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


Mikaela Cashman
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory
Berkeley, California, USA
mcashman@lbl.gov

Verónica G. Melesse Vergara
National Center for Computational Sciences, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
vergaravg@ornl.gov

John Lagergren
Biosciences Division, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
lagergrenjh@ornl.gov

Matthew Lane
University of Tennessee
Knoxville, Tennessee, USA
lanemj@ornl.gov

Jean Merlet
University of Tennessee
Knoxville, Tennessee, USA
merletjj@ornl.gov

Mikaela Atkinson
University of Tennessee
Knoxville, Tennessee, USA
matkin29@vols.utk.edu

Jared Streich
Biosciences Division, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
streichjc@ornl.gov

Christopher Bradburne
Johns Hopkins University Applied Physics Laboratory
Laurel, Maryland, USA
chris.bradburne@jhuapl.edu

Raina Plowright
Cornell University
Ithaca, New York, USA
rkp57@cornell.edu

Wayne Joubert
National Center for Computational Sciences, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
joubert@ornl.gov

Daniel Jacobson
Biosciences Division, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
jacobsonda@ornl.gov



## ABSTRACT

In this work we identify changes in high-resolution zones across the globe linked by environmental similarity that have implications for agriculture, bioenergy, and zoonosis. We refine exhaustive vector comparison methods with improved similarity metrics as well as provide multiple methods of amalgamation across 744 months of climatic data. The results of the vector comparison are captured as networks which are analyzed using static and longitudinal comparison methods to reveal locations around the globe experiencing dramatic changes in abiotic stress. Specifically we (i) incorporate updated similarity scores and provide a comparison between similarity metrics, (ii) implement a new feature for resource optimization, (iii) compare an agglomerative view to a longitudinal view, (iv) compare across 2-way and 3-way vector comparisons, (v) implement a new form of analysis, and (vi) demonstrate biological applications and discuss implications across a diverse set of species distributions by detecting changes that affect their habitats. Species of interest are related to agriculture (e.g., coffee, wine, chocolate), bioenergy (e.g., poplar, switchgrass, pennycress), as well as those living in zones of concern for zoonotic spillover that may lead to pandemics (e.g., eucalyptus, flying foxes).



## CCS CONCEPTS

• **Applied computing** → **Environmental sciences**; • **Computing methodologies** → **Parallel algorithms**.

## KEYWORDS

high performance computing, climate analysis

**ACM Reference Format:**
Mikaela Cashman, Verónica G. Melesse Vergara, John Lagergren, Matthew Lane, Jean Merlet, Mikaela Atkinson, Jared Streich, Christopher Bradburne, Raina Plowright, Wayne Joubert, and Daniel Jacobson. 2023. Longitudinal effects on plant species involved in agriculture and pandemic emergence undergoing changes in abiotic stress. In *Platform for Advanced Scientific Computing Conference (PASC '23), June 26–28, 2023, Davos, Switzerland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3592979.3593402


## 1 INTRODUCTION

Plant and animal species are under selective pressure to evolve and adapt to their natural habitats in order to maximize their chances of reproduction, which can lead to a dependence on particular environmental and climatic trends that constrain their geospatial distributions. However, land-use change and a rapidly changing climate are putting unprecedented pressure on such species, with


Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.



*PASC '23, June 26–28, 2023, Davos, Switzerland*



© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.



ACM ISBN 979-8-4007-0190-0/23/06…$15.00
https://doi.org/10.1145/3592979.3593402

potentially devastating consequences across agriculture, bioenergy, and the potential for pandemics. It is therefore important to identify geospatial zones linked by environmental similarity that are suitable for particular species of importance as well as to quantify the amount of change each area is experiencing. Such analyses not only identify potentially suitable areas not currently in use, but also detect the level of abiotic stress that relevant species may be experiencing in each habitat as the environment is changing.

One of the earliest methodologies of climatic clustering is the Köppen-Geiger method [18]. While an influential method, it relies heuristic decision rules based on only two source variables (temperature and precipitation) leading to low-precision of the resulting clusters. More recent related work present unsupervised classification methodologies as an improvement to Köppen-Geiger such as using k-means and PCA, or hierarchical clustering [21, 24, 30, 31]. In contrast to related work, the methodology used here incorporates 14 different climatic variables to leverage a richer feature space across a dense global coordinate grid. Furthermore, our methods are exact in nature by performing exhaustive all-to-all comparisons compared to other related unsupervised methods whose effectiveness can rely on parameter optimization. Finally, we provide longitudinal views in addition to static views of clustering.

In this work, we leverage and enhance a high-performance computing methodology to detect global regions correlated by environmental features from longitudinal and agglomerative perspectives [19]. We refine a number of technical aspects of the existing approach, resulting in improvements that increase the computational efficiency, remove bias against extreme climates, and measure levels of abiotic stress. Where previous methods considered up to 500,000 geolocations, our work is applied to more than 8.8 million points of dry land in a uniform grid across the globe. We demonstrate the applicability of the resulting similarity networks to species distribution models in agriculture (e.g., coffee, wine, chocolate), bioenergy (e.g., poplar, switchgrass, pennycress), as well as those living in zones of concern for zoonotic spillover that may lead to pandemics (e.g., eucalyptus, flying foxes).

In the following sections we first present background information on the vector similarity tool, CoMet, in Section 2. Then, in Section 3, we present our methodology, highlighting the advancements we have made to the workflow. In Section 4, we present results (i) comparing methods of measuring similarity, (ii) comparing an agglomerative perspective over 62 years versus yearly windows in a longitudinal format, (iii) comparing a 2-way versus 3-way perspective localized in Eastern Australia, and (iv) on biologically relevant species distributions. We conclude with a discussion in Section 5.

## 2 BACKGROUND

### 2.1 CoMet

In order to exhaustively compute similarity metrics at-scale, we leverage the **Co**mbinatorial **Met**rics (CoMet) library [14, 15] on high-performance computing (HPC) systems including the Oak Ridge Leadership Computing Facility (OLCF)'s Summit supercomputer. The CoMet library is a data analytics application that performs ultra-low precision mathematics by converting raw feature data to binary, thereby enabling 1-bit general matrix-matrix multiplications. CoMet enables efficient vector similarity metric computations and has demonstrated up to 98% weak scaling efficiency on leadership-class systems such as Summit. CoMet was also the first to exceed the ExaOp/s barrier ($10^{18}$ operations per second) with a production application using mixed precision calculations [16].

