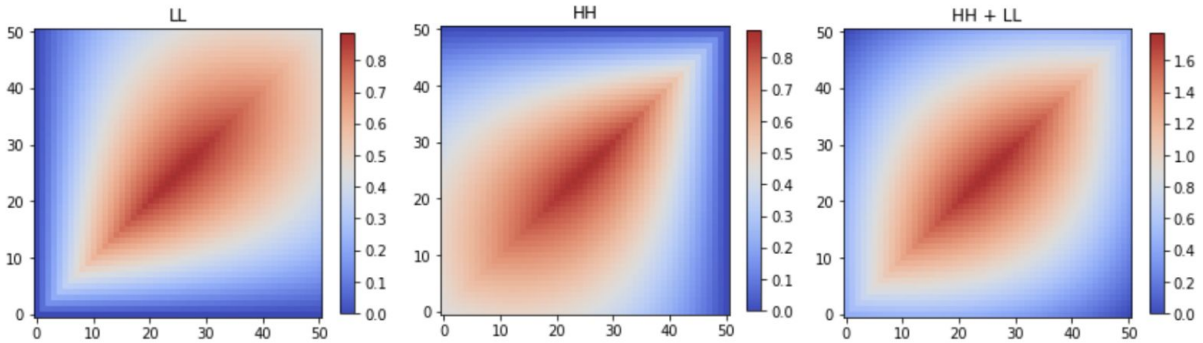
[1] Sharlee Climer, Alan R Templeton, Michael Garvin, Daniel Jacobson, Matthew Lane, Scott Hulver, Brittany Scheid, Zheng Chen, Carlos Cruchaga, and Weixiong Zhang. 2020. Synchronized genetic activities in Alzheimer's brains revealed by heterogeneity-capturing network analysis. bioRxiv (2020).

Methods | Enhancement #1

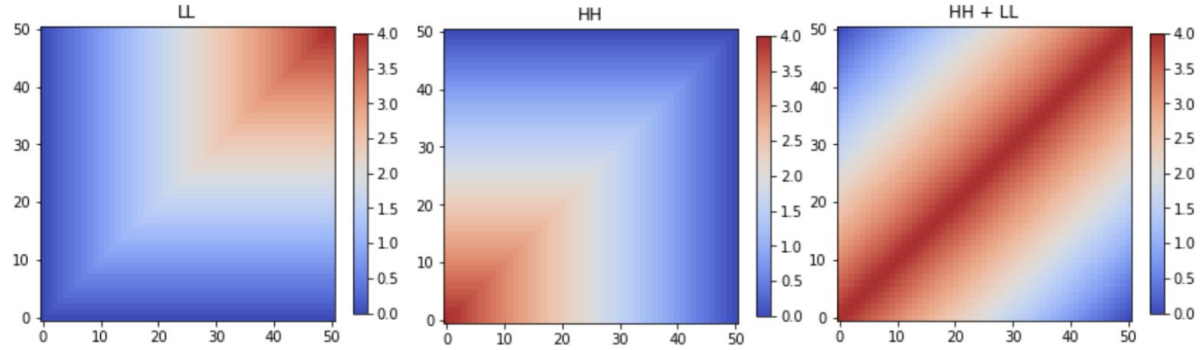
$$\mathrm{Duo}_{i,j}(r) = 4D_{i,j}(r)\left(1 - \frac{f_i(r)}{q}\right)\left(1 - \frac{f_j(r)}{q}\right)$$

Duo Frequency

- Penalize extremely rare case in genomics studies
- Irrelevant to this application
- Causes bias against extreme climate conditions
- Modification: adjust 1/q from 3/2 to 0 to eliminate
- Equivalent to Sørensen-Dice



(a) Original Duo scaling factor $1/q = 3/2$



(b) Modified Duo scaling factor $1/q = 0$

Methods | Correlation Output

High-High

Coordinate 1	Coordinate 2	**0.894**
Coordinate 1	Coordinate 3	**0.012**
Coordinate 1	Coordinate 4	**0.543**
.....		
Coordinate N	Coordinate N-3	**0.999**
Coordinate N	Coordinate N-2	**0.924**
Coordinate N	Coordinate N-1	**0.003**

High-Low

Coordinate 1	Coordinate 2	**0.894**
Coordinate 1	Coordinate 3	**0.012**
Coordinate 1	Coordinate 4	**0.543**
.....		
Coordinate N	Coordinate N-3	**0.999**
Coordinate N	Coordinate N-2	**0.924**
Coordinate N	Coordinate N-1	**0.003**

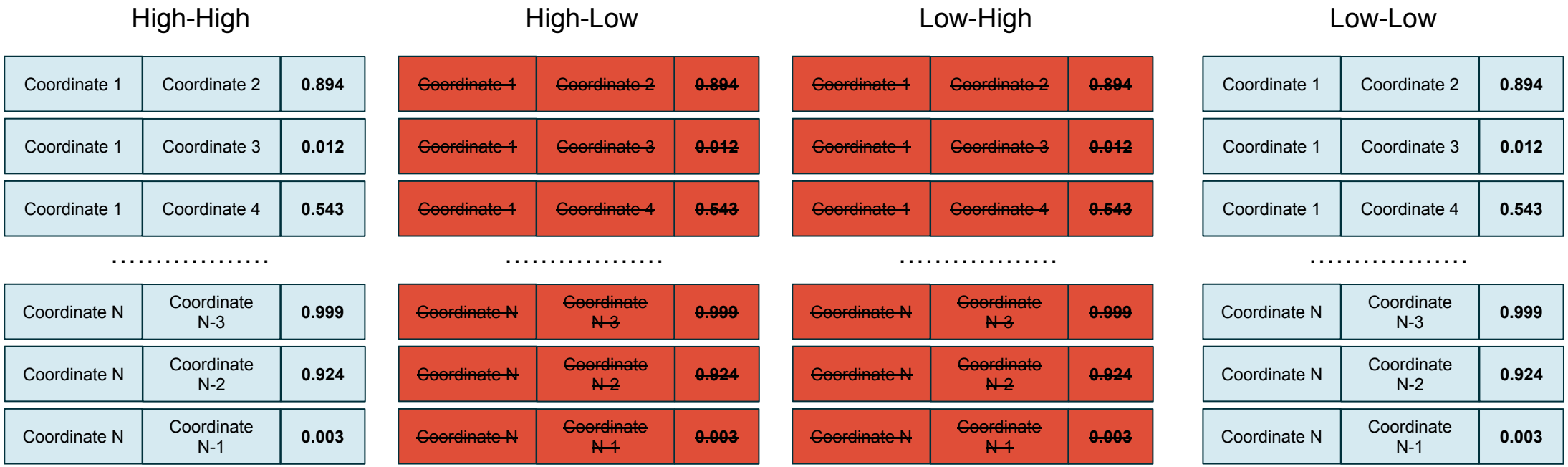
Low-High

Coordinate 1	Coordinate 2	**0.894**
Coordinate 1	Coordinate 3	**0.012**
Coordinate 1	Coordinate 4	**0.543**
.....		
Coordinate N	Coordinate N-3	**0.999**
Coordinate N	Coordinate N-2	**0.924**
Coordinate N	Coordinate N-1	**0.003**

