Publication Figures

Connor French

Setup

Load packages

Given that I will need to recreate figures independently, rather than in the order that they are positioned in the document, I am loading data for each figure independently. So, data may get loaded redundantly. The only data I'm reading in at the beginning are the two models

Main figurers

Posterior summary figure

I have two versions that I want to run by people.

Data processing

Read in data.

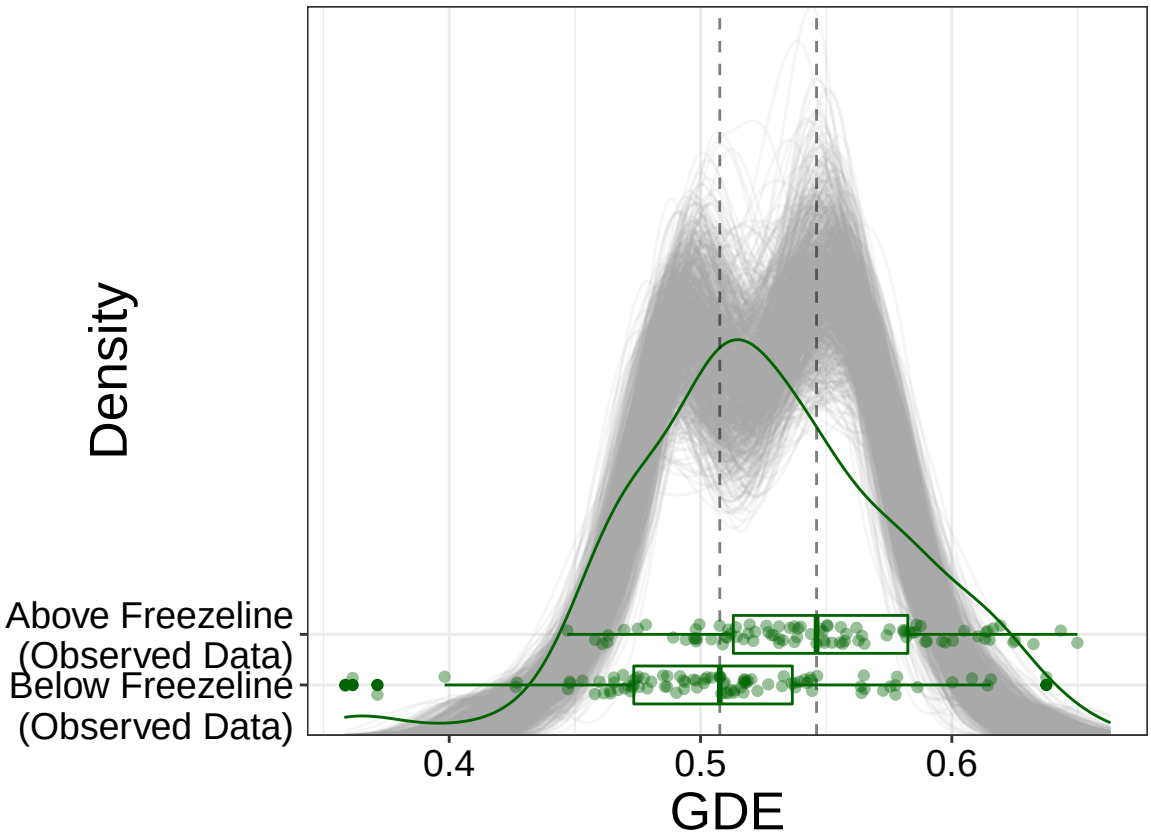
Sample posteriors for the figure. I'm retrieving the response posterior and the beta posteriors

New data frames for the GDE freeze line division and boxplot.

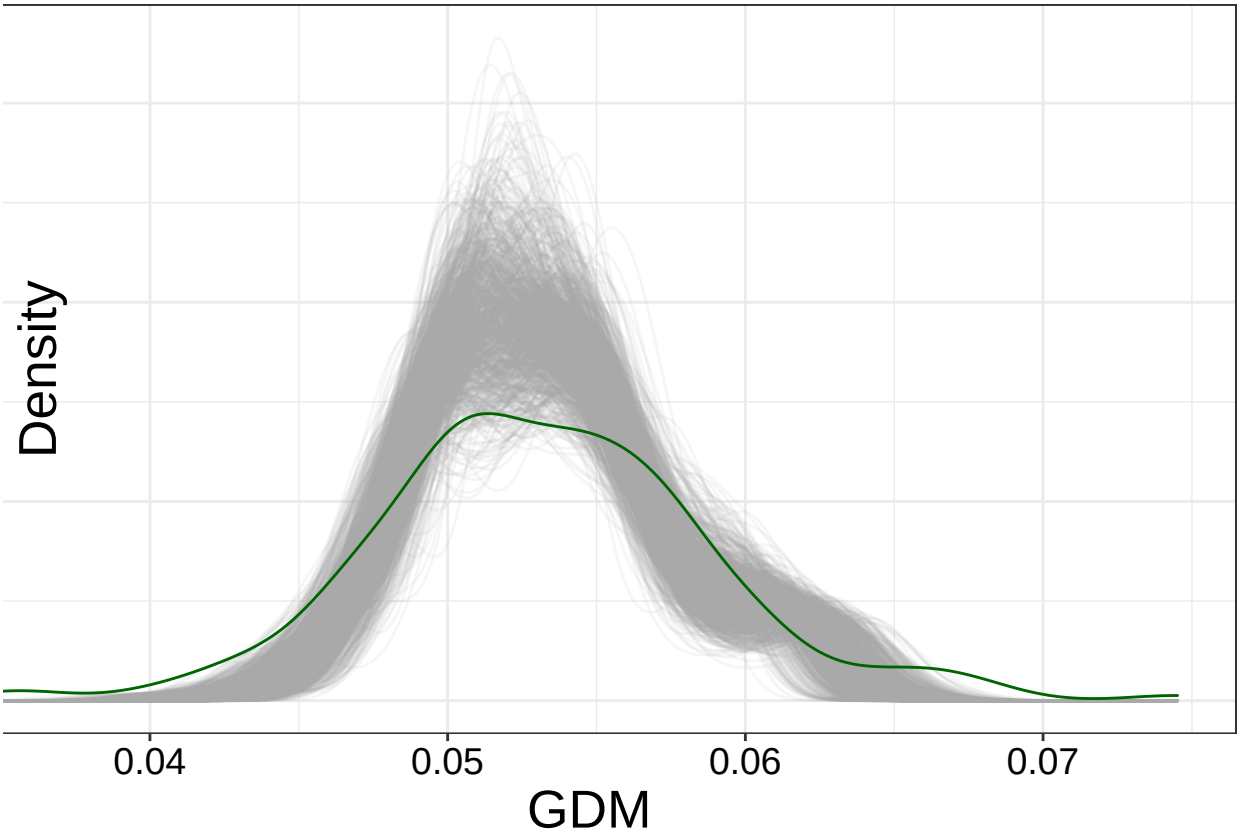
Figure

I have the panes, then the two options.

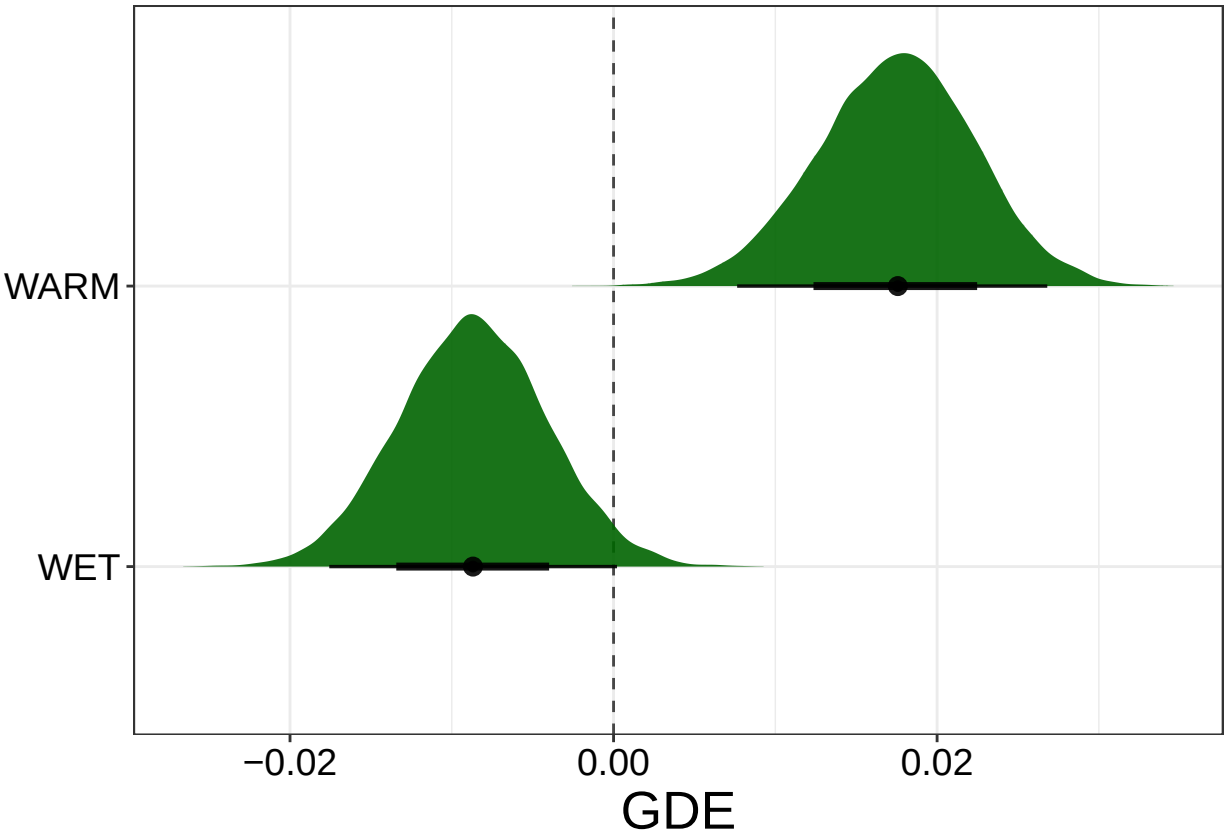
Posterior panes GDE posterior pane. Will need to move the y-axis label in inkscape.