As a measure of similarity between vectors, we use the Duo metric within CoMet [8]. The binary formatting of feature information translates into a High (1) feature value or a Low (0) feature value. In a 2-way comparison, this results in four possible categories of relationships: High-High (1, 1), High-Low (1, 0), Low-High (0, 1), and Low-Low (0, 0). We shorten these to HH, HL, LH, and LL moving forward. The Duo metric between two vectors $i$ and $j$ is defined as:

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

where $r$ is the given correlation relationship, $q$ is a scaling factor, and $D_{i,j}$ is the proportion of vectors with the given relationship $r$:

$$D_{i,j}(r) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{i_n=r_1, j_n=r_2\}}. \tag{2}$$

where $\mathbb{1}$ represents the indicator function which returns 1 when the condition is met (vector $i$ at position $n$ is equal to $r_1$ and vector $j$ at position $n$ is equal to $r_2$), and 0 otherwise. The summation in $D_{i,j}(r)$ represents the total number of positions between the two vectors $i$ and $j$ in which the relation $r$ is found. This summation is then divided by the total vector length to result in a proportion. The Duo metric also includes two frequency terms, $f_i$ and $f_j$, that account for the frequency of each value character {0,1} within each input vector. We will refer to these terms again in Section 3. The frequency terms are defined as:

$$f_i(r) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{i_n=r_1\}}, \tag{3a}$$

$$f_j(r) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{j_n=r_2\}}, \tag{3b}$$

where $N$ is the number of vectors, and $r = [r_1, r_2]$ is the correlation relationship. Note here that the indicator function $\mathbb{1}$ only checks for the corresponding relation value of interest ($r$) in the vector of interest ($i$ in $f_i$ and $j$ in $f_j$), returning 1 if they are equal and 0 otherwise. The HH and LL metrics represent positive correlations, and the HL and LH metrics represent anti-correlations. As our application is interested in defining correlated regions of geolocations, our derived correlation metric is the sum of positive correlations, HH+LL, which has not previously been considered.

The same Duo metric can be extended for use in 3-way vector comparisons (i.e., comparing all unique vector triplets instead of unique vector pairs). This extension results in eight possible categories of relationships: HHH, HHL, LHH, HLH, HLL, LLH, LHL, and LLL where the final correlation metric of interest in this work is the sum HHH+LLL, similarly a new consideration. We perform and compare both 2-way and 3-way Duo calculations in this work.

# 3 METHODS

Related work by Lagergren et al. [19] presents a methodology to utilize CoMet for climatic clustering. We follow a similar core methodology, however, we introduce several novel modifications. Next, we outline our methodology while highlighting our algorithmic refinements. Specific modifications we present are: (i) updating the vector similarity (correlation) metric to eliminate bias towards extreme conditions, (ii) implementing a `histogram` method within CoMet to eliminate the need for "test" runs to determine correlation threshold, and (iii) presenting a new analysis method highlighting absolute change among vectors in comparison to relativistic in related work. We detail each of these contributions in the following sections.

## 3.1 Vector Generation

CoMet utilizes ultra-low precision matrix-multiplications in order to scale to high-throughput functionality. To leverage such capability, the continuous-valued vector inputs are converted into a binary vector format. In this work, each vector represents one geolocation (i.e., point of dry land), and the elements comprising each vector are a representation of environmental features at that location.

To obtain these features, we extracted source data from the TerraClimate database [3]. This resource contains monthly observation data from January 1958 to December 2019, resulting in a total of 744 months of data across 14 environmental features: maximum temperature, minimum temperature, vapor pressure, precipitation accumulation, downward surface shortwave radiation, wind-speed, reference evapotranspiration (ASCE Penman-Montieth), runoff, actual evapotranspiration, climate water deficit, soil moisture, snow water equivalent, palmer drought severity index, and vapor pressure deficit. Following [19], to reduce correlations between climatic features, we replace maximum temperature with temperature range (i.e., trange = tmax − tmin), which makes use of the same information but reduces the correlation between minimum and maximum temperature.

Raw data from TerraClimate are in integer and float formats. For maximum computational efficiency within CoMet, these are converted into a binary representation. For each climatic variable, we use a uniform quantile transformation (i.e., a transformation from continuous to categorical distributions) to convert the continuous-valued climate data into $n$ categorical bins such that each bin contains the approximately same number of instances. This ensures extreme outliers are not disproportionately scaled in the vector comparison. Next, each bin is given a binary assignment of size $m = n-1$ such that for bin number $k$, the assignment is $0^{m-k+1}1^{k-1}$. For example, for $m = 5$, the assignment to bin 3 corresponds to the binary vector 00011. Since CoMet performs bit-wise comparisons, this ensures that correlations between bins are scaled appropriately. In this work, we choose $m = 50$ for optimal HPC performance.

To define the vectors, we uniformly sample geolocations from TerraClimate by taking equally spaced longitude and latitude steps and remove any ocean and Antarctic coordinates, resulting in a set of dry-land locations. We consider two geographic perspectives in this work: (i) a global view with a total of 8,834,910 dry-land geolocations, and (ii) a regional view of Eastern Australia with a total of 153,149 dry-land geolocations. The global perspective provides an order of magnitude finer resolution compared to the 500,517 geolocations in related work [19]. We utilize the Eastern Australia perspective as a comparison of 2-way and 3-way metrics in an area of high concern for zoonotic spillover events for Hendra virus, which has a 75% fatality rate in horses and a 57% fatality rate in humans [11, 20].

In addition to static views of historic climate patterns, we also provide several longitudinal perspectives. We first take a monthly mean perspective, in which all 62 years of data for each month are averaged into a single feature value. As we have 12 months, 14 feature variables, and a binary assignment size of 50 for each feature, this results in vectors of length $12 \times 14 \times 50 = 8400$ for each geolocation. We also take a yearly perspective, where each comparison contains all 12 months over a single year, and repeat this for all 62 years of available data. Similarly, this results in 62 vectors each of length 8400. The first approach provides an agglomerative view across all 62 years of data, while the second provides a longitudinal view at yearly resolution.

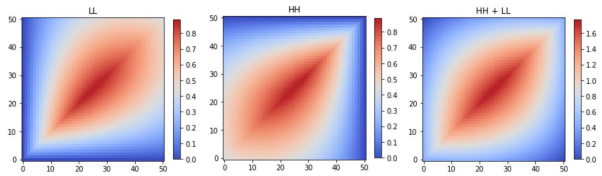
## 3.2 Correlation Metric

The Duo metric in Equation 2 contains two terms for capturing frequency effects: $f_i$ and $f_j$. These terms are beneficial for penalizing extremely rare cases in genomics studies, where cases of rare alleles can skew the resulting correlations. However, in our application, the effect of these terms causes extreme climate conditions (e.g., near the North Pole) to be filtered out, which is an undesirable consequence. To counteract this, in this work we present a modification to the Duo metric by setting the scaling factor $1/q = 0$, effectively canceling out the frequency terms. This modification of the scaling factor, $q$, results in the Duo metric being mathematically equivalent to the Sørensen–Dice coefficient, a set similarity metric that is not biased by the frequency of elements in a given vector [10, 26].