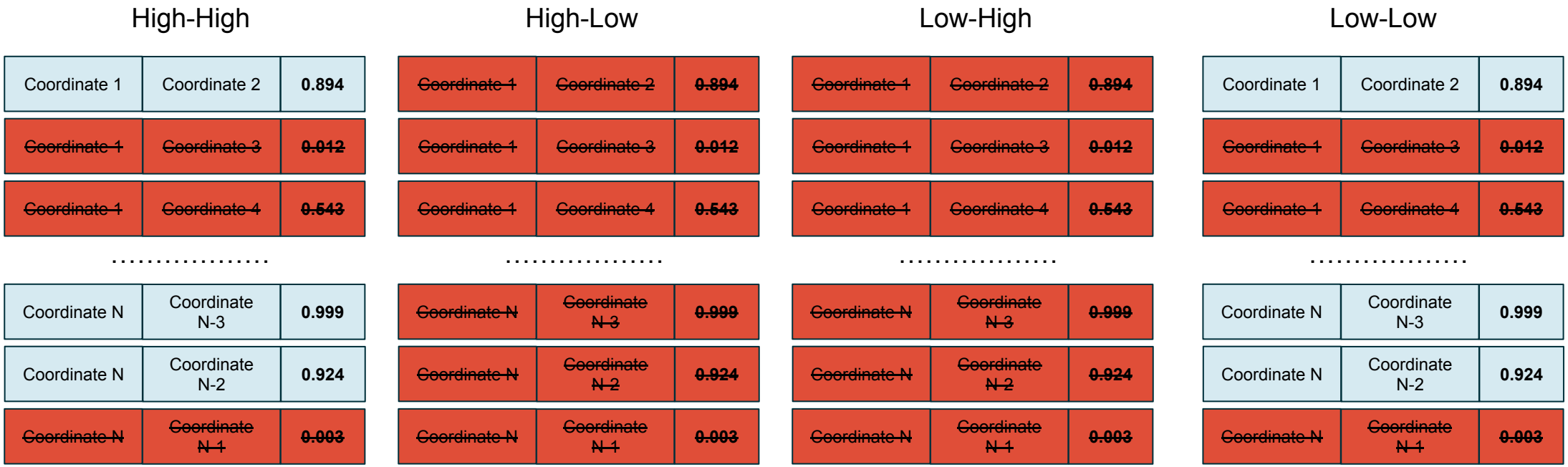
Low-Low

Coordinate 1	Coordinate 2	**0.894**
Coordinate 1	Coordinate 3	**0.012**
Coordinate 1	Coordinate 4	**0.543**
.....		
Coordinate N	Coordinate N-3	**0.999**
Coordinate N	Coordinate N-2	**0.924**
Coordinate N	Coordinate N-1	**0.003**

Methods | Correlation Output



Methods | Correlation Output

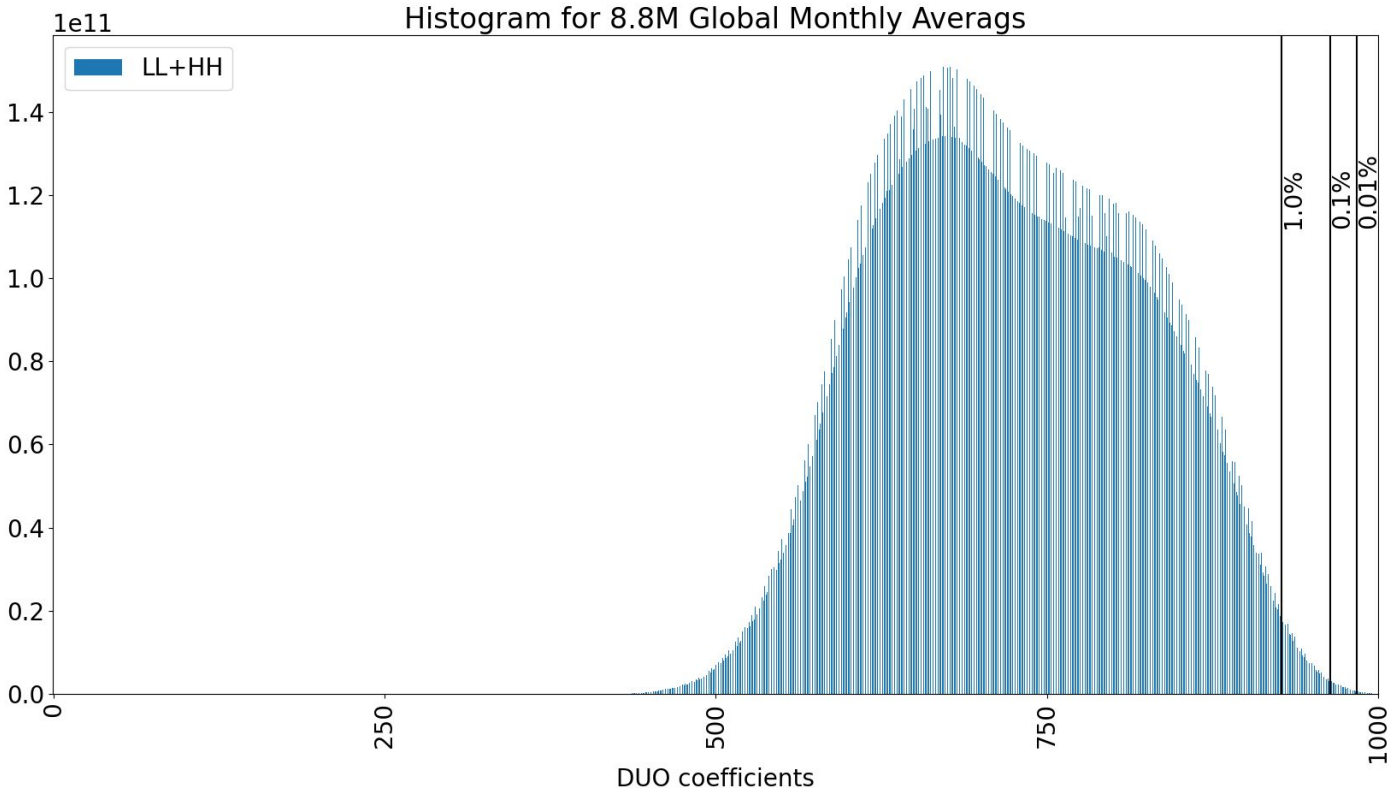


How to choose a sensible threshold?