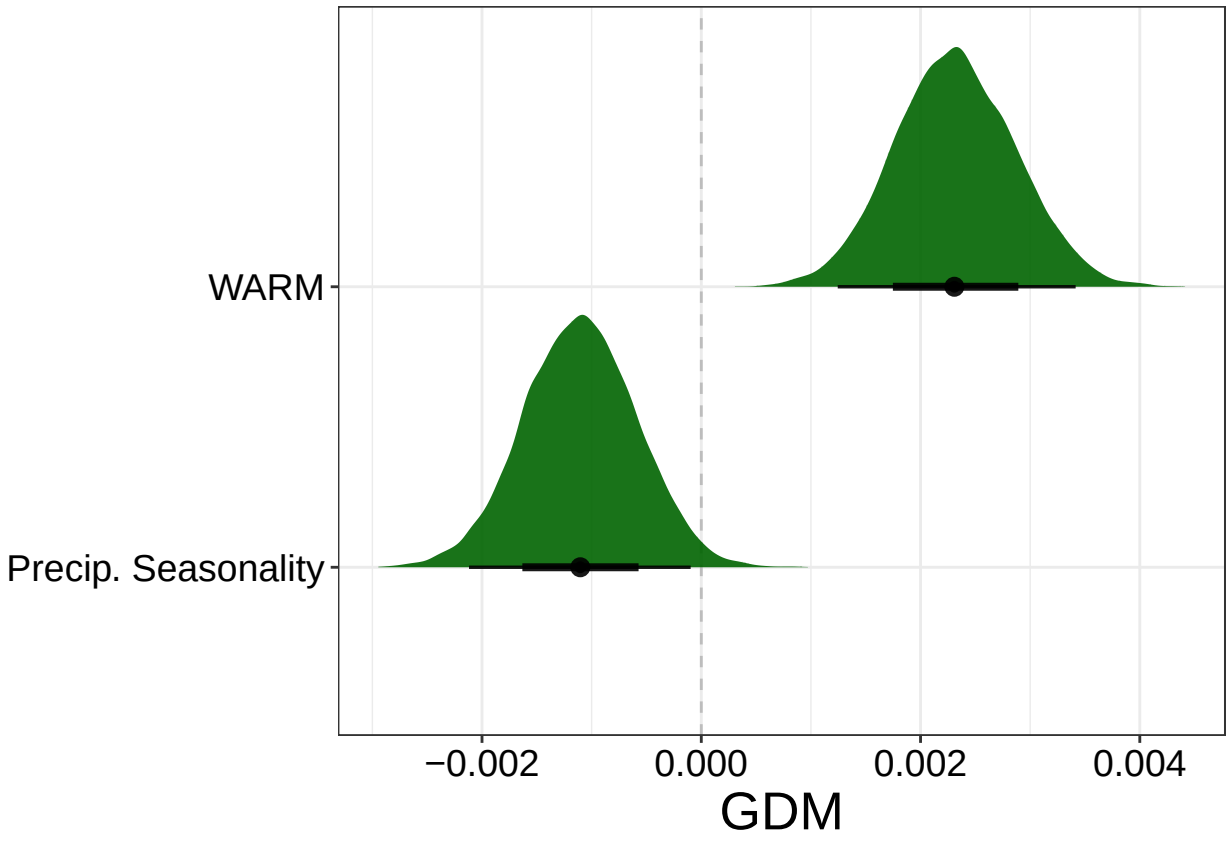
GDM posterior pane.

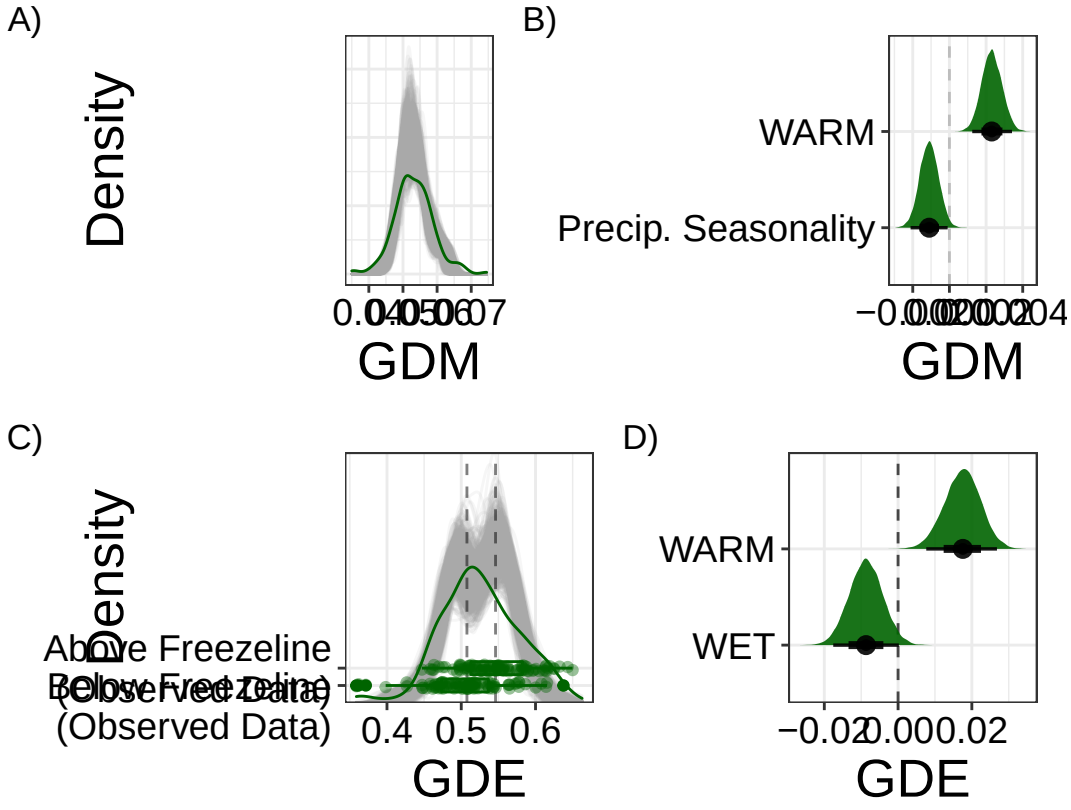


Beta posterior panes Posteriors for GDE



Posteriors for GDM





Option 1

Write to file

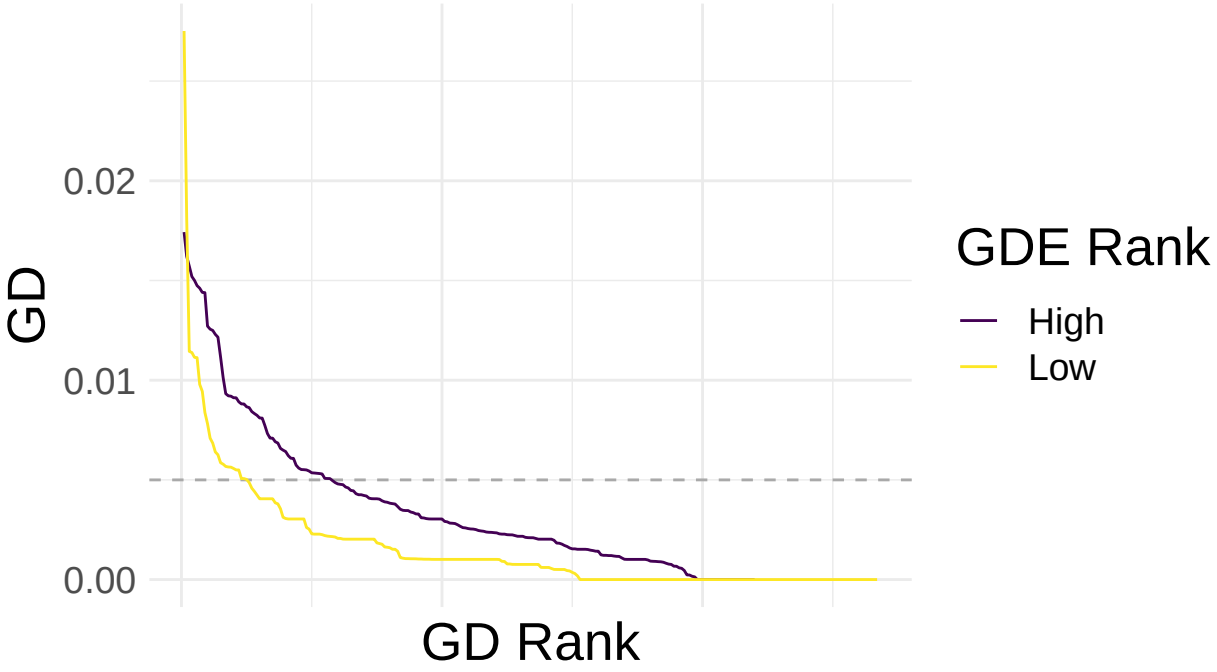
GDE conceptual figure

Read in data

```
## Reading layer `medium_150' from data source
##   `/Users/connorfrench/Dropbox/Old_Mac/School_Stuff/CUNY/BigAss-bird-phylogeography/BigAss-phylogeog
##   using driver `GeoJSON'
## Simple feature collection with 188 features and 125 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: -14472530 ymin: -5055287 xmax: 16986470 ymax: 7103713
## Geodetic CRS:  WGS 84
```