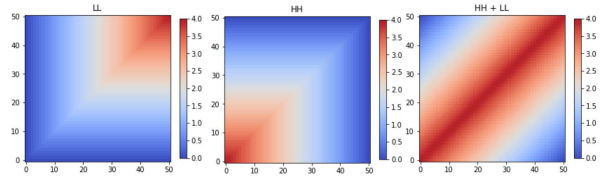
The effect of this modification is visualized in Figure 1. In particular, Figure 1a compares the Duo scores between the different correlation relationships with the default scaling factor $\frac{1}{q} = \frac{3}{2}$. If we observe the lower-left corner of the LL plot (left) and the upper-right corner of the HH plot (middle), we see that the resulting Duo correlation is lower, despite the large amount of agreement between the vectors being compared. In the combined plot HH+LL (right), we can observe that if a feature vector has a majority High (1) values or majority Low (0) values, then the correlation is lower compared to a feature vector with an equal number of High and Low values (middle of the plot). However, if we modify the scaling factor (Figure 1b), we observe that the correlations for HH+LL (right) are equally high regardless of the frequency of High or Low terms. Thus in this work we both implement the modified Duo metric and capture HH+LL relationships.

## 3.3 Correlation Thresholding

Writing every relationship between elements to disk is not practical for large scale applications. If we stored all relationships in our 8.8M vector application, the output would require more than 4PB of storage in binary form, and more than 37PB in plain text. In practice, a threshold is defined in which all correlations larger than the set threshold are stored. An ideal CoMet run is sizable enough to capture rich information (e.g. as many edges as possible to create




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




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

**Figure 1: Uniform score distribution of the Duo metric with varying scaling parameter, $q$. The horizontal and vertical axes represent the binary assignment of two vectors, $i$ and $j$, where the ticks represent the count of 1's in our binary encoding scheme. Thus, vectors $i$ and $j$ are exactly equal along the diagonal, and differ off of the diagonal. The origin represents vectors of all zeros while the upper right corner represents vectors of all ones. The color map represents the strength of the Duo score, ranging from low correlation in blue to a high correlation in red.**

the largest network) while avoiding an edge count that is too large for clustering (a limitation of clustering tools). Thresholding is performed at the tails of a distribution of similarity scores (edges). As it is often the case that small adjustments in a chosen CoMet threshold result in significant changes in the edge count, it can be non-trivial to identify an ideal threshold.

An additional contribution of this work is the implementation of new CoMet functionality that generates a distribution of all correlation metric values prior to generating output files, the writing of which can consume a significant amount of both compute (thousands of node hours) and storage (hundreds of terabytes) resources. By selecting the `histogram` method, CoMet will run the specified metric and store the distribution of scores of each relation type into user-defined bins without storing each resulting score. This allows users to observe the distribution and better inform selection of a similarity threshold value to use when filtering which metrics to store. Using the histogram feature, a user can know *a priori*, how many metrics will be saved for a particular threshold and relation type. This new methodology saves computational resources by eliminating the need to execute scaled "trial" runs with metric output enabled and varying thresholds. It also prevents the need to re-run CoMet if an incorrect threshold was chosen.

As an example of the implication of this new capability, consider the global scale run in this work containing 8.8M vectors. One run of CoMet utilizes 684 node hours. Raw output from CoMet is in binary form and must be post-processed back into plain text. That utilizes another 662 node hours for a total of 1,346 node hours. Using the previous CoMet workflow, if the incorrect threshold was chosen, a full CoMet run would have to be re-executed resulting in doubling the node hour usage to more than 2,692 node hours. It is possible this would be to be repeated additional times until the ideal sized output was confirmed. Smaller scaling runs may also be used, but similarly if the target output size is not achieved these runs will have to be repeated. With the new histogram method, CoMet only needs to be run once with minimal I/O using 252 node hours. With the resulting output, the user can know the exact threshold needed to fit any target output desired. If this target changes in the future, the same output can be used to select a new threshold. The storage needs of a CoMet run depend on the target output size. In this example the output was 888GB in binary and 7.6TB in plain text. Thus for any repeated run this storage could be roughly doubled for every run executed. In contrast, the histogram output is 48K of plain text.

Figure 2 illustrates the resulting histogram from a 2-way global-scale CoMet run. Using the score distribution, we observed that a threshold of 0.964 will result in storing approximately 0.1% of all metrics, which is equivalent to 36,712,590,809 edges that are then used for downstream network analysis. We provide these additions to CoMet at https://github.com/wdj/comet.

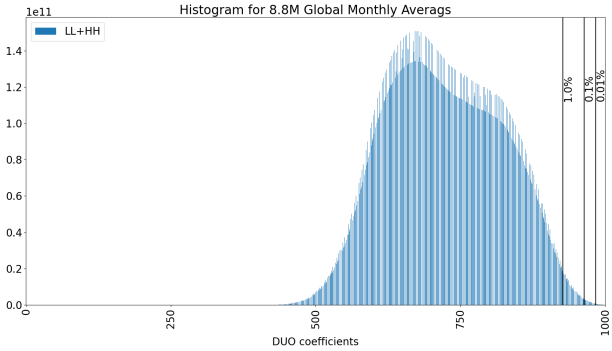



**Figure 2: Example histogram for the 8.8M Global Monthly Mean Dataset. The vertical axis represents the number of edges while the horizontal axis represents the correlation metric. The black vertical lines, going from left to right, represent thresholds that would yield 1%, 0.1%, and 0.01% of the total edges, respectively.**

## 3.4 Vector Comparisons in CoMet

Once a threshold is selected, we use CoMet to conduct an exhaustive vector comparison and generate vector pairs (or triplets in the case of 3-way) that have a similarity score that exceeds the chosen threshold. Given that we are interested in the HH+LL metric as part of this work, we also implemented additional tools that merge CoMet outputs, which by default reports two metrics (HH and LL) separately for each vector pair, into vector pairs with only the combined HH+LL score.

In order to scale to the number of vector comparisons targeted in this work, we performed experiments on the Summit supercomputer, which resides within the Oak Ridge Leadership Computing Facility (OLCF). Summit is an IBM AC922 system with over 4,600 compute nodes connected via EDR InfiniBand network. Each compute node has two 22-core IBM POWER9 processors, 512GB of

memory, and six NVIDIA V100 GPUs [1]. Summit has a theoretical peak performance of approximately 200 petaFLOPS and is currently the fifth fastest supercomputer in the world [2]. To demonstrate the computational load required of the applications presented in this work, the total node hour usage for the global monthly mean perspective was 1,973 node hours as well as 15.6 TB of storage. To compute the global yearly resolution, the total usage was 113,865 node hours and 1.3 PB of storage.

### 3.5 Analysis

We perform three kinds of downstream analysis in this work: network clustering, Correlations-of-Correlations, and species distribution modeling.