Methods | Enhancement #2

Histogram Method

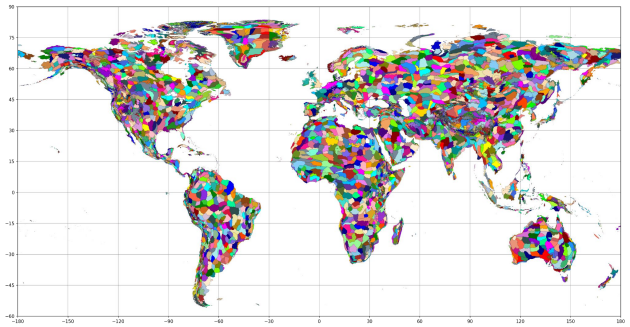
- Store distribution of relation types
- Minimal I/O
- Allows user to know *a priori* how many metrics will be stored for a given threshold and relation type



Methods | Analysis

Network Clustering

- **Source data:** graphical representation of correlated geolocations
- **Method:** High performance Markov Clustering (HipMCL[1])
- **Result:** high-resolution clusters defining climatic zones with similar characteristics

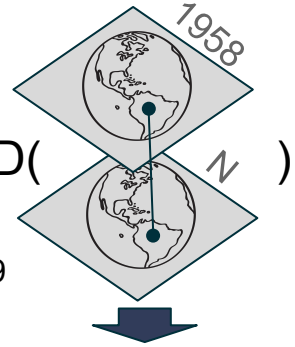


[1] Ariful Azad, Georgios A Pavlopoulos, Christos A Ouzounis, Nikos C Kyrpides, and Aydin Buluç. 2018. HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. Nucleic acids research 46, 6 (2018), e33–e33

Correlations-of-Correlations

- Time series of global networks

$$\forall_{\text{coordinates}}^{\frac{1}{N}} \sum_{N = 1959}^{2019} SD(\ 1958\ ,\ N\ )$$



Relative CorCor

Adjacency neighbor vectors

$$SD\left(\begin{matrix}1958\ \{V_3, V_5, V_6, V_{20}, V_{21}\}\\ N\ \ \{V_3, V_5, V_6\}\end{matrix}\right)$$

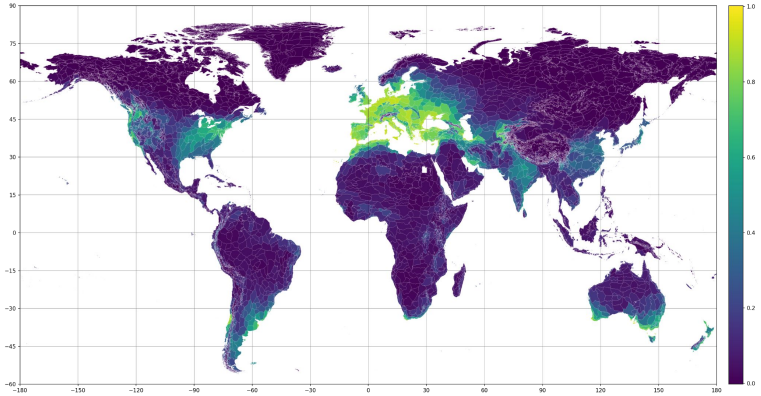
Absolute CorCor

Raw feature vectors

$$SD\left(\begin{matrix}1958\ \{14.0, 0.54, 1034, \ldots 50.3\}\\ N\ \ \{12.8, 0.53, 1034, \ldots 45.5\}\end{matrix}\right)$$

Species Distribution Modeling

- **Method:** Overlay climatype clusters + species distribution models
- **Models:** statistical machine-learning Maximum Entropy (Maxent[2]) model
  - species occurrence + environmental data
  - predicted probability distribution[3]



[2] Steven J. Phillips, Miroslav Dudík, and Robert E. Schapire. Maxent software for modeling species niches and distributions (Version 3.4.1). http: //biodiversityinformatics.amnh.org/open_source/maxent/

[3] JaneElith,StevenJPhillips,TrevorHastie,MiroslavDudík,YungEnChee,and Colin J Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and distributions 17, 1 (2011), 43–57.

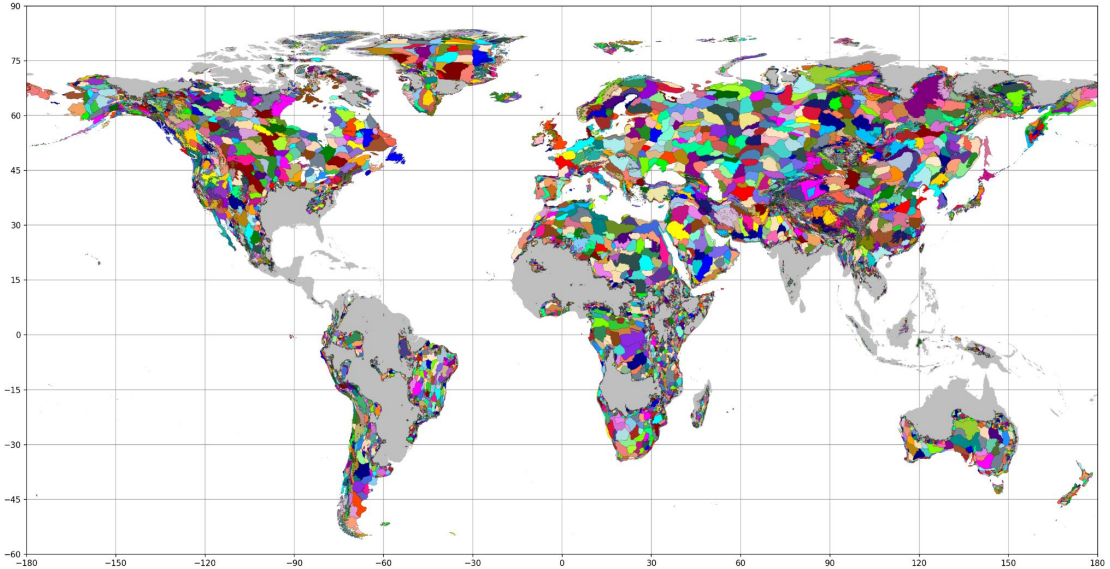
Results

Climatic clustering and analysis over multiple geological and longitudinal perspectives

