**Project Summary**

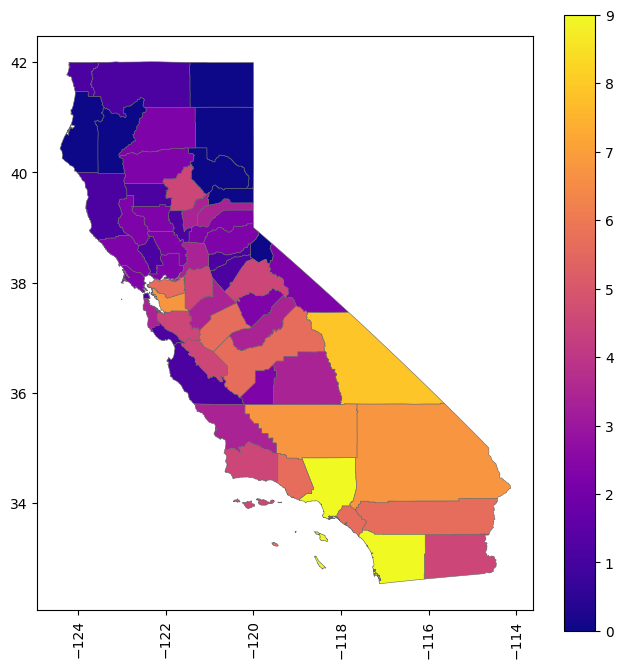
The alarming decline in global insect populations, known as the "Insect Apocalypse," is a critical issue attributed to human development, pesticides, and climate change. Insects play a pivotal role in ecosystem health and are essential for various ecological services. The California Insect Barcoding Initiative (CIBI), led by the Natural History Museum of Los Angeles County, aims to barcode every insect species in California, forming a comprehensive database to monitor population changes.

Our project aligns with CIBI by leveraging the Global Biodiversity Information Facility (GBIF) database to map the distribution of mantises (Mantodea) in California. We will identify distribution gaps and clusters, analyzing associated factors like ecoregion, human population density, and climate data. This analysis will help predict mantis habitats, providing insights into their ecological role. By overlaying these predictions with CIBI sampling sites, we'll propose targeted locations for future mantis barcode sampling, presenting a workflow applicable to other insect groups. Ultimately, this approach aids NHM in efficiently utilizing resources for insect population monitoring and conservation efforts.

For our analysis, we are investigating four types of data: biodiversity, ecoregion, climatological, and human population density. In preprocessing of each data type, data is disaggregated by counties as our primary key so that associations can readily be made between dataset. Counties were chosen for their relatively uniform and computationally manageable size across California state. For all analyses, county data was obtained from the Database of Global Administrative Areas because it is a source of highly accurate and high resolution administrative boundary geographical data (*Database of Global Administrative Areas 2022*).

**Biodiversity Update**

We were able to synthesize the biodiversity of mantises in each county as shown in Figure 1. This was accomplished by using county geographic data, which provided data on the borders of the counties in California, combined with the GBIF data that gave us access to the coordinates of each mantis observation in California. Our biodiversity data contains the observations of mantises by observers and has 4,556 records that include helpful columns such as species name, latitude, longitude, date, and elevation of the place and time of the observation. This dataset was sourced by the Global Biodiversity Information Facility (GBIF), an international data hub funded by governments worldwide to provide access to all types of data about life on Earth (*Telenius, Anders 2011*). Using the coordinates from the GBIF data, we mapped each observation to a specific county, and from there filtered for only unique species. This allowed us to see how many distinct species are found in each county, and potentially see where more observations might be needed.