To perform clustering on the resulting correlation vector pairs, a network is constructed in which nodes are geolocations, and edges represent pairs or triplets of geolocations whose similarity exceeded the chosen threshold. We then use high performance Markov Clustering (HipMCL), an HPC application for unsupervised network clustering that implements an efficient, distributed GPU-accelerated Markov clustering algorithm, to generate high-resolution clusters defining climatic zones with similar characteristics [4].

The original pipeline in related work [19] uses a method termed Correlations-of-Correlations (Cor-Cor), which utilizes a time series of global networks resulting from CoMet analysis of different time windows. Cor-Cor compares the vector in the adjacency matrix of each geolocation in a each time window to the vector of that same geolocation from the adjacency matrix in the first time window. This provides a view of relativistic change, i.e., how much the relationship of a geolocation to all other geolocations is changing over time. In this paper, we alter the Cor-Cor algorithm to compare the change in environment with respect to its own positional historical record, irrespective of other geolocations. Thus, we present an "Absolute" Cor-Cor metric as an additional view to compare with the original "Relative" interpretation. In particular, the binarization strategy described in Section 3.1 is applied to geolocations in each one-year time-window from 1958 to 2019. Then, the "Absolute" Cor-Cor computation is performed by computing the Sørensen–Dice coefficient between geolocations in different time-windows. For example, we construct a cumulative view by comparing each time window from 1959 to 2019 to the original time window from 1958 in order to measure cumulative environmental change over the time period. These values serve as a proxy for abiotic stress in plants, since it is not only important to identify zones (i.e., clusters) where species can thrive, but also to measure how much each zone is changing over time.

To measure the probability that various plant species may thrive in the identified climatic zones, we overlay our resulting clusters with several species distribution model outputs that describe potential habitable space of three bioenergy crops (pennycress, poplar, switchgrass), three species of agricultural crops (coffee, wine, and chocolate), and three pandemic-associated crop species of eucalyptus. Each species distribution model is generated using the statistical machine-learning Maximum Entropy (Maxent) model [23]. Maxent uses species occurrence data from geolocations across the globe combined with environmental data to generate a predicted probability distribution for a particular species [12]. In particular, the framework uses presence-only geolocations to extract biologically relevant climate patterns in order to build distributions in covariate space. Since the partitioned geography is constrained, so are the climate variables used to build null and predicted distributions. Random sampling of the provided geography is used to build a null distribution for probability comparisons. Through an iterative process, Maxent tests each provided environmental variables to determine their individual contribution to the final predictive model. In this work, we used standard thresholds for Maxent parameters with (i) multi-threading for accelerated time to completion, (ii) iterations set to 1000 (from 500) to increase accuracy of final models (since global models randomly sample broader climate diversity to create null distributions), (iii) 11 independent model replications (which are merged to create final models, i.e., 11,000 total iterations), (iv) "Equal Training Specificity and Sensitivity" to limit null distributions from oversaturation of sensitivity-specificity overlap, and (v) increase the random test percentage to 33%. An independent climatic data repository, Bioclim, was used as a source of environmental input data, which contains 19 climate variables of biologically relevant measures and combinations of temperature and precipitation [22]. Elevation was also included as an additional input feature. Sample locations were sourced from the global species archive, Global Biodiversity Information Facility (GBIF). Individual lists of coordinates were then filtered to be no closer than 10km from each other to correct for dosage effects to over-observed geography.

In each of the Cor-Cor and species distribution results, we aggregate the corresponding values within each climatic cluster so that each cluster represents an average view of the described phenomena.

## 4 RESULTS

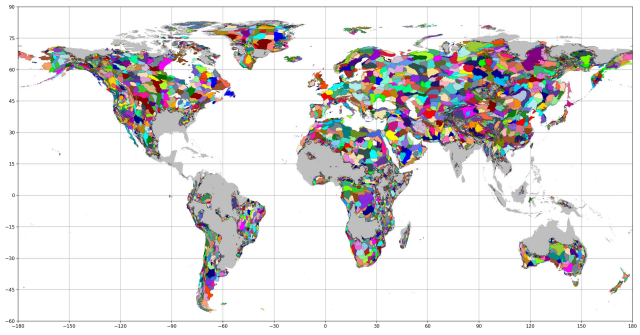
### 4.1 Duo vs Sørensen-Dice

In Section 3, we highlighted that one of the major refinements in this work is the removal of the frequency terms in Equations 3a and 3b to eliminate bias against extreme environmental conditions. To demonstrate its practical effect, we present a comparison of the original Duo metric which includes all types of relationships to the modified Sørensen-Dice metric with HH+LL. Both runs have been thresholded to retain approximately the same number of edges in the resulting network (37,522,455,884 for the Duo metric and 36,712,590,809 for the Sørensen-Dice metric).

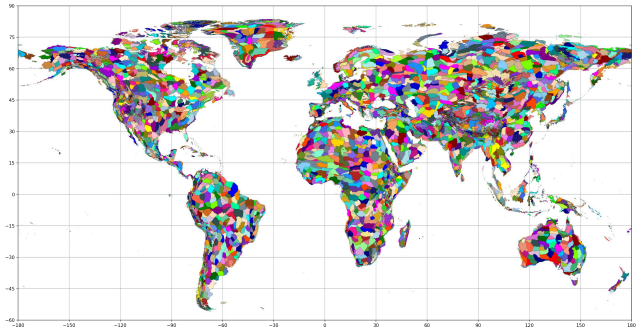
Despite the two climatic similarity networks containing approximately equal numbers of edges, the resulting clusters are dramatically different. In particular, the effects of the original Duo metric penalizing extreme climatic values (e.g., in polar and equatorial regions) can be seen in the number of missing geolocations in Figure 3a, compared to the Sørensen-Dice metric in Figure 3b, which retains nearly all of the geolocations.

### 4.2 Longitudinal Perspectives

In addition to the agglomerative view in Figure 3b, we also apply two variants of the Cor-Cor algorithm for a longitudinal perspective on the 8.8M Global Monthly Mean Dataset. The Absolute Cor-Cor analysis enables the identification of absolute environmental change over the time course. We compare this to the original Relative Cor-Cor method. Both results can be seen in Figure 4. Both