Filter pi df for cells that are within the final data frame

Figure



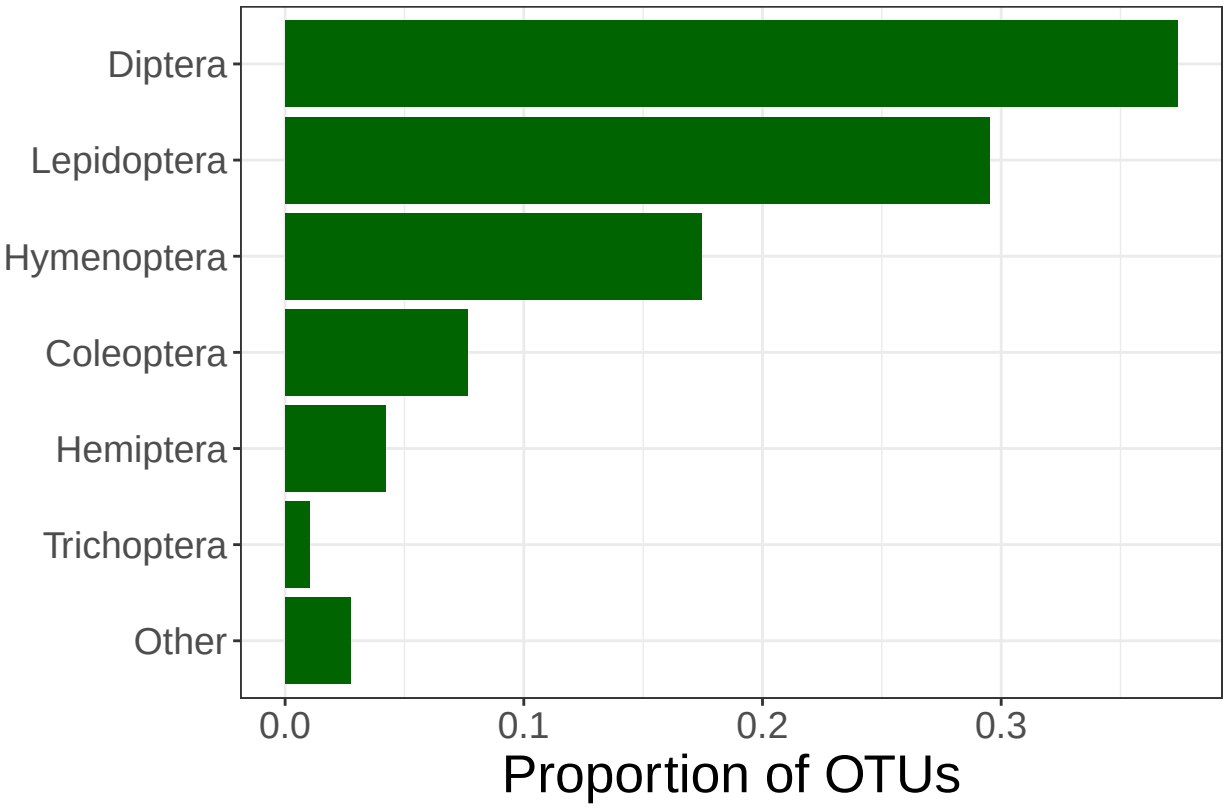
GDE conceptual fig: Illustration of distributions described by the first order Hill number of average pairwise nucleotide diversity across OTUs (GDE). Each line summarizes the distribution of genetic diversity for the cells with the maximum and minimum GDE values in the current dataset. The values are ordered from highest GD to lowest GD. The horizontal dashed line indicates a theoretical GDE value of 1.0, where all OTUs have the same GD.

Save to file

Order sampling figure

Read in data and do some wrangling

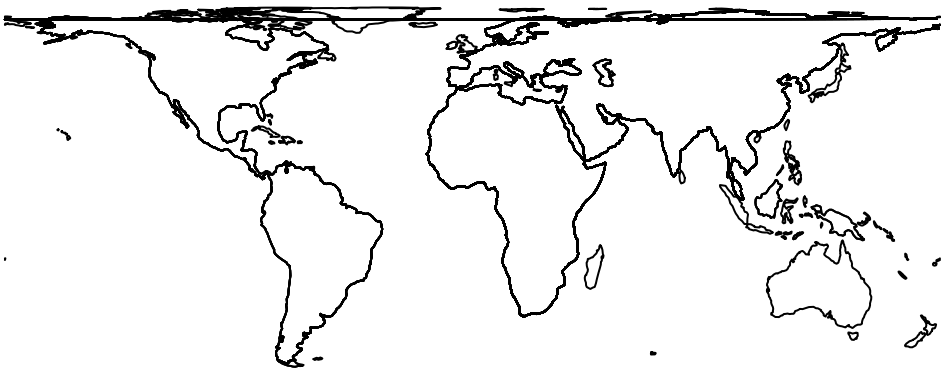
Plot the figure



Save to file

Prediction Maps

Map helpers



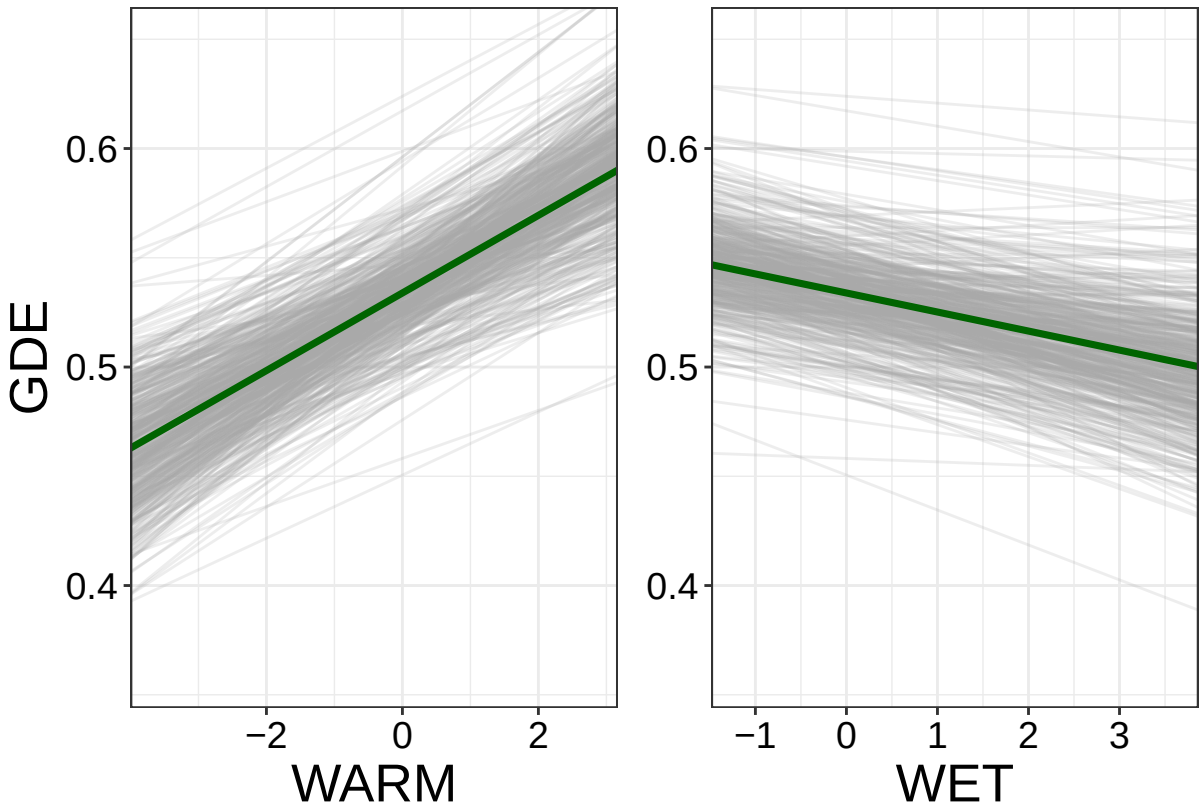
Read in and wrangle data

The global freezeline raster needs to be aggregated. It takes forever, so I write it to a file to read in after the first pass.

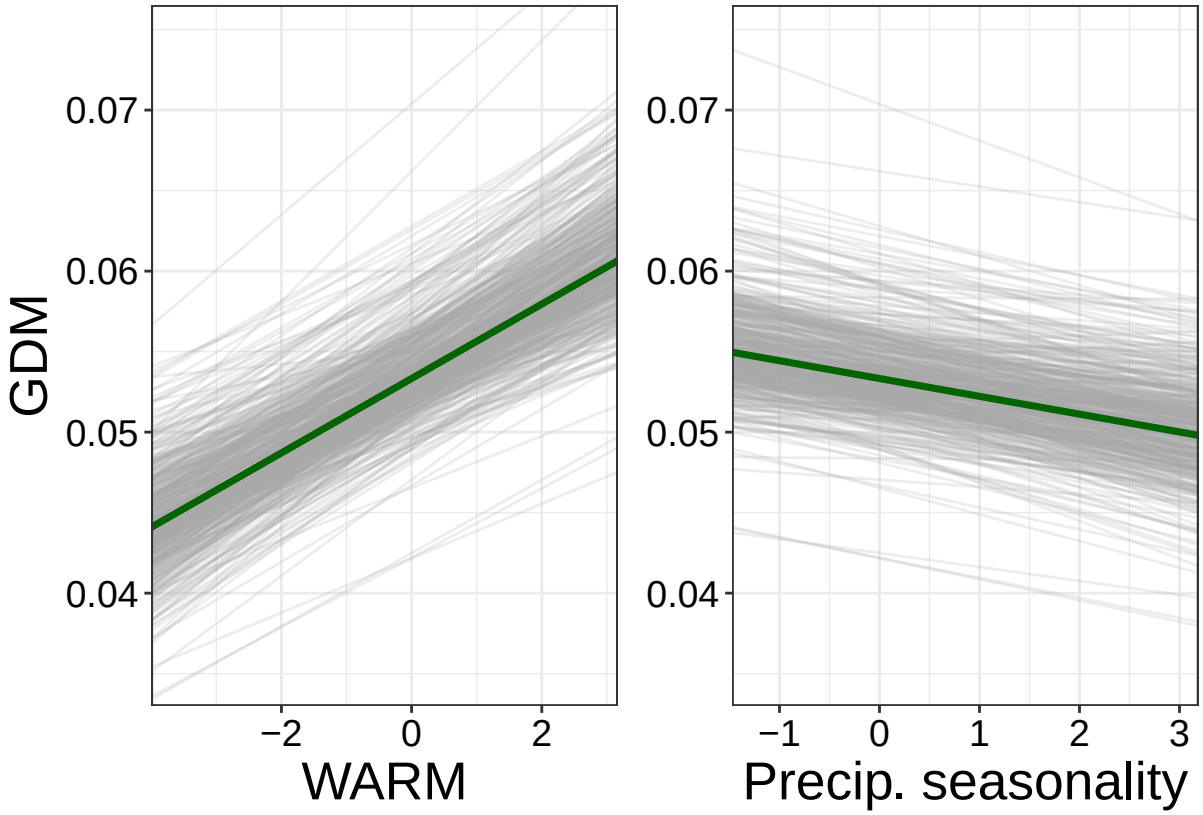
Read in if it's already been aggregated.