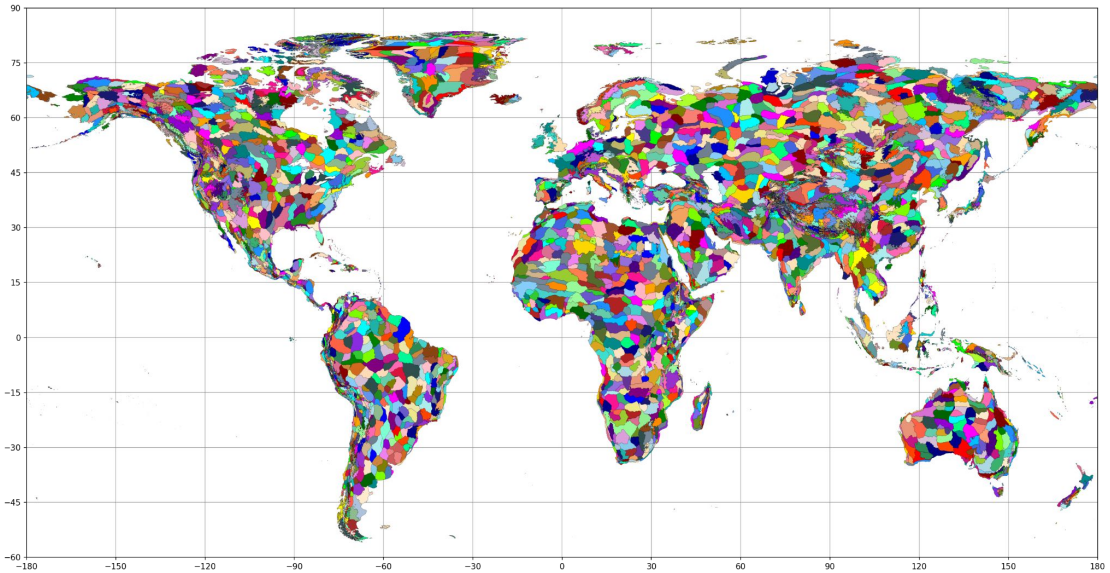
Results | Duo vs Sørensen-Dice

Duo
(37,522,455,884 edges)

*Gray indicates coordinate not present in any edge

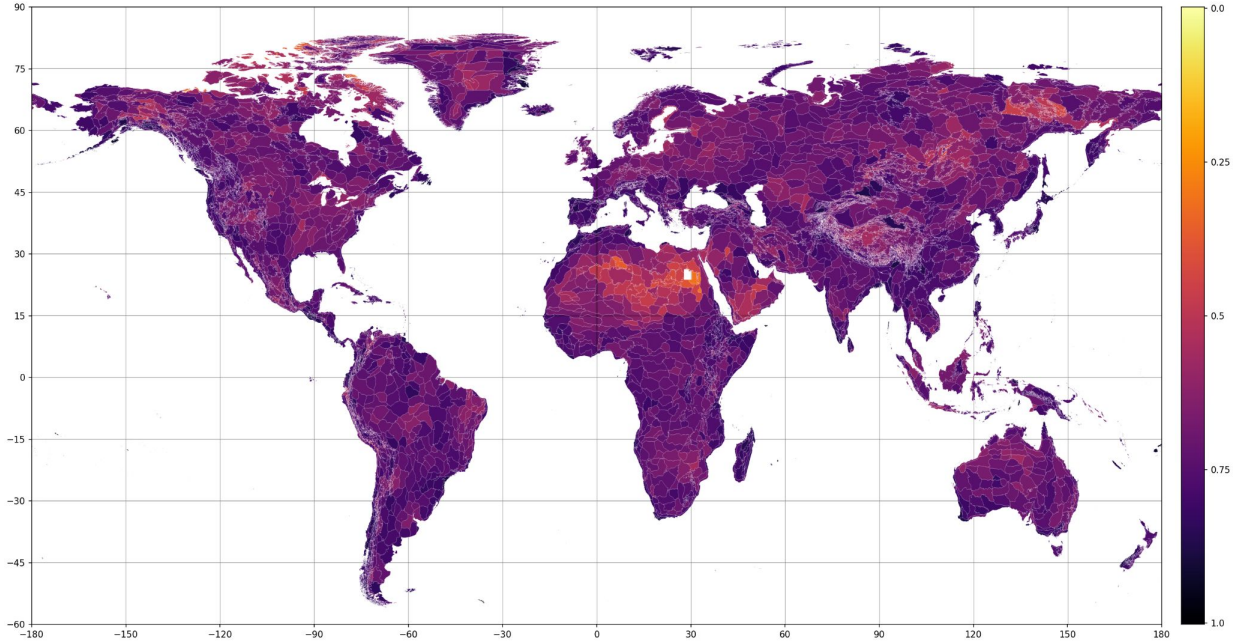


Sørensen-Dice
(36,712,590,809 edges)