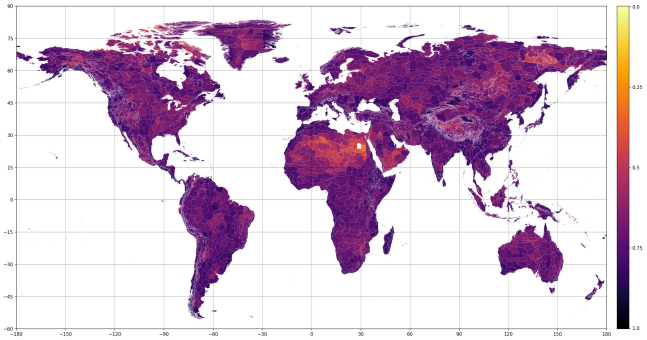
**(a) Original Duo Metric**



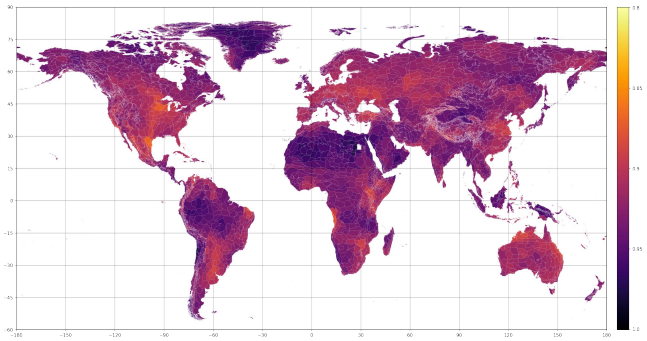
**(b) Sørensen-Dice Metric**

**Figure 3: Comparison of climatic clusters emerging from different binary vector similarity metrics applied to the 8.8M Global Monthly Mean Dataset. The clustering in (a) and (b) represent the HipMCL clusters that arise from the CoMet similarity network using the original Duo metric and the Sørensen-Dice metric, respectively. Different colors represent different climatic clusters linked by environmental similarity. Grey color indicates geolocations that were dropped due to frequency effects of the original Duo metric. Note that we are limited in the number of colors provided by our chosen visualization package (matplotlib) to 113 total colors (with greyscale and extremely bright colors removed). Thus, it is possible that distinct groups with the same color may correspond to the same climatype cluster, however, we find this is an infrequent occurrence. Therefore, in general, clusters that share the same color are distinct. We show the colored version in this paper for easier viewing, but also provide the cluster boundary maps on our supplementary site.**

Cor-Cor metrics were calculated at yearly resolution. We comment that the computational time required to compute Relative Cor-Cor is non-trivial compared to Absolute Cor-Cor. The Absolute Cor-Cor results highlight the geolocations that have experienced significant environmental changes compared to themselves over the time course, which serves as a proxy for shifts in abiotic stress in this work. In particular, this analysis reveals larger climatic shifts in central North America and Europe, southern Africa and South America, as well as East Australia. Absolute Cor-Cor is a different view of change than that considered in the Relative Cor-Cor method which is focusing on how a geolcations environmental relationships with other other geolocations are changing and is thus an indication of how climatype zones (i.e., the underlying network topologies) are changing.



**(a) Relative Correlations-of-Correlations Cumulative**



**(b) Absolute Correlations-of-Correlations Cumulative**

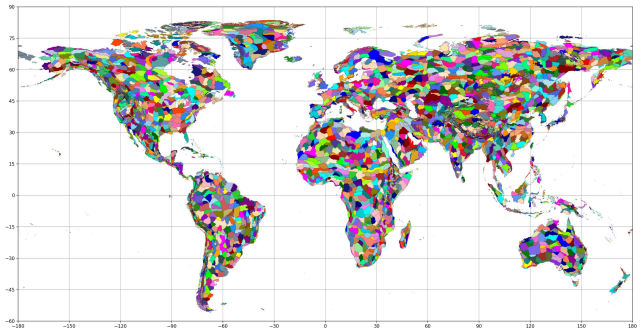
**Figure 4: Comparison of two Correlations-of-Correlations Cumulative methodologies (Relative and Absolute). In the color scheme, dark color indicates areas experiencing a small amount of environmental change while bright color indicates larger levels of change.**

We also present a longitudinal perspective of the resulting CoMet clusters over each individual one-year time-window. This perspective enables the observation of historical climatic effects and changes in climatic relationships over time. We show a sample of two years of clusters in Figure 5. All 62 years worth of clusters and boundaries as well as a video composition can be found at https://github.com/mikacashman/PASC23_Climatypes_SupResources.

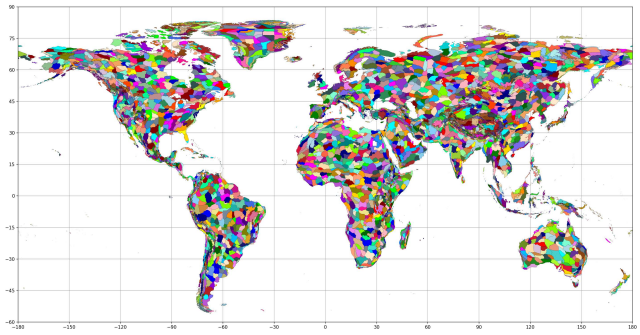
These longitudinal perspectives provide a rich analysis on how geolocations are changing over time which may contribute to increased levels of abiotic stress in the context of a changing climate.

## 4.3 2-way vs 3-way in Eastern Australia

In order to compare 2-way to 3-way vector comparisons, we restrict the region of interest to Eastern Australia for a more focused perspective. The Eastern Australia data set consists of 153,149 geolocations, each including the 14 climatic variables extracted from TerraClimate as described in Section 3. The network resulting from the 2-way vector comparison of this subset was clustered using HipMCL. The results from using an inflation value of 2 can be seen

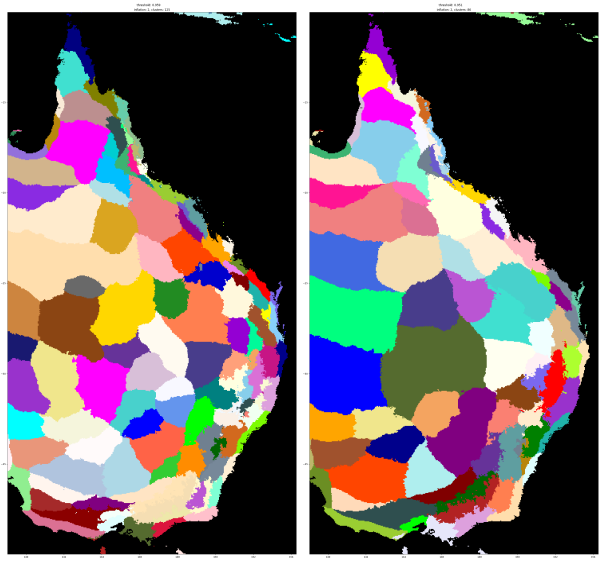


(a) 1958



(b) 2019

**Figure 5: Sample of Yearly Vector Clusters. Showing the first year (1958) and the last year (2019). Full set of clusters (colored and boundaries) can be found on our supplementary site https://github.com/mikacashman/PASC23_Climatypes_SupResources.**



(a) Cluster Results at 2-way (b) Cluster Results at 3-way

**Figure 6: Agglomerative Eastern Australia (a) 2-way with 113,686,591 edges and 125 cluster (b) 3-way with 57,151,731,825 edges and 86 clusters.**

in Figure 6a resulting in a total of 125 clusters. In contrast, the network resulting from the 3-way vector comparison using the same inflation value results in a total of 86 clusters as shown in Figure 6b. We can observe a difference in the number of total clusters and cluster sizes in different geographic regions. We further found the number of edges and nodes as well as edge density and inflation can play a role in resulting clusters, but find more analysis into each factor warrants future work. We utilize these clusters further in Section 4.4.