Partial dependence curves

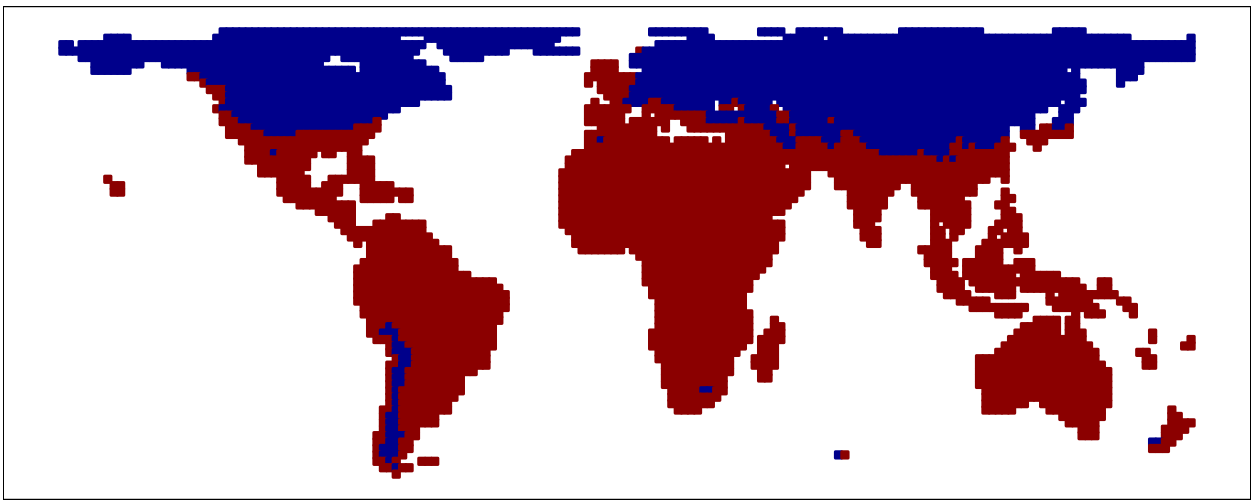
GDE partial dependence plots



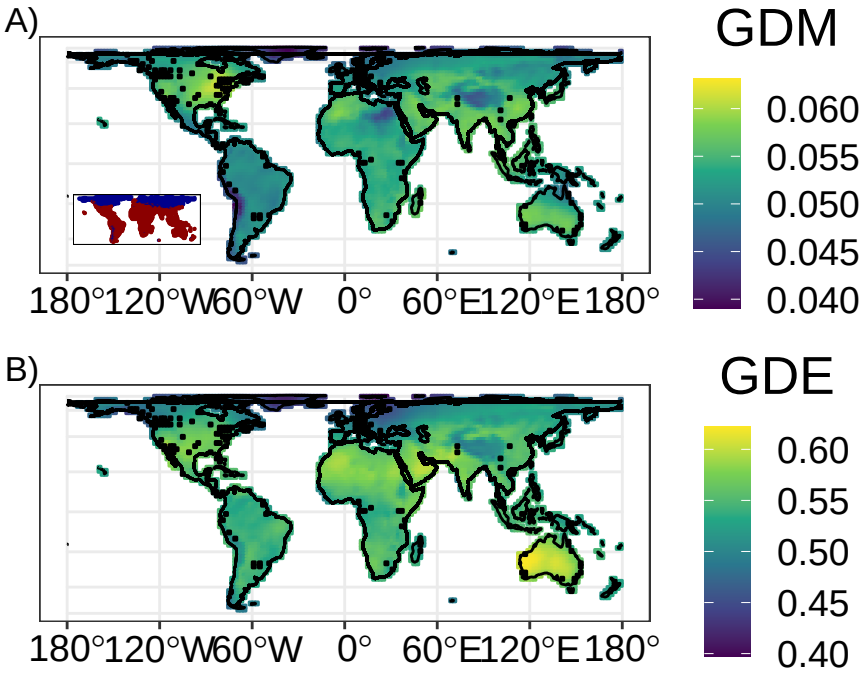
GDM partial dependence plots



Freeze line map for inset.



Make the map. USE BASE MAP WITH MORE ISLANDS. ADD PARTIAL DEPENDENCE CURVES UNDER EACH MAP



Predicted genetic diversity mean (GDM) and genetic diversity evenness (GDE). Genetic diversities are predicted from spatial linear models. For GDE A), the top model included MTWM and MPWM (see also posterior summary fig), while for GDM B), the top model included MTWM and precipitation seasonality. Black boxes indicate grid cells with oberved data (n = 187) used to build the models.