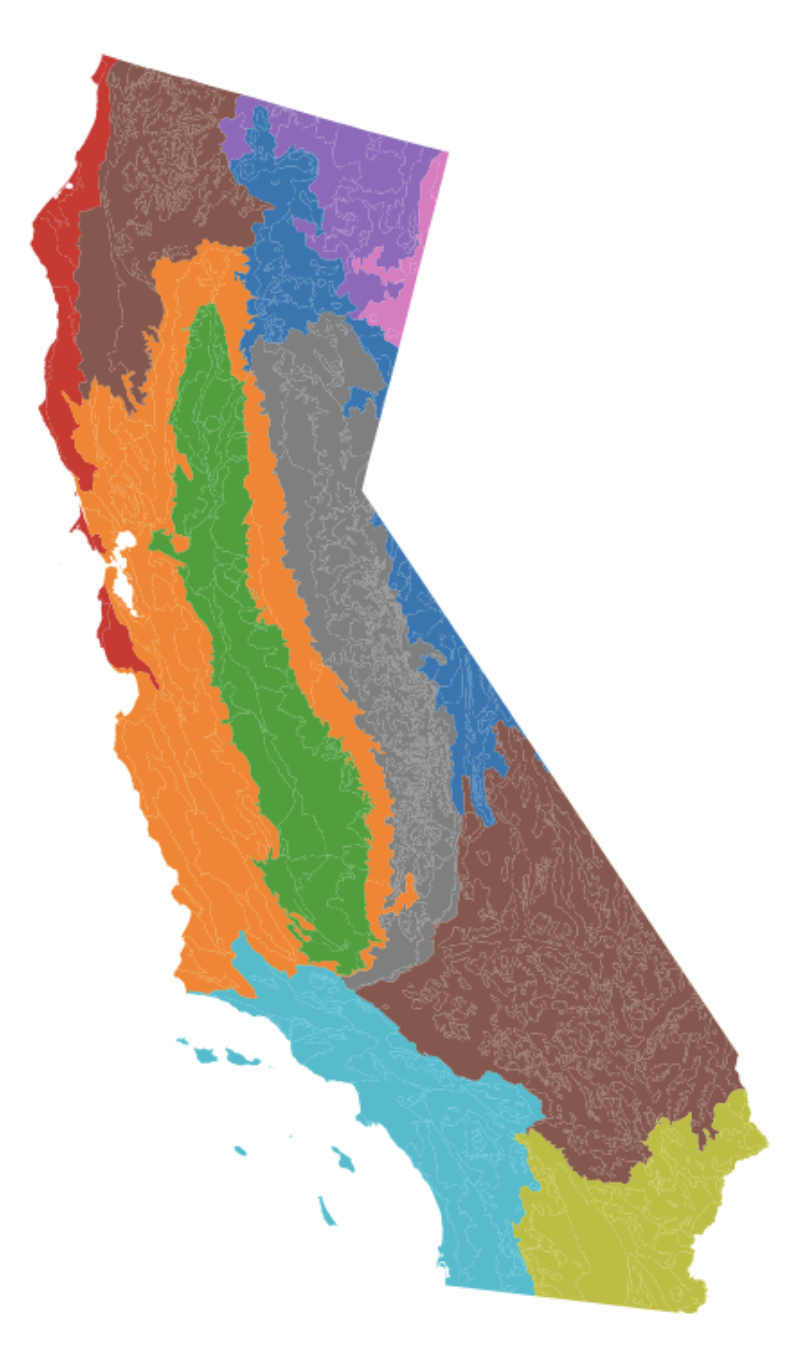
****

***Figure 1:*** *A heatmap of the diversity of mantis (Mantodea) species in California counties.*

On initial observation, it appears that heavily populated areas (such as places near Los Angeles, San Diego, and San Francisco) have more mantis diversity, which is something we will continue to investigate as the project progresses. We may find this is an artifact of sampling bias.

**Ecoregion Update**

The California ecoregion data, which comes from the US Environmental Protection Agency (EPA), contains level 3 and level 4 shapefiles which differ in their granularity (*Griffith et al. 2022*). After plotting both shapefiles, it appears that the more broadly defined level 3 ecoregions (*Figure 2*) would be better suited for analysis; the level 4 data splits California into an excessive number of ecoregions which do not relate well to counties.