## 4.4 Species Distributions

In this work, we focused on species distributions in three distinct application categories to highlight a diverse set of use cases that the refined methodology described here enables. The three biological applications we focus on are: (i) bioenergy, (ii) agriculture, and (iii) zoonotic spillover.

Understanding the climate zones that are currently compatible with bioenergy deployment is important for the commercialization of viable bioenergy feedstocks as well as providing climatype targets for genome optimization and selection as well as breeding efforts that could develop bioenergy feedstock lines that will thrive in specific regions/climatype zones.

*Thlaspi arvense* (also referred to as pennycress) is an emerging covercrop and a bioenergy feedstock for sustainable aviation fuel. As such, it is compatible with, and can be used on the same land used for annual row crop food agriculture. In addition, as it helps to prevent soil erosion and nutrient runoff, it has impacts on increasing carbon sequestration, preventing nutrification of water systems and thus contributes to the overall sustainability profile of row crop agriculture. In Figure 7a, we show an overlay of the pennycress species distribution over the agglomerative global mean set of clusters. Note, this species is used across continents and finding similar cliamtypes across geography is crucial to find optimal varieties for precision breeding. These overlays are created by taking the mean over a fixed sample size of probability scores for each cluster.

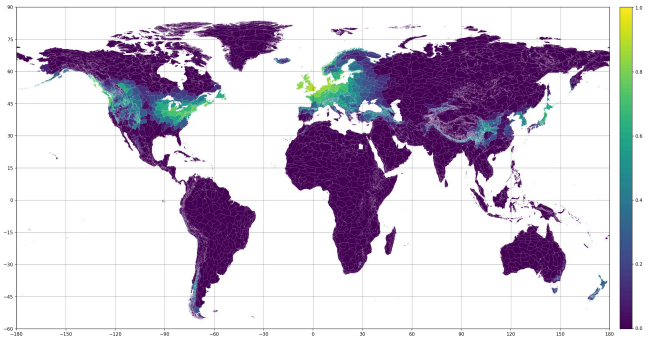
Many existing energy systems will likely rely on combustion based fuels, but the need for carbon neutral sources of combustible material will be needed to mitigate global changes to climate [9, 28]. Switchgrass is one of the leading biofuel feedstocks invested in by commercial entities as well as the United States Department of Energy to create heating fuel for homes or jet fuel for air travel. Switchgrass is a warm-season broadly adapted C4 grass species with growing regions that could fill in many current agricultural gaps in the central to central eastern North American Continent primarily in the United States [13, 17]. Switchgrass is a perennial crop plant, thus replanting events are uncommon. Switchgrass has a manageable genome size at 1.13Mb, allotetraploid composition, the genome is near fully sequenced with genome annotations of both subgenomes [6]. Substantial meta data is available and could connect omic layers to genomic features, as well as ample geographic sampling of genotypes that can be used for experimentation ranging from south east Canada to north east Mexico. Results for the

switchgrass species *Panicum virgatum* overlaid with global clusters can be seen in Figure 7b.

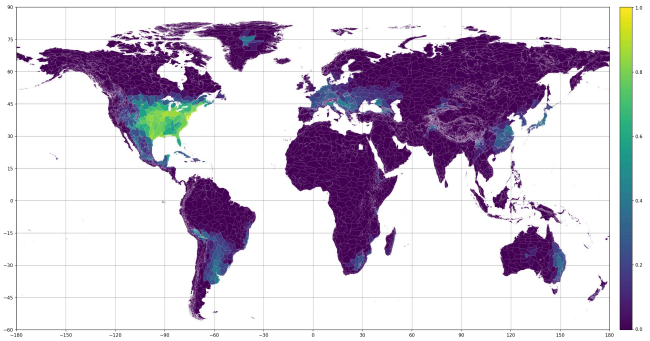
*Populus trichocarpa* (native to northwestern North America) and *Populus deltoides* (native to eastern North America), more commonly known as poplar, are fast growing high biomass yielding tree species with great potential to create bioenergy products [29]. A future bioenergy industry will need to expand its agricultural range beyond current growing regions and/or improve agricultural yield in current marginal landscapes [25]. Populus species often can be cut back to ground level for harvesting and regenerate trees from existing ground tissue and root systems, thus minimizing the need for replanting [29]. *Populus trichocarpa* is also one of the most studied model organisms for trees and has four version updates of its genome at approximately 380Mb, full genome annotation, as well as substantial amounts of biological omics data, phenotype data, and meta-data to improve breeding [7, 29]. Figure 7c displays results for *Populus trichocarpa* overlaid with global clusters.

For the agriculture of perennial crops, this climatype clustering methodology can be used to identify potential alternate climatic zones well suited for specific crops. In this work, we studied species distributions for *Coffea arabica* (coffee), *Theobroma cacao* (chocolate), and *Vitis vinifera* (grape vine) as shown in Figures 8a-8c. Each of these species origins are different than there current trans-continental range. Much of the initial legwork to grow these species interanationally is finished, but cultivar sub-varieties are not well developed. By examining yield differences of agricultural cultivars means changing crop varieties to optimal performance based on the same or similar climatypes could increase overall farm yield, aiding local farmers and improving global food securities. Overall, understanding the changes that regions targeted for both bioenergy and food agriculture are undergoing is crucial for breeding/bioengineering of cultivars that will have resilience to increasing levels of environmental variation and abiotic stress. Failure to do so will lead to increasing levels of crop failures with the associated negative impacts on economic, food, and energy security. Similarly, the ability to observe climatic trends using these methods may help to predict future ranges which could accommodate bioenergy and food agriculture, as well as provide dynamic ranges of climate conditions for regional crop optimization.