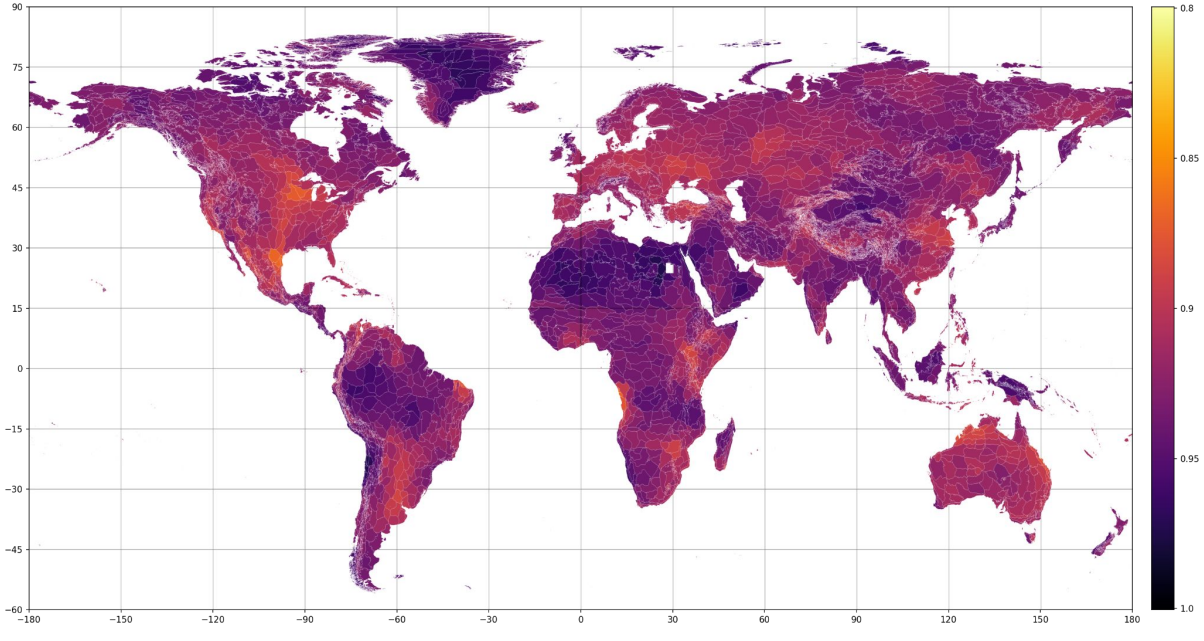


Results | Longitudinal Perspectives

Relative Cor-Cor



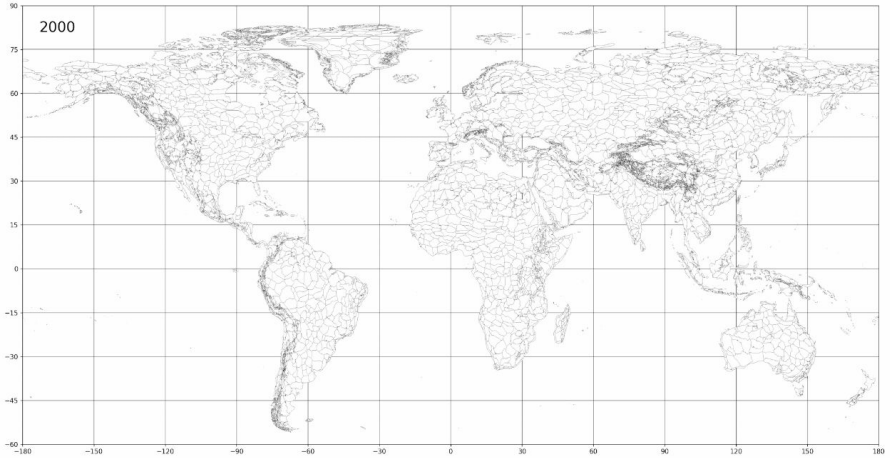
Absolute Cor-Cor



Results | Longitudinal Perspectives | Yearly

Gif doesn't render on downloaded pdf. Please see:

https://github.com/mikacashman/PASC23_Climatypes_SupResources



Results | Overlay w/ Species Distributions

- Global species occurrence data combined with environmental data from the climatic clusters
- Identify what climatic regions (features) contribute to thriving species
- Identify similar regions not currently utilized

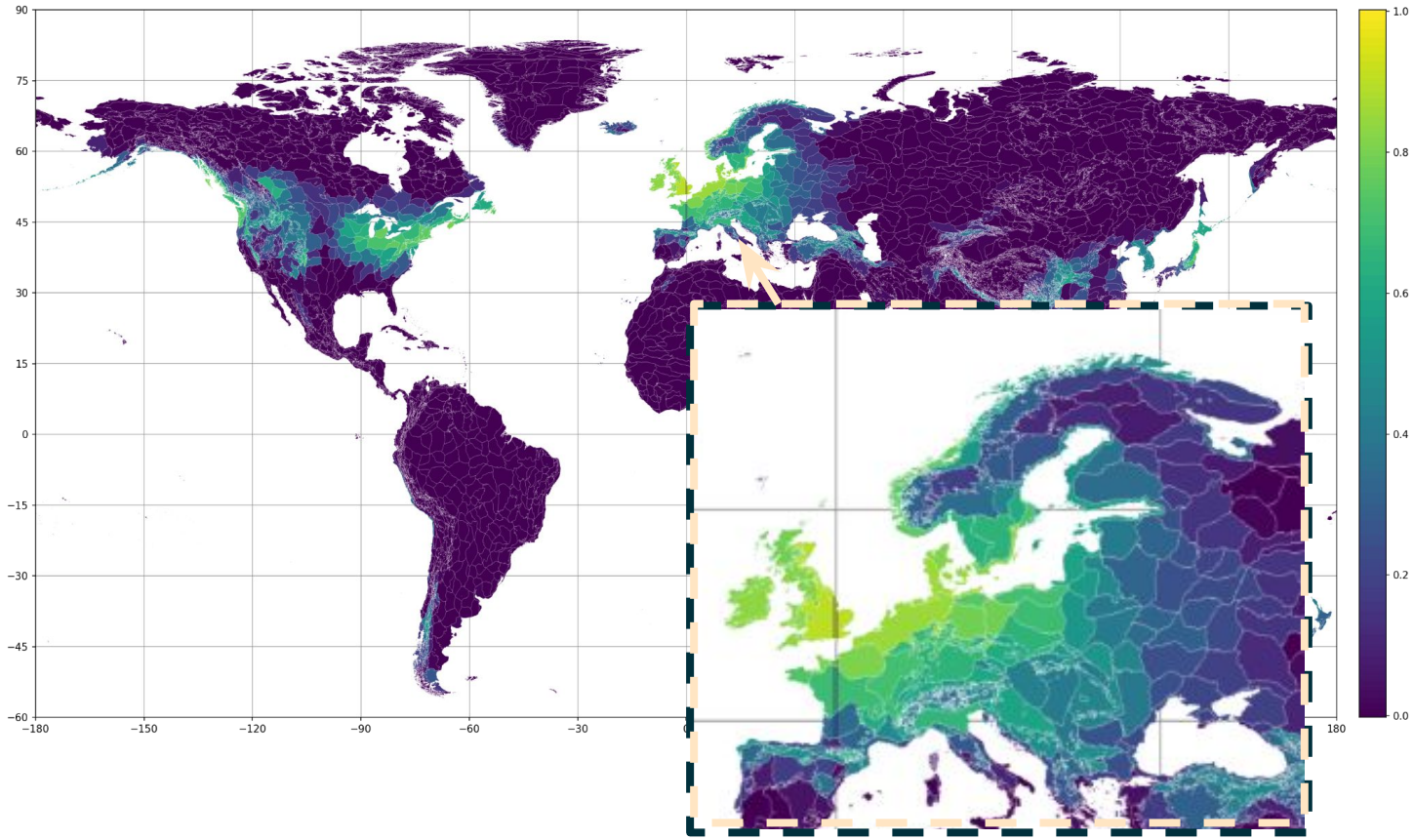
- Species distributions across three biological applications:

bioenergy

agriculture

zoonotic spillover

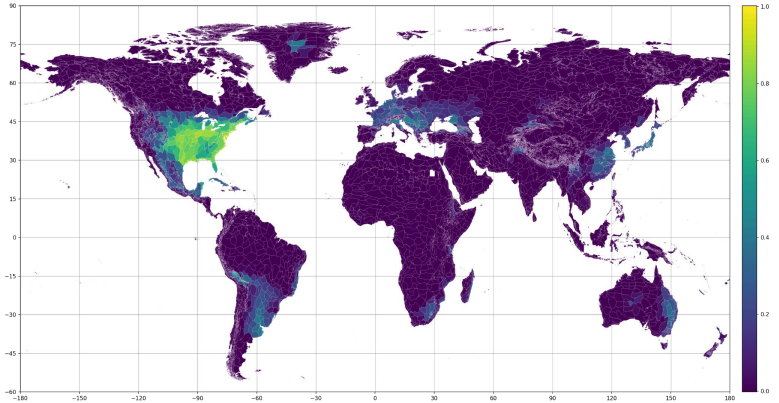
Results | Overlay w/ Species Distributions | Bioenergy