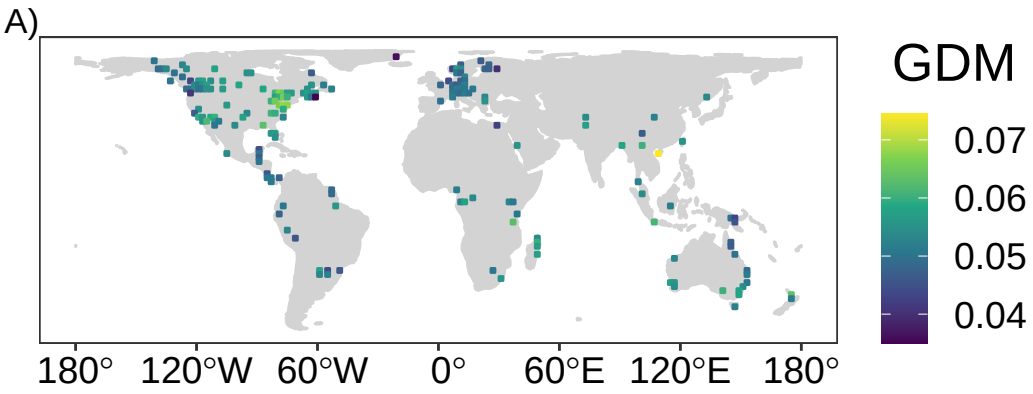
Write figure to file

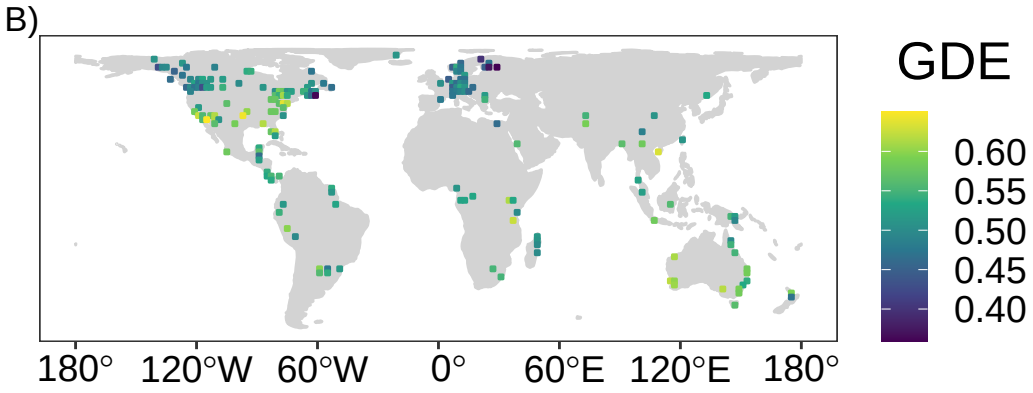
Observed Maps

ADD LATITUDE TO OBSERVED MAPS

Map helpers

Read in data

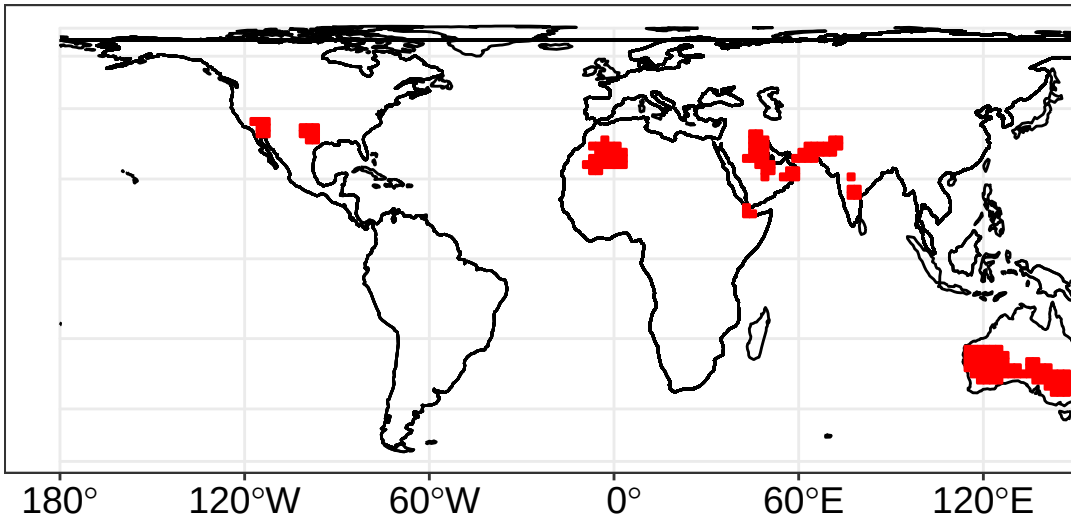




Hotspot maps

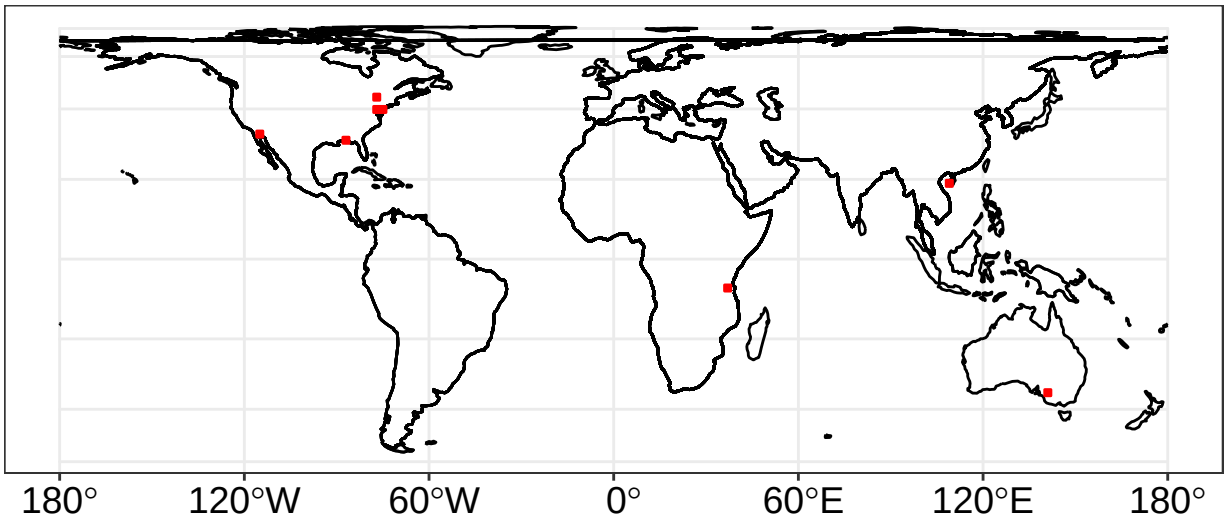
Map helpers

Read in and wrangle data



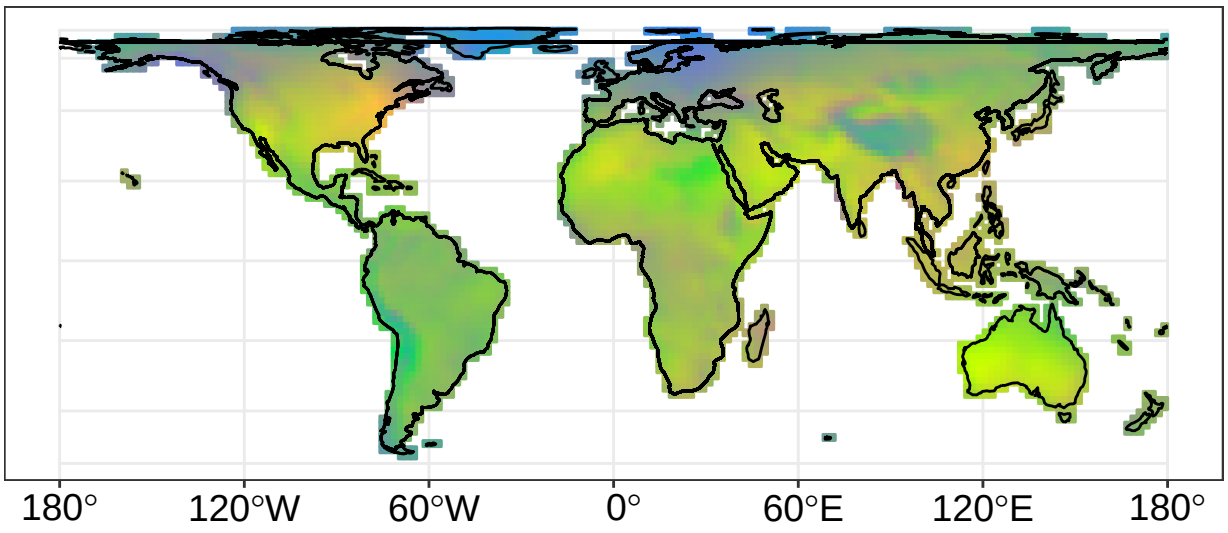
Hotspot map of predicted values

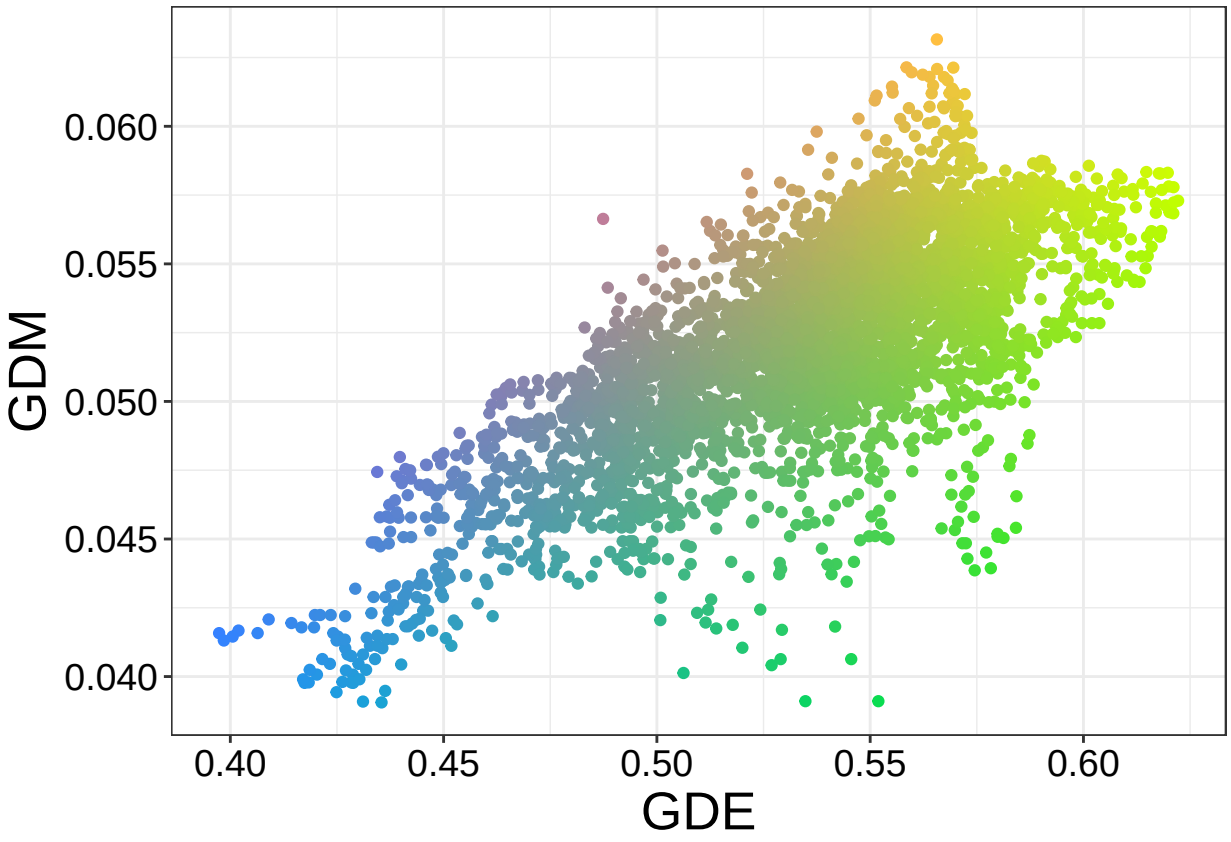
Observed hotspot map



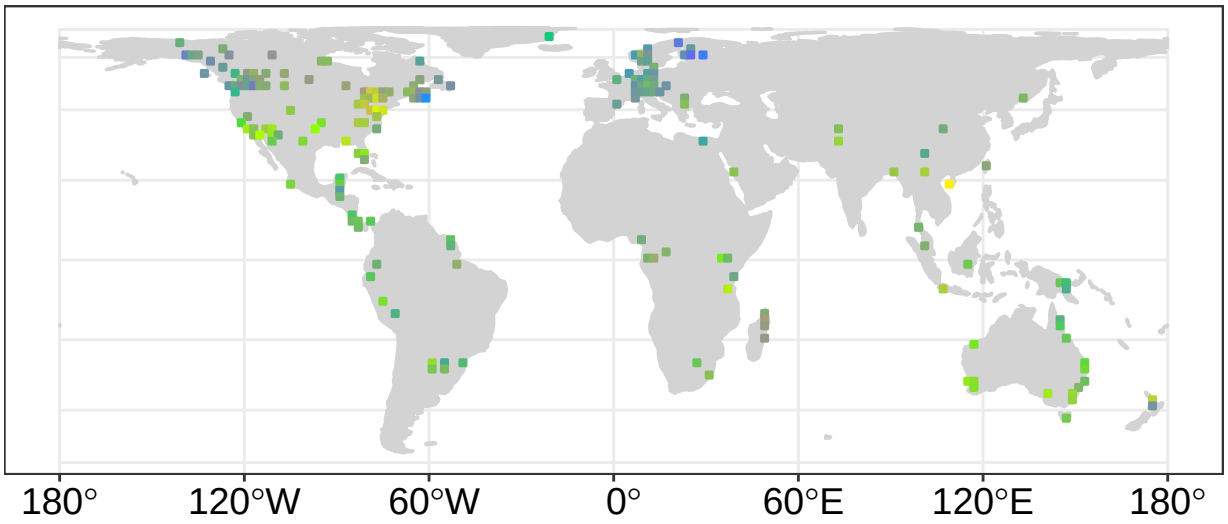
2D map

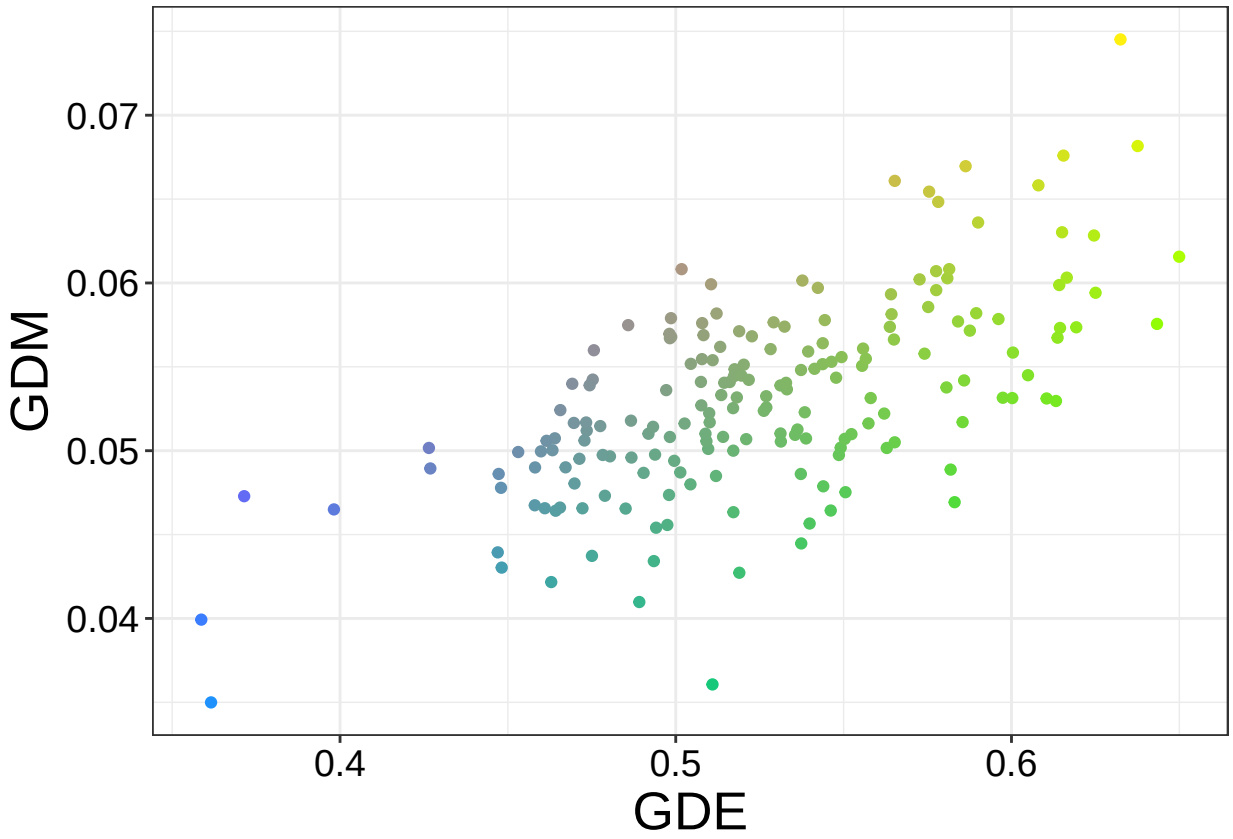
Projected values





Observed values





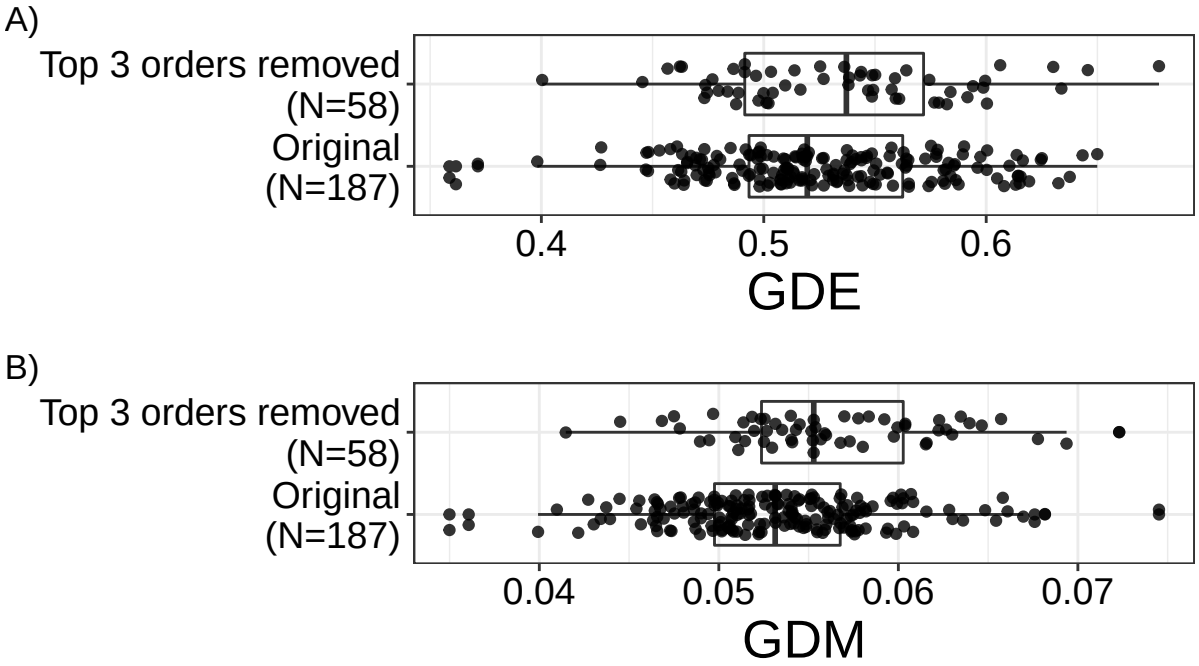
Supp Figs

Order outliers

Read in data and filter, removing outlier orders. I'm only considering the outlier data set that retains the same filtering regime as the original data to keep things comparable. I initially explored multiple filtering regimes since the number of cells is drastically reduced, but the distributions look similar and it's more logical to only compare the data sets with the same filtering regime.

```
## Rows: 245
## Columns: 4
## $ cell    <dbl> 636, 643, 943, 944, 1292, 1294, 1297, 1356, 1357, 1470, 1471, ~
## $ hill    <dbl> 0.4454309, 0.4622727, 0.5744267, 0.5767245, 0.4863921, 0.47329~
## $ avg_pi  <dbl> 0.04147382, 0.04450441, 0.05429685, 0.05352699, 0.05145907, 0.~
## $ min_otu <chr> "Top 3 orders removed\n(N=58)", "Top 3 orders removed\n(N=58)"~
```

Boxplots comparing the GDE and GDM distributions.



Distribution of per−cell GDE A) and GDM B) with the three orders that make up the largest numbers of OTUs in the data set either removed or kept. The distributions are not significantly different for GDE (Cohen's D = −0.15 [−0.44, 0.15]; p = 0.335) but they are for GDM (Cohen's D = −0.50 [−0.79, −0.19]; p = 0.002).