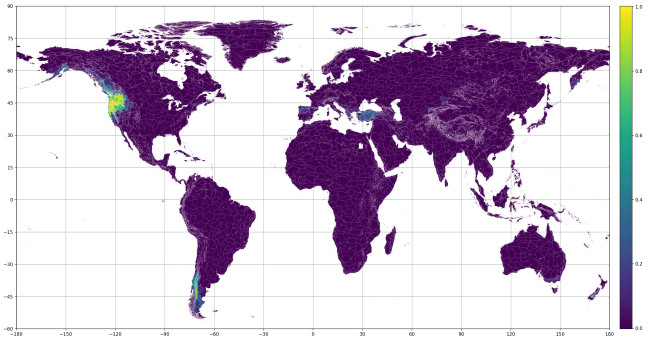
Furthermore, this climatype methodology can be used to identify regions with pathogen reservoir species whose food supply is significantly impacted by climatic changes and could result in zoonotic spillover events. In this work, we focused the study of potential zoonotic spillover to the Eastern Australia region and eucalyptus species. Here we focus on three species of eucalyptus trees (*Eucalyptus robusta, Eucalyptus tereticornis, Melaleuca quinquenervia*) which are among the main food supplies for flying foxes, and climatic changes to flying fox habitats can encourage migration to more populated areas thus increasing the risk of spillover [11]. Furthermore, the concomitant nutritional stress on flying foxes causes them to shed higher levels of Hendra virus, again increasing the likelihood of a zoonotic spillover event [5]. Figure 9 shows the three species distributions over the eastern Australia 2-way clusters. These results demonstrate that there are specific environmental regions in Eastern Australia where each Eucalytptus species tends to thrive, in which the areas of lighter green could be of higher importance for abiotic stress monitoring and modeling. These distributions can be examined in conjunction with distributions of the flying fox species, as these Eucalytptus species are linked to a primary food source for flying foxes. This could in turn stress these bats and drive potential spillover events, as described in [11]. The black flying fox is the reservoir of Hendra virus variant 1 and thus it is important for pandemic prevention to study the relationship between Eucalyptus species and flying foxes. The coastal regions are shown to have the highest distribution of flying foxes [27].



**(a) *Thlaspi arvense* (pennycress)**

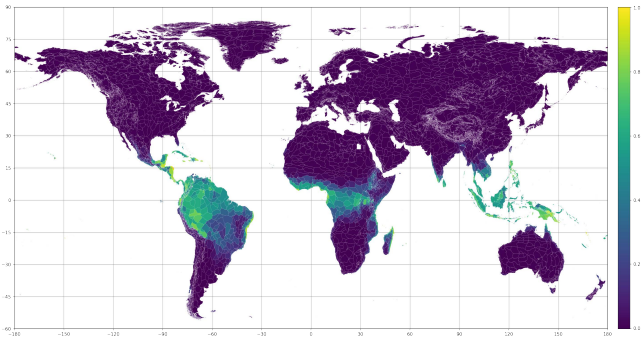


**(b) *Panicum virgatum* (switchgrass)**

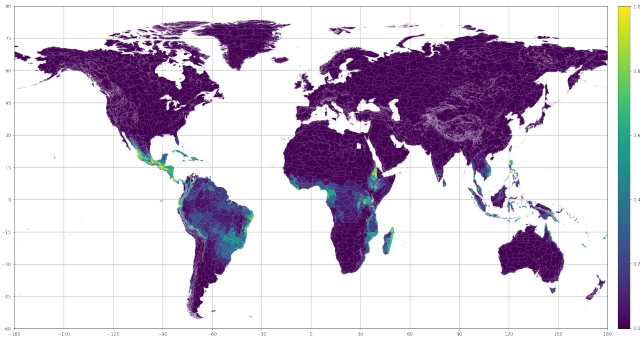


**(c) *Populus trichocarpa* (poplar)**

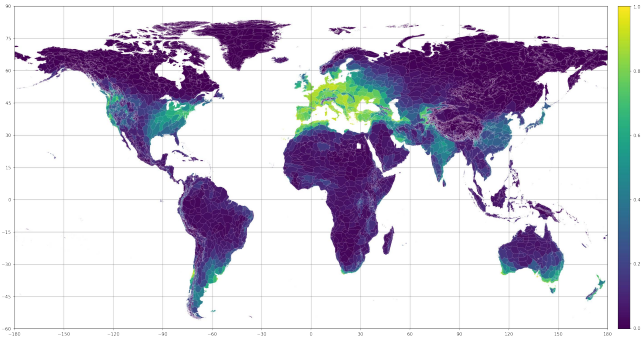
**Figure 7: Species distributions for three species related to bioenergy: (a) pennycress, (b) switchgrass, and (c) poplar. Probability models are overlaid with the resulting agglomerative CoMet clusters.**



(a) *Theobroma cacao* (Chocolate)



(b) *Coffea arabica* (Coffee)

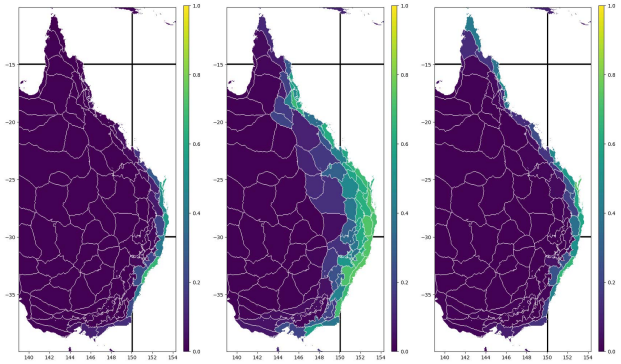


(c) *Vitis vinifera* (Grape Vine)

**Figure 8: Species distributions for three species related to agriculture: (a) chocolate, (b) coffee, and (c) grape vine. Probability models are overlaid with the resulting agglomerative CoMet clusters.**

## 5 CONCLUSIONS AND FUTURE WORK

The methodology originally described in [19] introduced a novel way to identify climatype networks using the Duo similarity metric in the CoMet software package. The refinements presented here to the correlation metric, methodology of determining an ideal threshold, and new analysis method not only allow us to save on computational time and storage resources, but further allow us to more accurately identify similarities when using climatic data by customizing the correlation metric and relationship types to match our application. These refinements were implemented with



(a) *Eucalyptus robusta* (b) *Eucalyptus tereticornis* (c) *Melaleuca quinquenervia*

**Figure 9: Species distributions for three species of Eucalyptus related to zoonotic spillover. Probability models are overlaid with the resulting 2-way CoMet clusters in eastern Australia.**

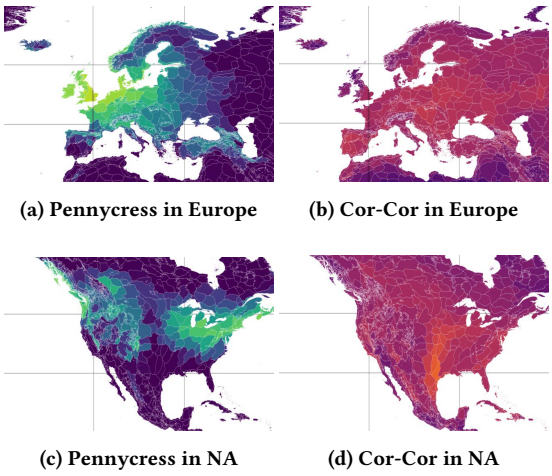
modularity so they may be applied to future applications that may require alternative metrics or relationships to be captured.

In this work we demonstrated the modified Duo metric is better suited in the climatype identification context by showcasing three distinct biological applications that can leverage the insights that emerge from climatype networks and Cor-Cor analysis presented.