***Figure 2:*** *Level 3 Ecoregions of California disaggregated by county*

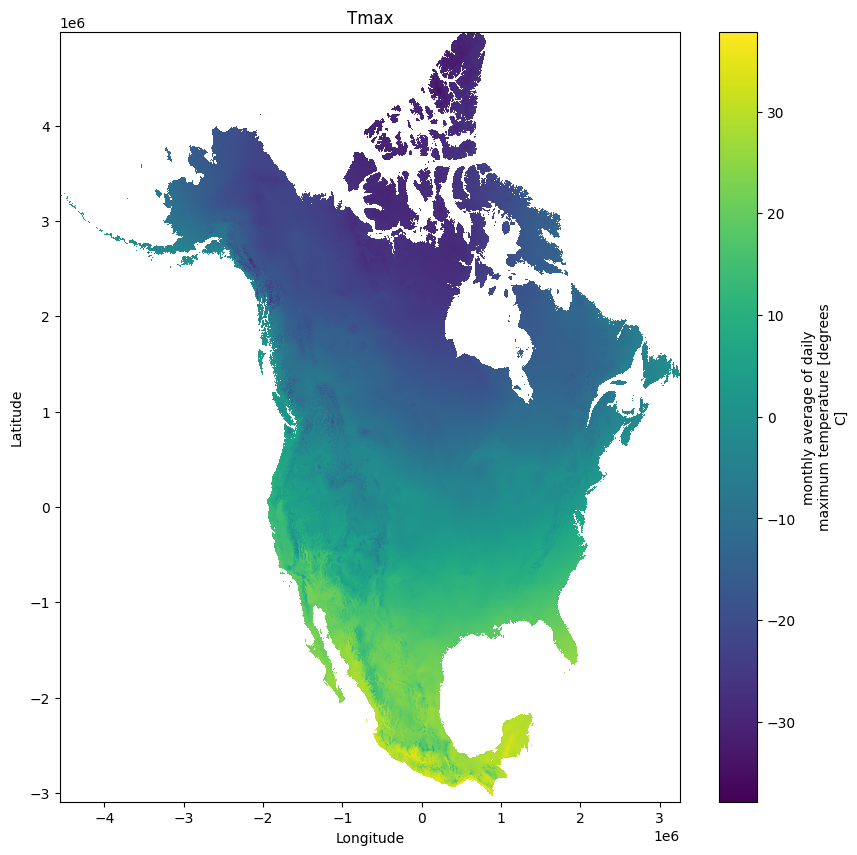
After looking at the ecoregion data in Figure 2, we wanted to integrate it with county data. After cleaning and filtering the data, we were left with another shapefile splitting California into its 58 counties. We combined the county and ecoregion data by intersection points and created a table with county (the primary key) and a list of that county’s ecoregions. We will be utilizing county data as a linking factor across different datasets (human population, biodiversity, climatological).