Save plot to file

Welch's t-tests to determine whether the distribution of GD of the complete data set is significantly different from the distribution of GD with the top three orders removed, while keeping the minimum number of OTU filter the same.

GDE does not significantly differ (Cohen's D = -0.15, p = 0.335), but GDM does (Cohen's D = -0.50, p = 0.002).

```
##
##  Welch Two Sample t-test
##
## data:  hill by min_otu
## t = -0.96996, df = 92.347, p-value = 0.3346
## alternative hypothesis: true difference in means between group Original
## (N=187) and group Top 3 orders removed
## (N=58) is not equal to 0
## 95 percent confidence interval:
##  -0.024289275  0.008348739
## sample estimates:
##            mean in group Original\n(N=187)
##                                  0.5256608
## mean in group Top 3 orders removed\n(N=58)
##                                  0.5336311
```

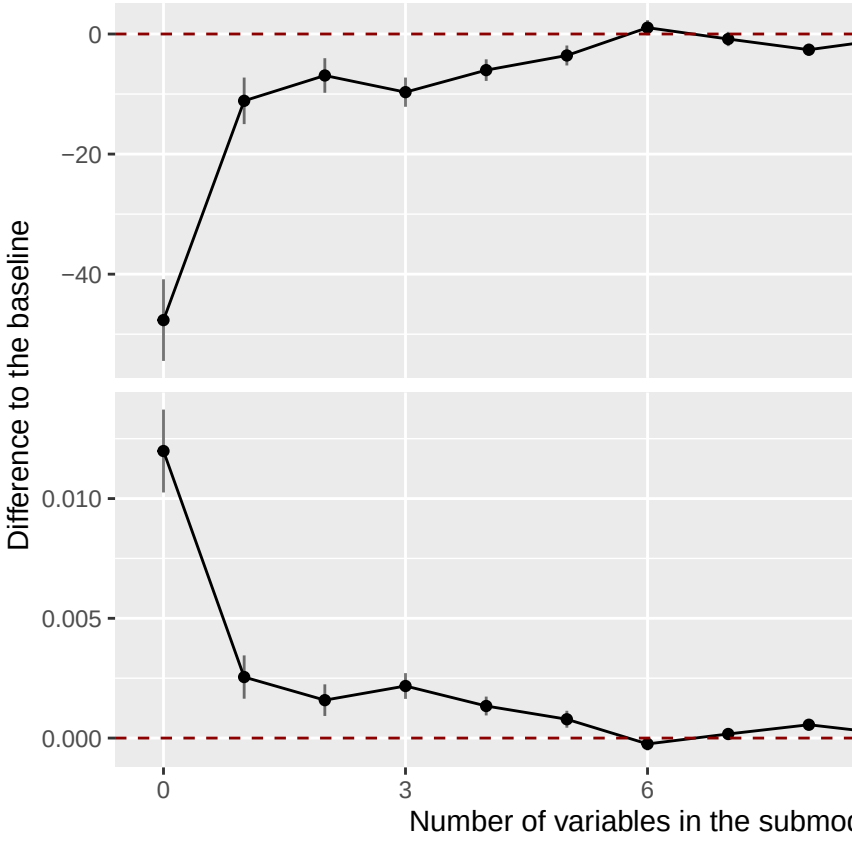
```
## Cohen's d |        95% CI
## -------------------------
## -0.15     | [-0.44, 0.15]
##
## - Estimated using un-pooled SD.
```

```
##
##  Welch Two Sample t-test
##
## data:  avg_pi by min_otu
## t = -3.2628, df = 91.195, p-value = 0.001553
## alternative hypothesis: true difference in means between group Original
## (N=187) and group Top 3 orders removed
## (N=58) is not equal to 0
## 95 percent confidence interval:
##  -0.004749268 -0.001154899
## sample estimates:
##            mean in group Original\n(N=187)
##                                 0.05316686
## mean in group Top 3 orders removed\n(N=58)
##                                 0.05611894
```

```
## Cohen's d |         95% CI
## -------------------------
## -0.50     | [-0.80, -0.19]
##
## - Estimated using un-pooled SD.
```