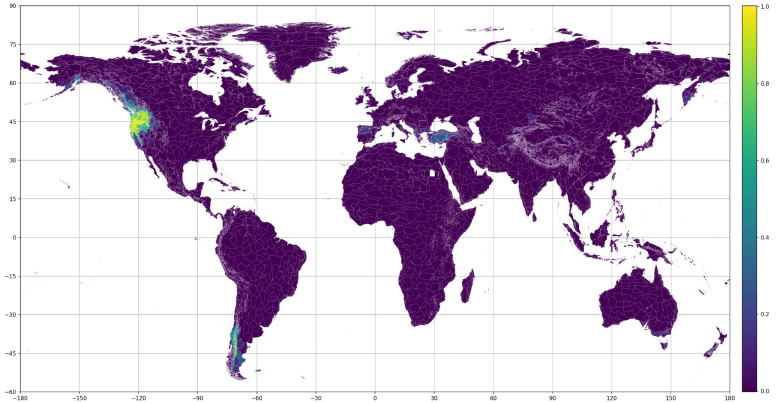


Pennycress (*Thlaspi arvense*)

- Bioenergy feedstock for sustainable aviation fuel
- As an emerging cover crop can help assess profile of row crop agriculture

Switchgrass (*Panicum virgatum*)

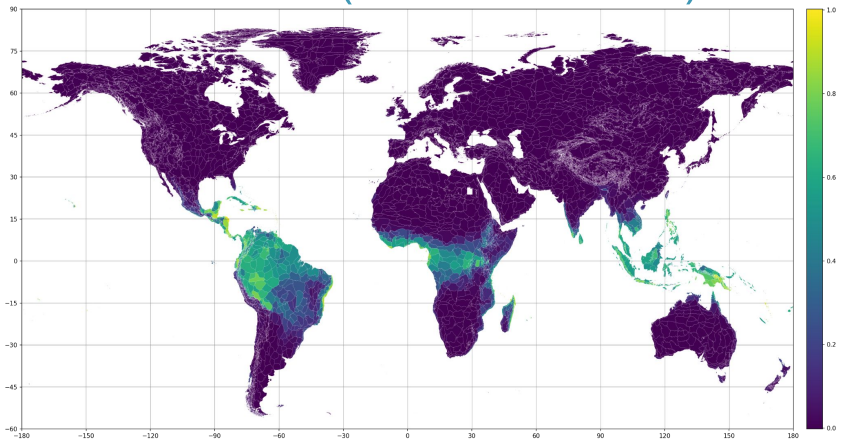


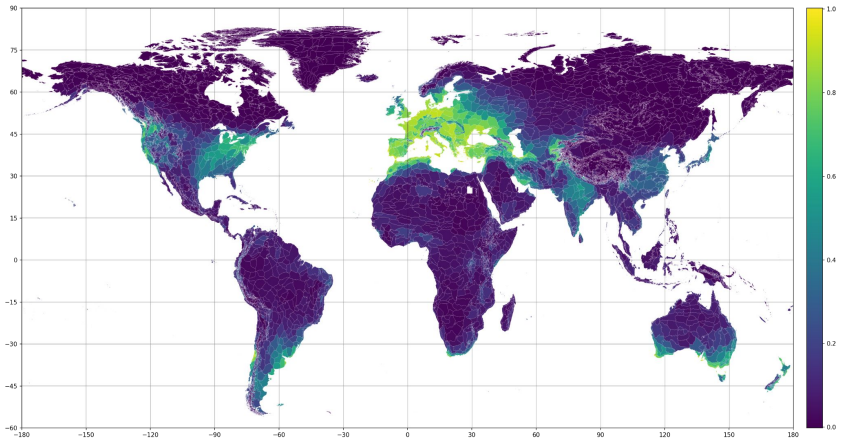


Poplar (*Populus trichocarpa*)

Results | Overlay w/ Species Distributions | Agriculture

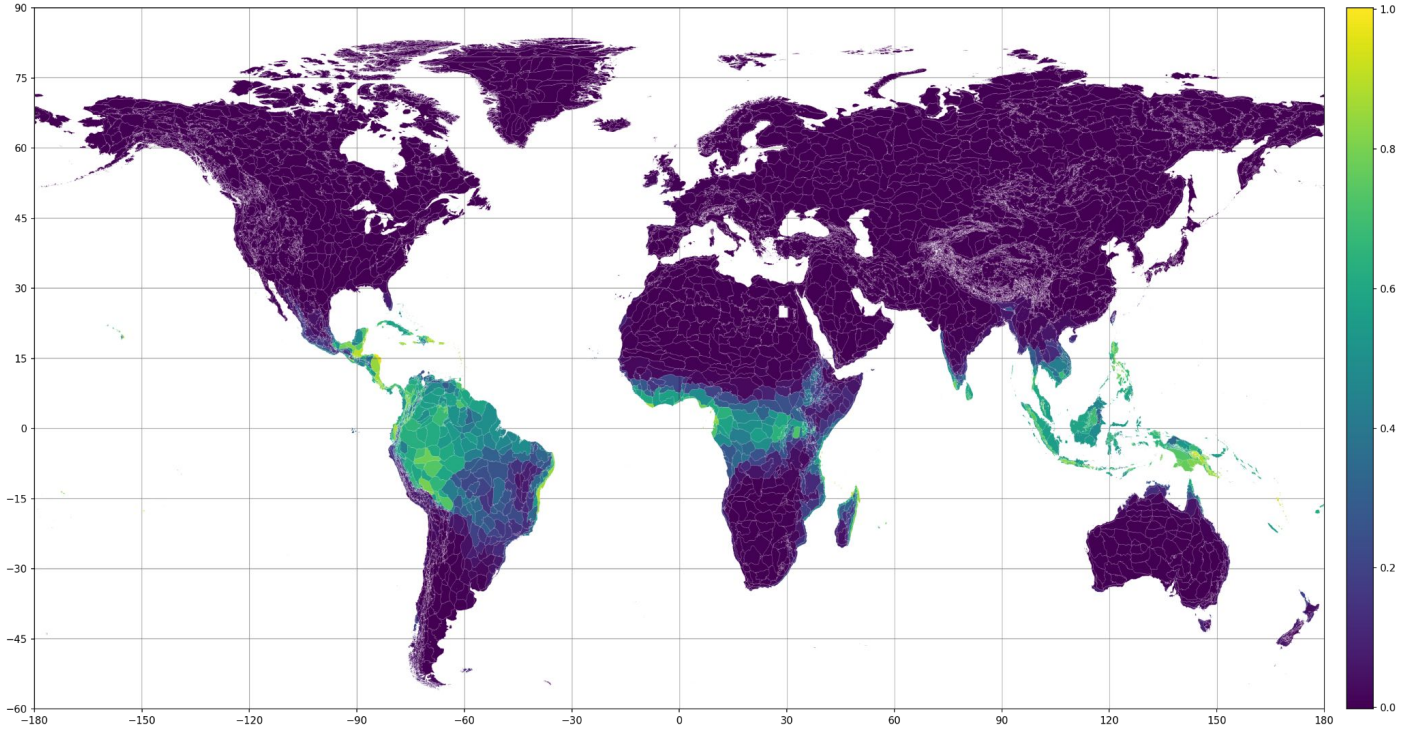
Coffee (*Coffea arabica*)