For example, Figure 10 narrows in on the major regions of growth for the bioenergy crop pennycress. We compare these regions with Cor-Cor results demonstrating the risk of the effect of changing climate has on this crop. In the comparison, we can see an indication that the potential for a shift in optimal geographic growth areas for Pennycress, from Western to Eastern Europe, and from northern longitudinal to more central North American regions. Similar analyses can be done in the primary locations in each of the species distributions. In future work, species-specific analyses of these indicators of abiotic stress can be conducted to assess what variables are the primary drivers of environmental change. These indicators could then be combined with other known drivers of pandemics, such as loss and fragmentation of habitats, in order to refine pandemic risk assessment.

## 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



(a) Pennycress in Europe (b) Cor-Cor in Europe

(c) Pennycress in NA (d) Cor-Cor in NA

**Figure 10: Regional comparison of the species distribution of the bioenergy crop Pennycress to Cor-Cor.**

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).

## REFERENCES

[1] [n. d.]. Summit: Scale new heights. Discover new Solutions. https://www.olcf.ornl.gov/summit/
[2] [n. d.]. TOP500. https://top500.org/
[3] 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.
[4] 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.
[5] 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.
[6] Michael D Casler, Christian M Tobias, Shawn M Kaeppler, C Robin Buell, Zeng-Yu Wang, Peijian Cao, Jeremy Schmutz, and Pamela Ronald. 2011. The switchgrass genome: tools and strategies. *The Plant Genome* 4, 3 (2011).
[7] Hari B Chhetri, David Macaya-Sanz, David Kainer, Ajaya K Biswal, Luke M Evans, Jin-Gui Chen, Cassandra Collins, Kimberly Hunt, Sushree S Mohanty, Todd Rosenstiel, et al. 2019. Multitrait genome-wide association analysis of Populus trichocarpa identifies key polymorphisms controlling morphological and physiological traits. *New Phytologist* 223, 1 (2019), 293–309.
[8] 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).
[9] SA de Jong, ETA Hoefnagels, Joost van Stralen, HM Londo, Raphael Slade, André Faaij, HM Junginger, et al. 2017. Renewable jet fuel in the European Union: scenarios and preconditions for renewable jet fuel deployment towards 2030. (2017).
[10] Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 3 (1945), 297–302.
[11] 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.
[12] Jane Elith, Steven J Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J Yates. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and distributions* 17, 1 (2011), 43–57.
[13] Jeffrey C Hartman, Jesse B Nippert, Rebecca A Orozco, and Clint J Springer. 2011. Potential ecological impacts of switchgrass (Panicum virgatum L.) biofuel cultivation in the Central Great Plains, USA. *biomass and bioenergy* 35, 8 (2011), 3415–3421.
[14] 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. https://doi.org/10.1016/j.parco.2019.02.003
[15] Wayne Joubert, James Nance, Deborah Weighill, and Daniel Jacobson. 2018. Parallel accelerated vector similarity calculations for genomics applications. *Parallel Comput.* 75 (2018), 130–145. https://doi.org/10.1016/j.parco.2018.03.009
[16] 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* (Dallas, Texas) *(SC '18)*. IEEE Press, Piscataway, NJ, USA, Article 57, 14 pages. http://dl.acm.org/citation.cfm?id=3291656.3291732
[17] Charles Kwit and C Neal Stewart. 2012. Gene flow matters in switchgrass (Panicum virgatum L.), a potential widespread biofuel feedstock. *Ecological applications* 22, 1 (2012), 3–7.
[18] Wladimir Koppen. [n. d.]. ([n. d.]).
[19] 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). https://doi.org/10.1094/PBIOMES-02-22-0007-R
[20] Michael Letko, Stephanie N Seifert, Kevin J Olival, Raina K Plowright, and Vincent J Munster. 2020. Bat-borne virus diversity, spillover and emergence. *Nature Reviews Microbiology* 18, 8 (2020), 461–471.
[21] Marc J. Metzger, Robert G. H. Bunce, Rob H. G. Jongman, Roger Sayre, Antonio Trabucco, and Robert J. Zomer. 2013. A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography* 22 (2013), 630–638.
[22] Richard G Pearson, Jessica C Stanton, Kevin T Shoemaker, Matthew E Aiello-Lammens, Peter J Ersts, Ned Horning, Damien A Fordham, Christopher J Raxworthy, Hae Yeong Ryu, Jason McNees, et al. 2014. Life history and spatial traits predict extinction risk due to climate change. *Nature Climate Change* 4, 3 (2014), 217–221.
[23] Steven J. Phillips, Miroslav Dudík, and Robert E. Schapire. [n. d.]. Maxent software for modeling species niches and distributions (Version 3.4.1). http://biodiversityinformatics.amnh.org/open_source/maxent/. Accessed: 2022-12-20.
[24] Jean Philippe Praene, Bruno Malet-Damour, Mamy Harimisa Radanielina, Ludovic Fontaine, and Garry Riviere. 2019. GIS-based approach to identify climatic zoning: A hierarchical clustering on principal component analysis. *Building and Environment* 164 (2019), 106330.
[25] Jana Reinhardt, Pia Hilgert, and Moritz Von Cossel. 2022. Yield performance of dedicated industrial crops on low-temperature characterized marginal agricultural land in Europe–a review. *Biofuels, Bioproducts and Biorefining* 16, 2 (2022), 609–622.
[26] Thorvald Julius Sørensen. 1948. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.* I kommission hos E. Munksgaard.
[27] Libby A Timmiss, John M Martin, Nicholas J Murray, Justin A Welbergen, David Westcott, Adam McKeown, and Richard T Kingsford. 2021. Threatened but not conserved: flying-fox roosting and foraging habitat in Australia. *Australian Journal of Zoology* (2021).
[28] Rahul Tiwari, Rahul Mishra, Akansha Choubey, Sunil Kumar, AE Atabani, Irfan Anjum Badruddin, and TM Yunus Khan. 2023. Environmental and economic issues for renewable production of bio-jet fuel: A global prospective. *Fuel* 332 (2023), 125978.
[29] Gerald A Tuskan, Stephen Difazio, Stefan Jansson, Jörg Bohlmann, Igor Grigoriev, Uffe Hellsten, Nicholas Putnam, Steven Ralph, Stephane Rombauts, Asaf Salamov, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *science* 313, 5793 (2006), 1596–1604.
[30] Xianliang Zhang and Xiaodong Yan. 2014. Spatiotemporal change in geographical distribution of global climate types in the context of climate warming. *Climate Dynamics* 43 (2014), 595–605.
[31] Jakob Zscheischler, Miguel D. Mahecha, and Stefan Harmeling. 2012. Climate Classifications: the Value of Unsupervised Clustering. *Procedia Computer Science* 9 (2012), 897–906. https://doi.org/10.1016/j.procs.2012.04.096 Proceedings of the International Conference on Computational Science, ICCS 2012.

Received 23 December 2022; revised 1 February 2023; accepted 6 April 2023