Model selection fig

GDE



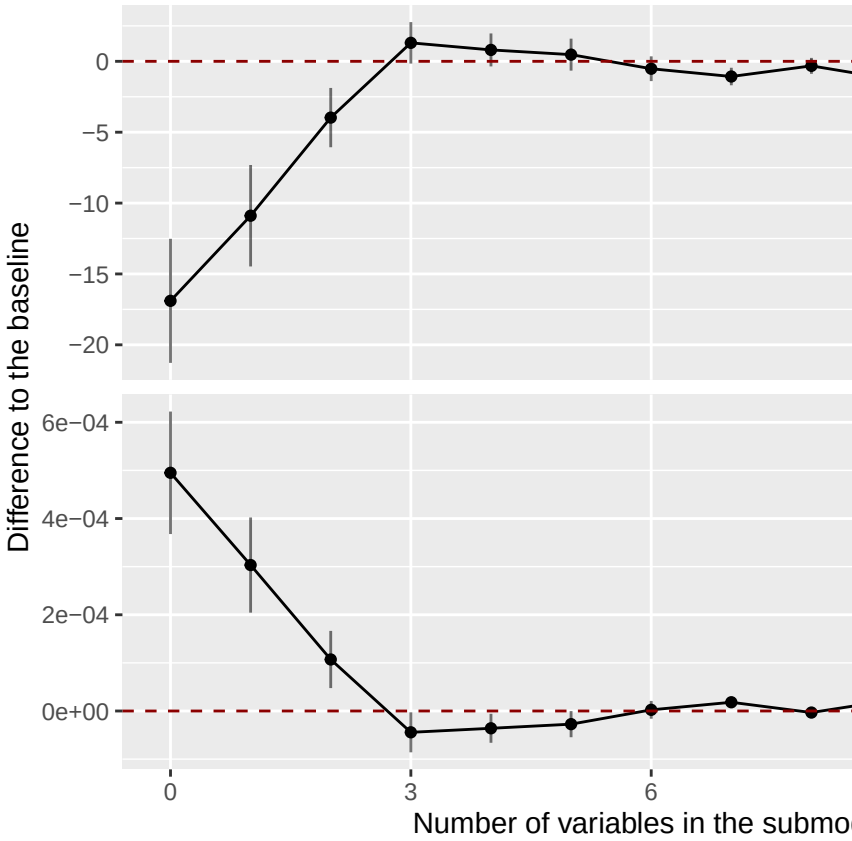
Plot the validation results relative to the full model

Write to file.

```
## pdf
```

```
##   2
```

GDM



Plot the validation results relative to the full model

Write to file.

```
## pdf
##   2
```

MESS Maps

Multivariate Environmental Similarity Surfaces comparing the environmental space in the observed data compared to the global distribution that we projected to.

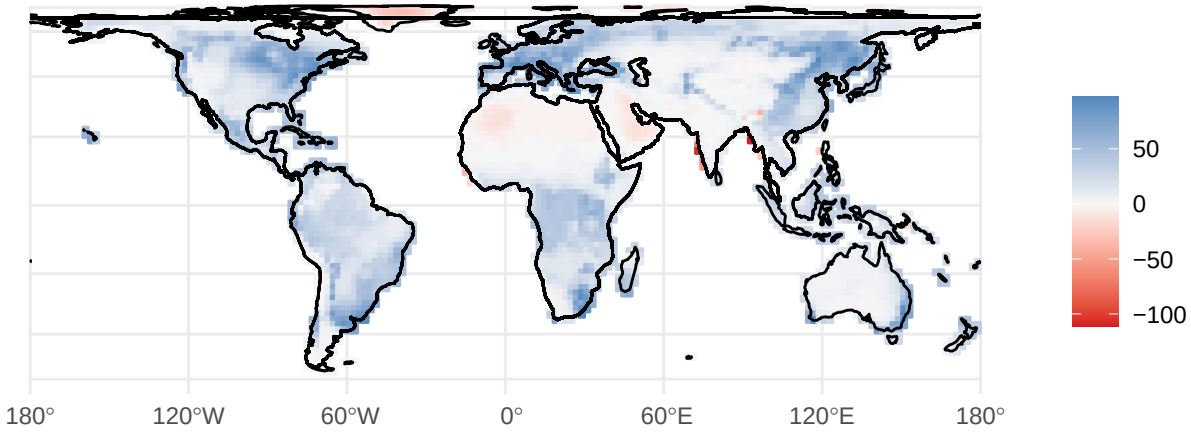
Map helpers

GDE

Read in and filter data.

Plot the MESS map for GDE.

Multivariate environmental similarity surface (GDE)

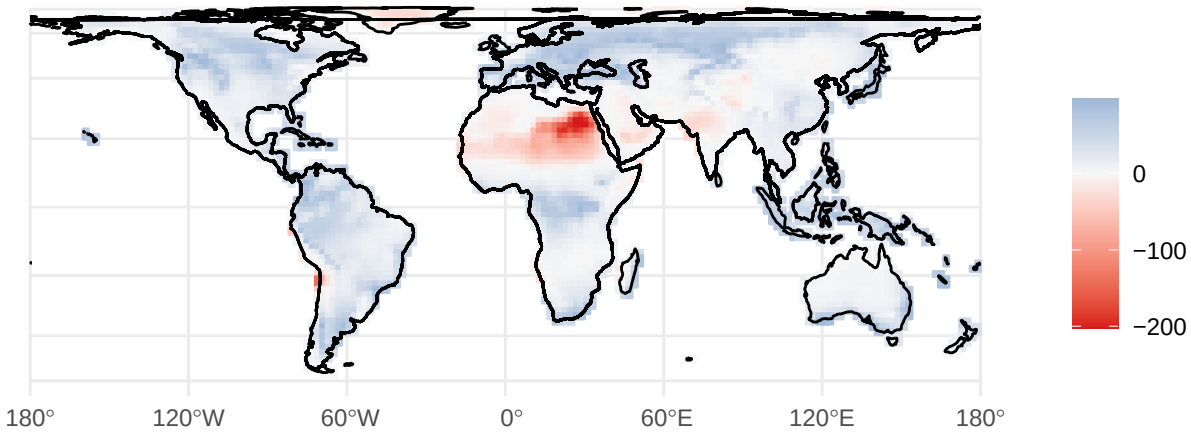


GDM

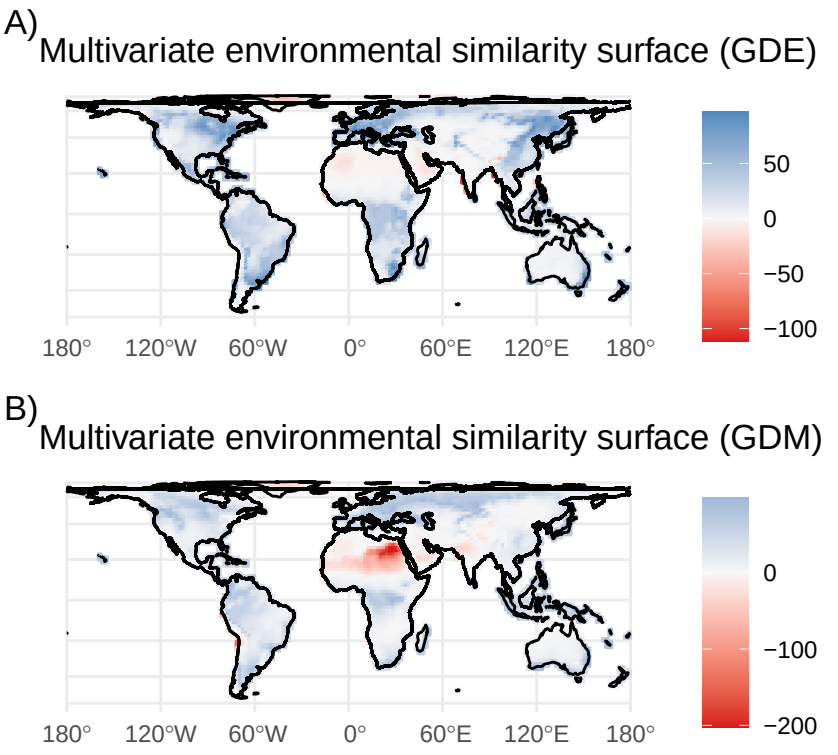
Read in and filter data.

Plot the MESS map for GDM.

Multivariate environmental similarity surface (GDM)



Final figure



Write to file.