Grape Vine (*Vitis vinifera*)

- Applications
  - identify related zones not currently utilized
  - predict future ranges
  - provide dynamic ranges of climate conditions for regional crop optimization

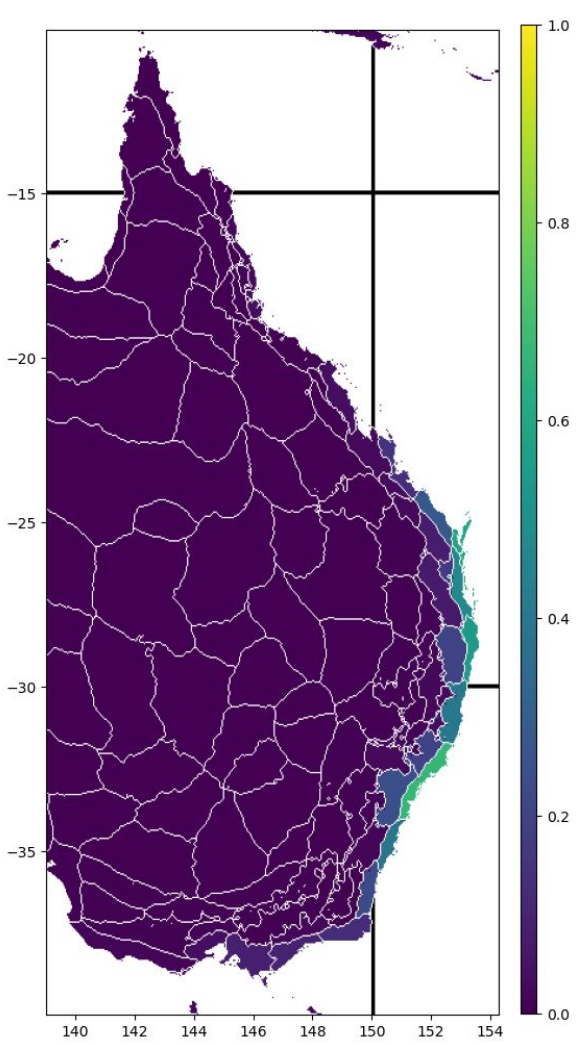


Chocolate (*Theobroma cacao*)

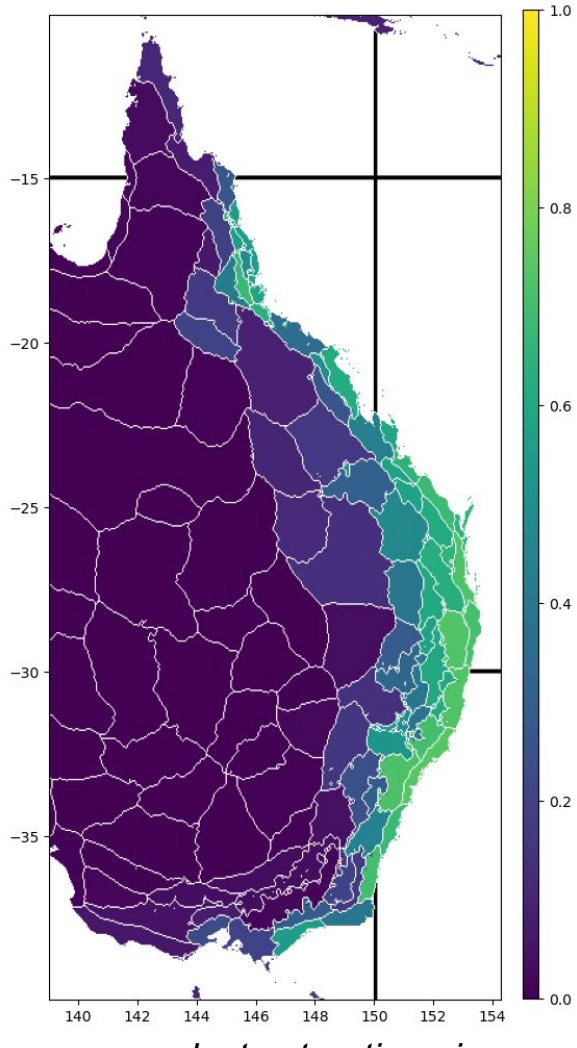
Chocolate (*Theobroma cacao*)

Results | Overlay w/ Species Distributions | Zoonosis

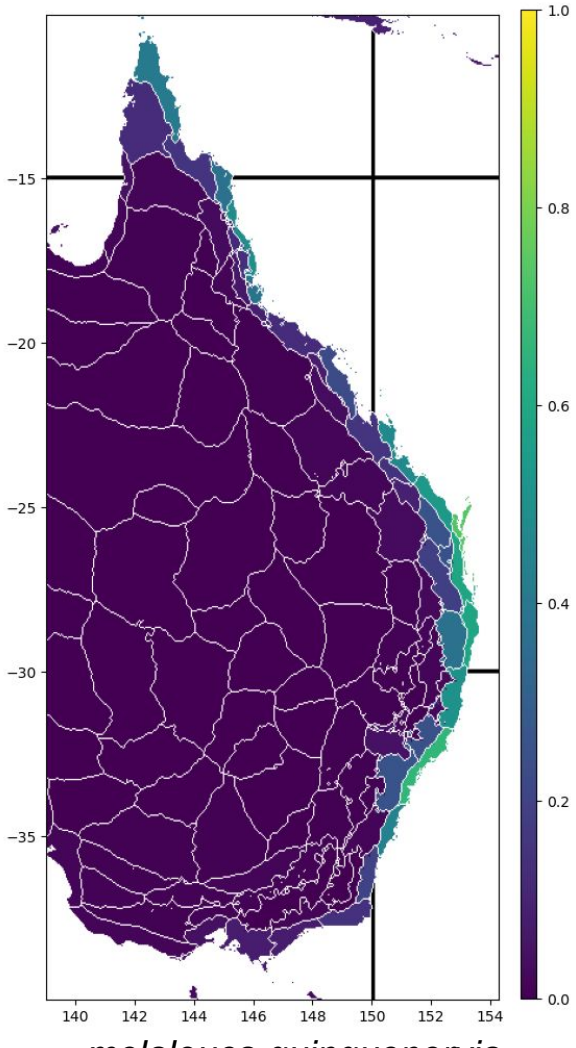
- Identify regions with pathogen reservoir species
- Food supply impacted by climatic changes
- Could result in zoonotic spillover events
- Use-Case: Potential zoonotic spillover to the Eastern Australia region and eucalyptus species



eucalyptus robusta



eucalyptus tereticornis



melaleuca quinquenervia

[1] Peggy Eby, Alison J Peel, Andrew Hoegh, Wyatt Madden, John R Giles, Peter J Hudson, and Raina K Plowright. 2023. Pathogen spillover driven by rapid changes in bat ecology. Nature 613, 7943 (2023), 340–344.