**Climatological Update**

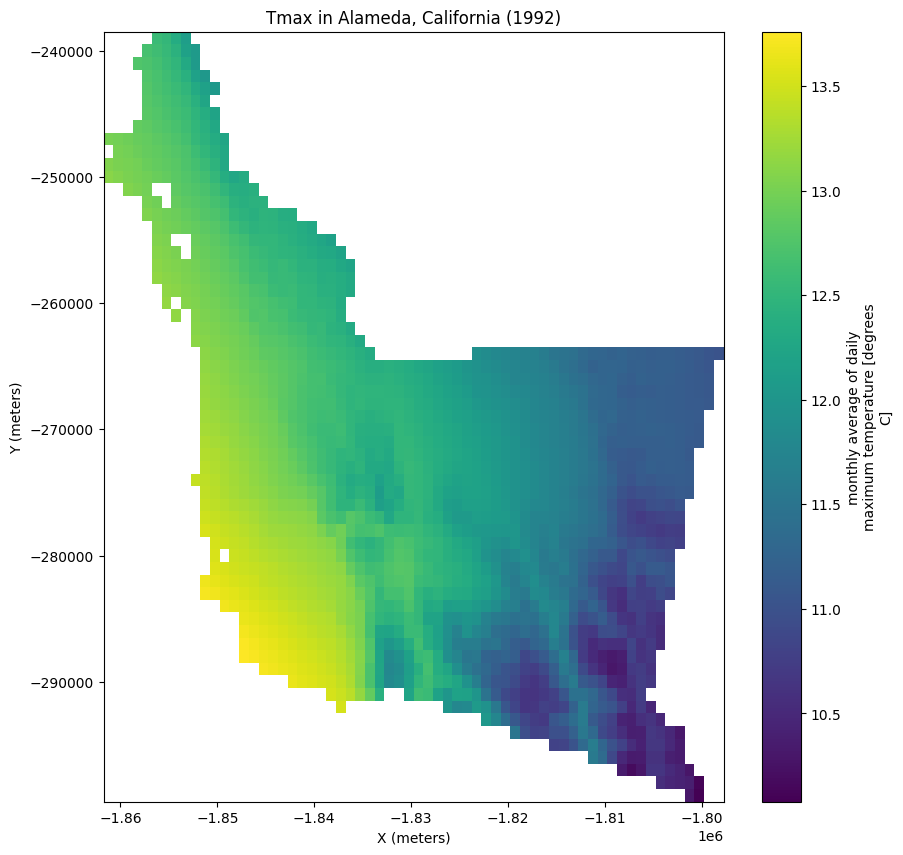
Daymet is a comprehensive data product hosted by the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) that provides gridded estimates of daily weather parameters across North America, Hawaii, and Puerto Rico, offering critical insights for understanding terrestrial biogeochemical processes (*Thornton et al. 2022*). The dataset is generated through sophisticated algorithms and computer software that interpolate and extrapolate from daily meteorological observations, resulting in detailed grids of weather parameters such as temperature, precipitation, vapor pressure, and snow water equivalent. Developed and curated with support from NASA and the U.S. Department of Energy, Daymet stands out for its high spatial (1 km x 1 km) and temporal (daily) resolution, stretching from 1980 to the present for North America and from 1950 for Puerto Rico. The dataset is indispensable for areas lacking meteorological instrumentation, providing valuable data for ecological and climate research. This data source was chosen for its high level of accuracy, high temporal resolution, and professional interpolation across remote areas without meteorological stations, which is the case in areas of California state.

Access to Daymet data is facilitated through the Thematic Real-time Environmental Distributed Data Services (THREDDS) Data Server, which offers a variety of remote data access protocols and dataset aggregation capabilities (*Unidata 2023*). Specifically, the Daymet: Monthly Climate Summaries on a 1-km Grid for North America, Version 4 R1 dataset provides monthly climate summaries derived from the daily data, allowing for detailed seasonal analysis. This dataset includes monthly averages and totals for key weather parameters, and is available in netCDF and GeoTIFF formats, ensuring easy integration with various research and analysis tools. The NetCDF format was chosen in our workflow because of the large amounts of data we would need to process over the large area of California state. Metrics on monthly total precipitation, average snow water equivalent (SWE), maximum temperature, minimum temperature, and average vapor pressure are available. Following domain expertise, total precipitation, maximum temperature, and minimum temperature are the most biologically relevant to mantises, so these metrics were used in our analyses.

After data was downloaded, it was visualized to verify that the desired data area was obtained. Figure 3 shows an example of maximum temperature data visualized using Python. After reprojecting county polygons to match the projections of Dayment data, the data was clipped to county boundaries. An example is shown in Figure 4. The average values across counties was then calculated in order to produce a table with county as the primary key.



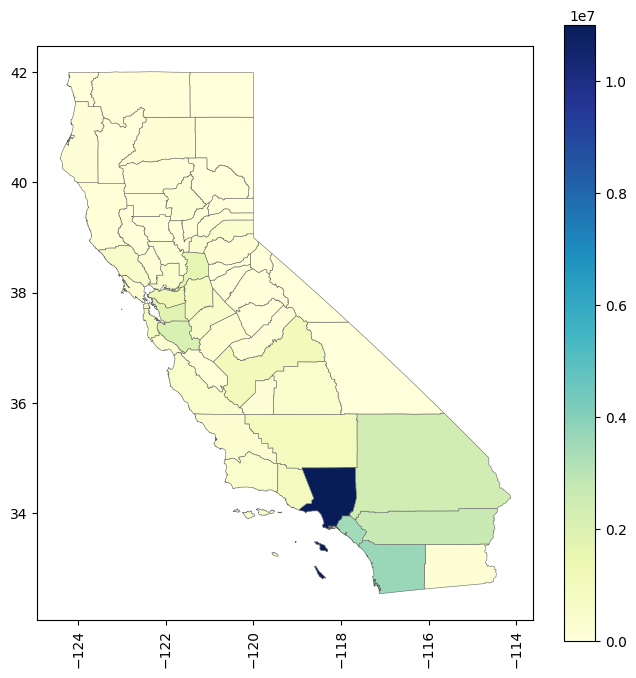
***Figure 3****: Visualization of Daymet monthly average of maximum temperature (°C) in the year 1980 across North America.*