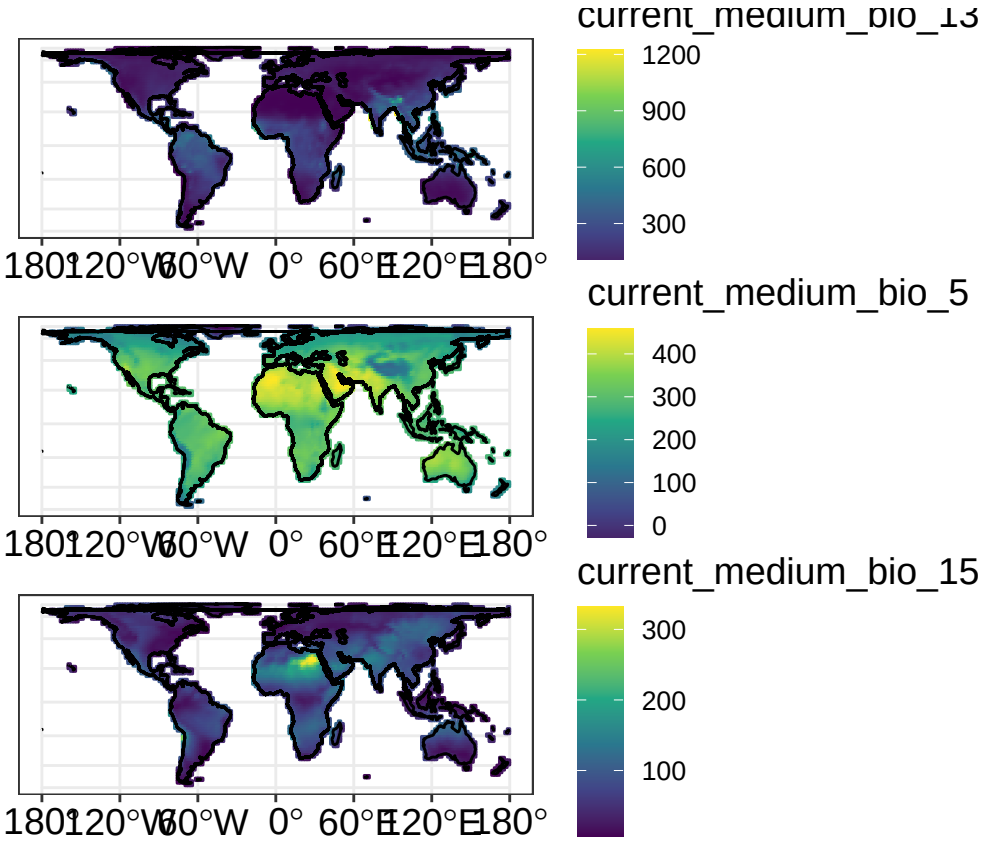
Predictors

Map helpers

Read in data

Mapping function

Make maps

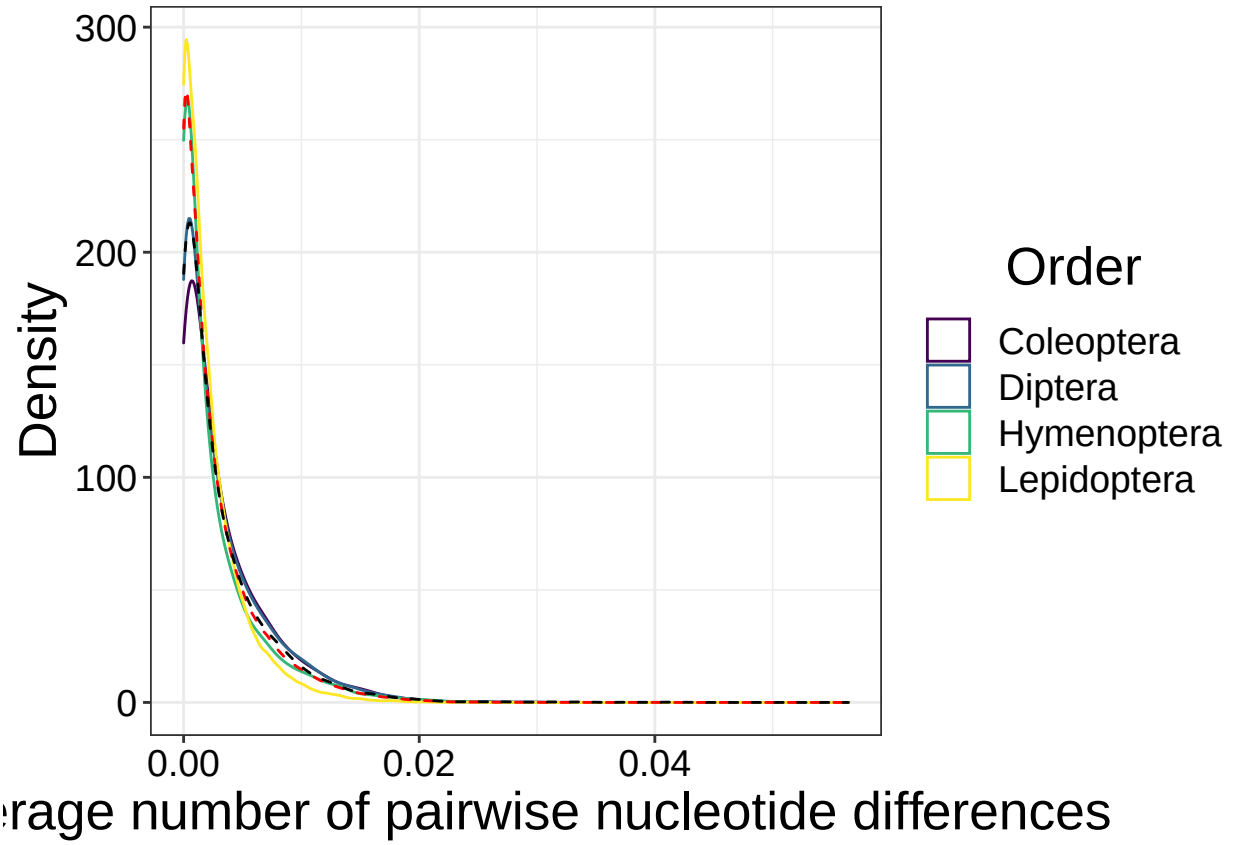


Per-order GD

Map helpers

Read in and wrangle the data

Visualize GD as density plots. Need to add a legend for the complete data set and data set without the top three orders.



Order data sets

Function to filter data sets

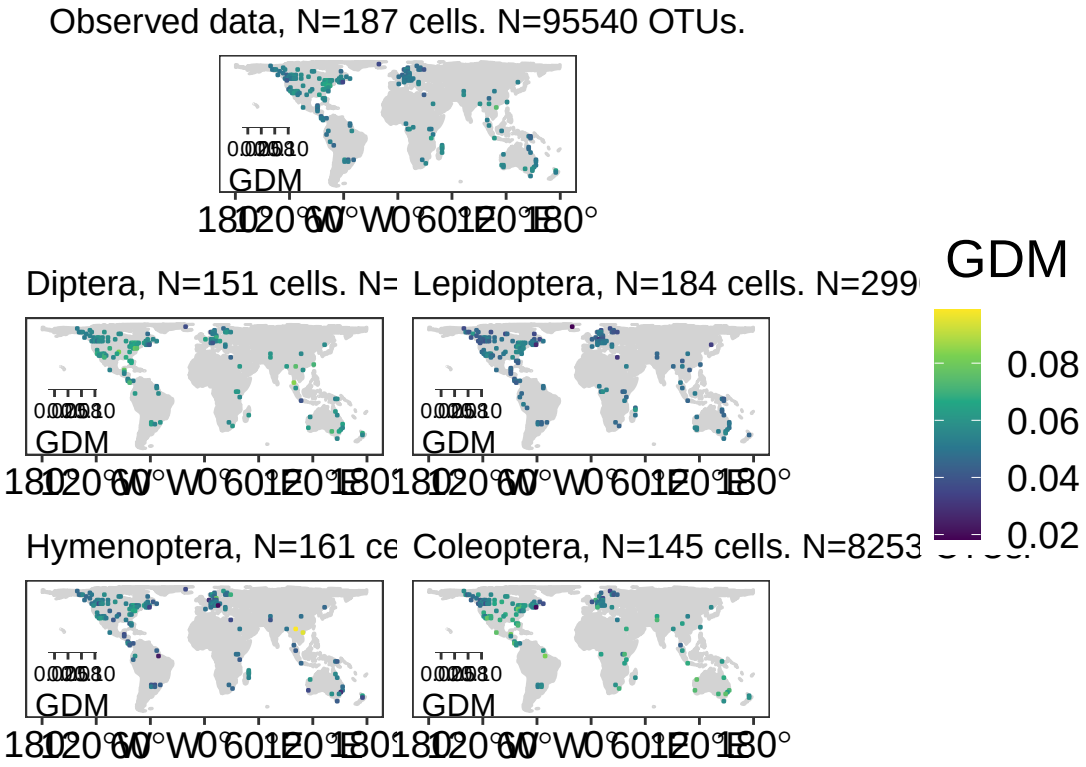
Get data sets for each order

Convert the data sets to sf for mapping

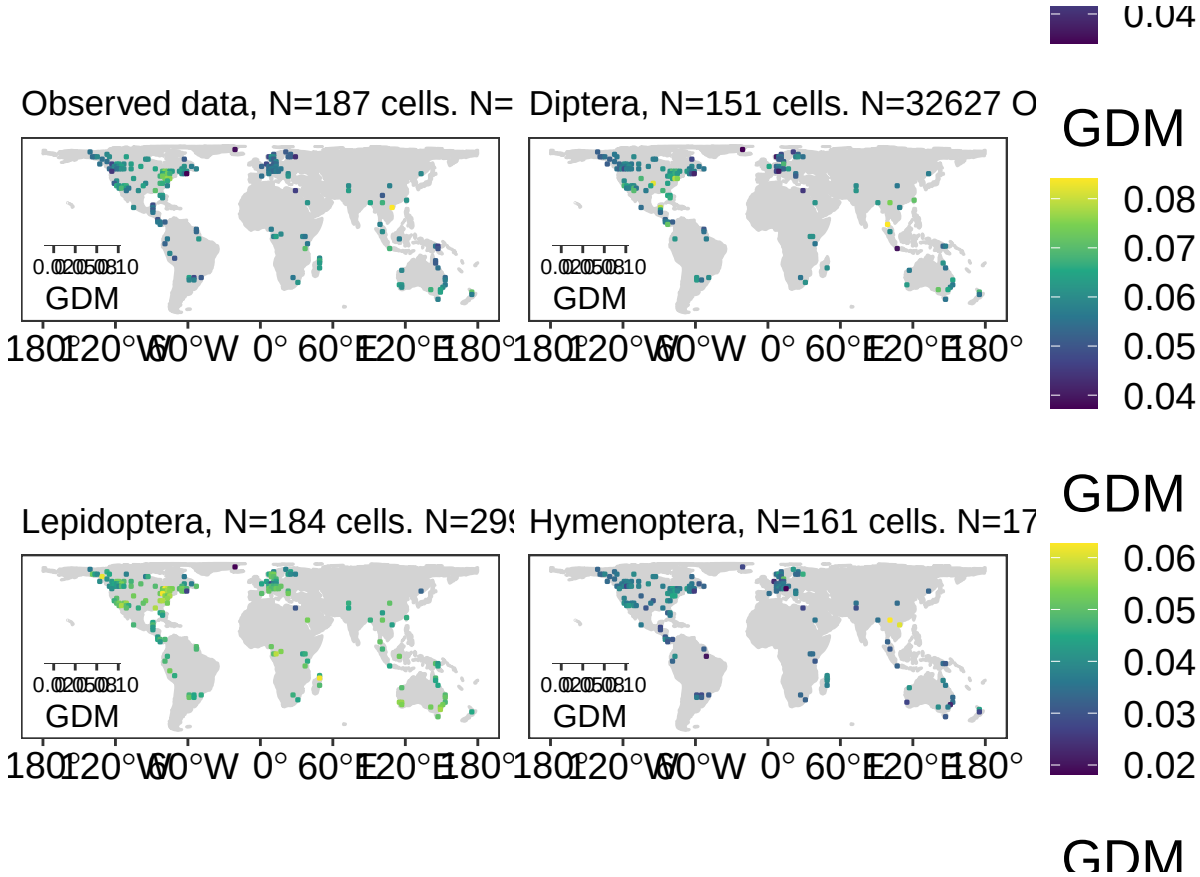
Make maps

These are maps where the minimum and maximum GDM are scaled to the minimum and maximum across all data sets.

Colors are scaled to the min and max values across data sets



These are maps with independent scales for each data set.



Sampling bias

Per the great suggestion by Rob Anderson (comment here for posterity): > I wonder what you can do to test for artifacts of sampling bias and potentially correct for it. I see that you focused only on cells that meet thresholds of data availability. However, those pixels still surely vary considerably in how complete the sampling was. The question is whether or not that affects the results. It occurs to me that you could take data from single cells that have the most data - and rareify that (randomly) to see if the GD estimates vary. If they don't, you've documented robustness to that. If they do (and in a predictable way), then you potentially could correct for it (even if the relationship is not linear, rather for example following some allometric scaling relationship). What do you think about all this?