[2] Daniel J Becker, Peggy Eby, Wyatt Madden, Alison J Peel, and Raina K Plowright. 2023. Ecological conditions predict the intensity of Hendra virus excretion over space and time from bat reservoir hosts. Ecology Letters 26, 1 (2023), 23–36.

Summary

Climatic clustering and analysis over multiple geological and longitudinal perspectives

Summary

Identify changes in high-resolution zones across the globe linked by environmental similarity

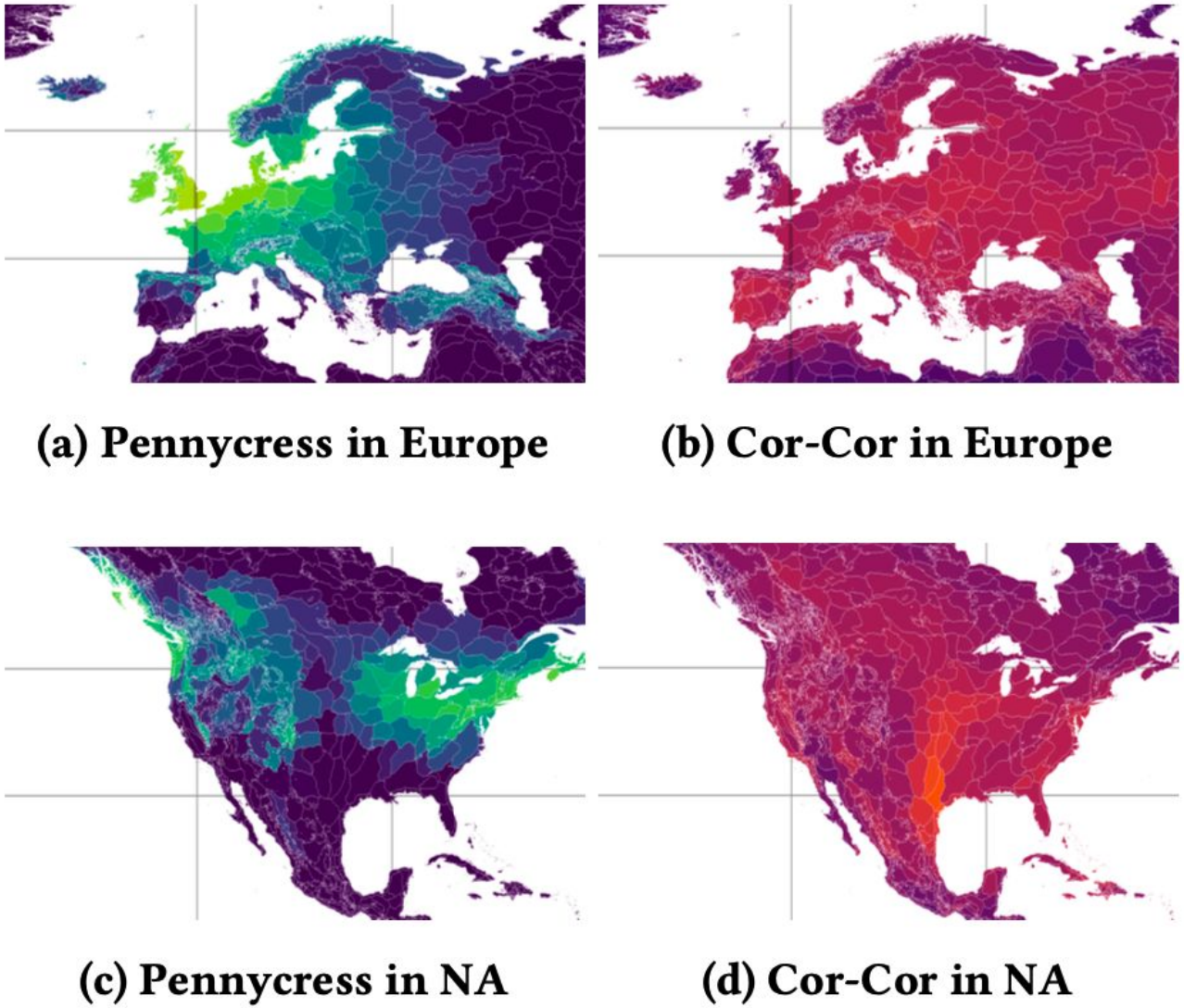
Refine exhaustive vector comparison methods & apply across 744 months of climatic data

- updated similarity metrics
- new histogram feature for resource optimization
- compare agglomerative and longitudinal views
- compare 2-way and 3-way vector comparisons
- demonstrated use on a diverse set of applications

Implications: reveal locations around the globe experiencing changes in abiotic stress

Future Work

- Combine species distributions with Cor-Cor
- Species specific analyses of indicators of abiotic stress → assess what variables are the primary drivers of environmental change
- Further analyze yearly data w.r.t. species of interest
  - e.g. identify how regional changes might correlate with successful species



BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

ACKNOWLEDGMENTS

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. Funding was also provided by the Integrated Pennycress Resilience Project (IPReP), the Center for Bioenergy Innovation (CBI), and the DOE Systems Biology Knowledgebase (KBase), all of which are supported by the Genomic Sciences Program of Office of Biological and Environmental Research in the DOE Office of Science. KBase is funded under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886. The authors would also like to acknowledge funding from the U.S. National Science Foundation (EF-2133763). This manuscript has been co-authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Mikaela Cashman[1], Verónica G. Melesse Vergara[2], John Lagergren[2], Matthew Lane[3], Jean Merlet[3], Mikaela Atkinson[3], Jared Streich[2], Christopher Bradburne[4], Raina Plowright[5], Wayne Joubert[2], Daniel Jacobson[2]

➡ **Jean Merlet is giving a talk tomorrow at 2:30** ⬅

BERKELEY LAB
Bringing Science Solutions to the World

Thank You

U.S. DEPARTMENT OF ENERGY
Office of Science

Longitudinal effects on plant species involved in agriculture and pandemic emergence undergoing changes in abiotic stress

Mikaela Cashman[1], Verónica G. Melesse Vergara[2], John Lagergren[2], Matthew Lane[3], Jean Merlet[3], Mikaela Atkinson[3], Jared Streich[2], Christopher Bradburne[4], Raina Plowright[5], Wayne Joubert[2], Daniel Jacobson[2]

OAK RIDGE National Laboratory

THE UNIVERSITY OF TENNESSEE KNOXVILLE

[1]Lawrence Berkeley National Laboratory
[2]Oak Ridge National Laboratory
[3]University of Tennessee
[4]Johns Hopkins University Applied Physics Laboratory
[5]Cornell University

JOHNS HOPKINS APPLIED PHYSICS LABORATORY

Cornell University

June 26th 2023 - PASC'23