***Figure 4****: Visualization of Daymet monthly average of maximum temperature (°C) in the year 1980 clipped to the boundary of Alameda County, California.*

**Human Population Density Update**

We acquired our human population density data from the US Census Bureau 2020 American Community Survey, which we cleaned to only focus on the total population of each county in California (US Census Bureau 2020). We used Python to plot a heatmap of the population for each county, as can be seen below.



***Figure 5****: Population density by county in California*

The most populated county by a large margin was Los Angeles, while a few other counties in Southern California and the Bay Area had higher populations (Orange, San Diego, San Bernardino, Riverside, Santa Clara, Alameda, Sacramento). Generally, the rest of the counties have proportionally low human populations.

**Next Steps**

Having successfully acquired and organized our diverse datasets—biodiversity data, ecoregion data, climatic data, and human population data—into structured tables with 'county' serving as a common primary key, our next steps involve the creation of an Entity-Relationship (ER) diagram to visually represent the relationships and dependencies between these datasets. This diagram will serve as a foundational tool for our subsequent analysis, aiding in the identification of associations and patterns that could predict the presence of mantises across different counties in California. Leveraging machine learning models such as the Apriori Algorithm and Decision Trees, we aim to uncover meaningful relationships and dependencies between mantis presence and various ecological, climatic, and anthropogenic factors. The insights garnered from these models will not only enhance our understanding of mantis distribution but also pinpoint potential gaps in their observed presence. By identifying counties where mantises are predicted to exist, yet are currently unobserved, we can provide valuable recommendations to the California Insect Barcoding Initiative (CIBI), guiding their future sampling efforts and contributing to the broader goal of biodiversity conservation and monitoring.

**Citations**

*Database of Global Administrative Areas (2022). GADM database of Global Administrative Areas, version 4.1. [online] URL:* [*https://*](https://doi.org/10.3334/ORNLDAAC/2131)[*gadm.org*](http://www.gadm.org)*.*

*Griffith, Glenn E., et al. (2016). "Ecoregions of California." US Geological Survey Open-File Report 1021: 1-45.*

*Telenius, Anders. "Biodiversity information goes public: GBIF at your service." (2011) Nordic Journal of Botany 29.3: 378-381.*

*Thornton, M.M., R. Shrestha, Y. Wei, P.E. Thornton, S-C. Kao, and B.E. Wilson. 2022. Daymet: Monthly Climate Summaries on a 1-km Grid for North America, Version 4 R1. ORNL DAAC, Oak Ridge, Tennessee, USA.* [*https://doi.org/10.3334/ORNLDAAC/2131*](https://doi.org/10.3334/ORNLDAAC/2131)

*Unidata, (2023): Integrated Data Viewer (IDV) version 6.2u1 [software]. Boulder, CO: UCAR/Unidata.* [*http://doi.org/10.5065/D6RN35XM*](http://doi.org/10.5065/D6RN35XM)

*U.S. Census Bureau (2020). Total Population, 2020 American Community Survey 5-Year Estimates Detailed Tables. Retrieved from* [*https://data.census.gov/*](https://data.census.gov/map?t=Populations+and+People&g=040XX00US06$0500000&tid=ACSDT5Y2020.B01003&layer=VT_2020_050_00_PY_D1&mode=thematic&loc=37.4190,-120.0923,z5.1617)