Local climatic effects on colonisation and extinction drive changes in mountain butterfly communities

Ursul, Guim[1,2], Mingarro, Mario[1,3], Castro-Cobo, Sara[1,2,4], Cancela, Juan Pablo[1,5,6], Romo, Helena[2,7] and Wilson, Robert J.[1]

[1] *Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain*

[2] *Universidad Autónoma de Madrid (UAM), Madrid, Spain*

[3] *Estación Experimental de Zonas Áridas (EEZA-CSIC), Almería, Spain*

[4] *Tragsatec, Madrid, Spain*

[5] *Centre for Ecology, Evolution and Environmental Changes (CE3C), Lisbon, Portugal*

[6] *Universidade dos Açores, Campus Universitário de Angra do Heroísmo, Terceira, Portugal*

[7] *Centro de Investigación en Biodiversidad y Cambio Global (CIBC), Madrid, Spain*

This dataset contains all data necessary to replicate all methods and results found in the manuscript with the title "Local climatic effects on colonisation and extinction drive changes in mountain butterfly communities"

Glossary:

STI: Species Temperature Index

CTI: Community Temperature Index

SC: Sample Coverage

qD: estimated species richness

Description of the data and file structure

All data is stored in one Excel file with several sheets. Below there is an explanation of the use and information contained on each sheet:

1. Coordinates_list: longitude and latitude coordinates for the 75 study sites to run Microclima in R and extract the modeled temperature values for each site. Also, the coordinates are used to conduct Moran's I test on spatial autocorrelation on model residuals.

2. iNEXT_spp_richness: contains the size-based output from *iNEXT* function in R (see https://cran.r-project.org/web/packages/iNEXT/vignettes/Introduction.pdf(opens in new window) for more information about iNEXT functionality). This data was used to extract estimated species richness (qD) for the several SC filters used in the paper. For each location and period (historical and recent) the model is iterated by equal increments on sampling size (column named m) up to double it. Then qD is either rarefacted if its calculated for a sample size below the observed value or extrapolated if its calculated for a sample size above the observed value. All information included is listed below:

- Code: location code which can also be referred as a "community"

- m: sample size of the assemblage or community

- Method: Rarefaction, Observed or Extrapolation

- Order.q: 0 refers to the hill number selected (q = 0 in our case because we are interested in species richness)

- qD: diversity estimate of order q or in this case, estimated species richness (q = 0)

- qD.LCL: 95% lower boundary confidence limits of diversity

- qD.UCL: 95% upper boundary confidence limits of diversity

- SC: Sample Coverage

- SC.LCL: 95% lower boundary confidence limits of sample coverage

- SC.UCL: 95% upper boundary confidence limits of sample coverage

- Survey: either Historical (1984-2005) or Recent (2017-2022)

- Region: region of study (Guadarrama, Gredos, Meridional or Javalambre)

3. Main_analysis_data: contains all variables used in the statistical analysis conducted for Figures 3, 4 and 5. All columns are explained below:

- Code: location code which can also be referred as a "community"

- Region: region of study (Guadarrama, Gredos, Meridional or Javalambre)

- Variable:

    o *CTI_IB* for Iberian CTI (average across all STI of species in a community)

    o *CTI_EU* for European CTI

    o *CTI_sd_IB* for CTIsd for Iberian CTI (SD across all STI of species in a community)

    o *CTI_sd_EU* for CTIsd for European CTI

    o *STI_sd_IB* for STIsd for Iberian STI (average across STIsd for all species in a community)

    o *STI_sd_EU* for STIsd for European STI

- Survey:

    o *Historical* (butterfly communities from 1984-2005 period)

    o *Recent* (buttferly communities from 2017-2022 period)

    o *diff* (difference between Recent and Historical) for each variable

- *mean to q75*: values extracted from the randomisation process for each variable including the average (*mean*), standard error (*se*), the median (which is used in the main analysis), 25th percentile (*q25*) and 75th percentile (*q75*).

- *recent_10y to diff_surv_y*: maximum temperature values extracted from Microclima R package for two different methods: 10-year periods (*10y*) 1980-89 and 2013-22 and for surveyed years (from September previous year of survey to August from the surveyed year). There are three columns for each method, either historical, recent and the difference (recent - historical Tmax).

- Estimated species richness (qD): there are three sets of columns here, one for each SC filter used (*qD.LCL* refers to the lower bound of qD (95%), *qD.UCL* refers to the upper bound of qD (95%) and *qD* refers to the estimated species richness which is used as an independent variable in the manuscript)

4. Persistence category analysis: data used for figure 6A from main paper about the different persistence categories (colonize, persist and extinct): follows the same pattern as sheet 3 with main analysis data but includes a column with persistence category (if a species has colonised that particular site, persisted over time or became extinct). *Variable* column contains either *CTI* of all species in each persistence category and site; and *STIsd* of all species in each persistence category. Three variables of modelled temperature are added here to compute test for Figure 6A.

5. Occupancy analysis: data used for any occupancy analysis related to their historical occupancy (Nsites historical) and the temporal change in their occupancy (ΔNsites):

- *Name*: Complete species name following Wiemers et al. (2018) nomenclature to be able to add the phylogenetic tree into the analysis.

- *Survey*: this refers to which variable mean, se, median, q25 and q75 refer, in this case, these variables refer to temporal change in Nsites (e.g. occupancy) for each species (N = 129). Median is used in all analysis.

- *IBER_mean_temp* and *IBER_sd_temp*: contains Species Temperature Index (STI) and the standard deviation on STI based on Iberian Peninsula distributions of butterfly species (Mingarro et al. 2021).

- *EU_temp.mean* and *EU_temp.sd*: contains Species Temperature Index (STI) and the standard deviation on STI based on European distributions of butterfly species (Platania et al. 2020).

- *N_sites_hist*: contains historical occupancy of species as the number of sites where they were present.

6. Paired test table: contains historical, recent and the difference for each of the 11 variables used in the analysis. This table is used to compute paired test to check for significant differences over time.

7. Microclima data: contains mean annual maximum temperatures for the two methods used to compute them: two 10-year period (1980-1989 and 2013-2022) and the surveyed years with butterfly data. This data is used to run spearman tests on microclima data and to compute Figure S1 to compare temperature data from both methods.

Sharing/Access information

Data was derived from the following sources:

- Species nomenclature from Wiemers et al. (2018): https://doi.org/10.3897/zookeys.811.28712(opens in new window)

- Phylogenetic tree for European butterflies from Wiemers et al. (2019): https://zenodo.org/records/3531555(opens in new window)

- Iberian Species Temperature Index from Mingarro et al. (2021): https://doi.org/10.1111/icad.12498(opens in new window)

- European Species Temperature Index from Platania et al. (2020): https://doi.org/10.1111/geb.13154(opens in new window)

More information about the methods behind estimated especies richness (*iNEXT*) and *Microclima* can be found in the following links:

- https://cran.r-project.org/web/packages/iNEXT/vignettes/Introduction.pdf(opens in new window)

- https://rdrr.io/github/ilyamaclean/microclima/man/runauto.html(opens in new window)

R code availability

R code is available in this Dryad DOI to replicate all analyses of the paper. More code related to this work can be found at https://github.com/guimursul/Public_MNCN