Detailed description of column headers of all tables

Table 'DataSources'

This table provides information at the lowest level, the data source. Each data source has a number of plots, years, and a spatial location and one or more references to the source of the data.

"DataSource_ID" – Unique identifier for each data source (numerical, 165 entries, not continuous)
"DataSource_name" - Unique descriptive name for each data source. Used for easy reference of the authors
"Realm" – Realm in which samples were collected. (factor with 2 levels: 'Terrestrial' and 'Freshwater')
"InvertebrateGroup" – coarse description of taxon/ taxa studied in data source (factor with 47 levels)
"AbundanceOrBiomass" – Does the data source provide information on insect biomass (B), abundance (A) or both (AB) (factor with 3 levels)
"Start" – First year of sampling in data source (numeric)
"End" – Last year of sampling in data source (numeric)
"DurationDataSource" – Time between first and last sample
"NrYrsData" – Number of years in which data was collected.
"NrSites" – Number of sites studied
"Continent" - Continent where samples were collected (factor with 7 levels)

- Africa: African continent
- Asia: Eurasia east of the Urals, Caucasus and Bosporus, including the Middle East and the Indian subcontinent. The eastern boundary lies west of New Guinea.
- Europe: Eurasia west of the Urals, Caucasus and Bosporus.
- South America: South America.
- Central America: Central America including Mexico and the Caribbean.
- North America: USA and Canada.
- Oceania: Australia and New Zealand.

"Region" – arbitrary grouping of countries and states into geographical regions (factor with 27 levels)
"NationState" – Nation state in which the samples were collected (factor with 41 levels)
"CountryOrState" – Geographic unit in which samples were collected. 'State' level is only used in large countries such as Russia, Brazil, Canada and the United States. (factor with 100 levels)
"OpenAccessLicense" – License for access, use and republishing of the original data source
Open access licenses:
PD: public domain (all data extracted from papers),
OGL: Open Government License (UK),
CC-BY, CC0, CC-BY-NC, CC-BY-ND,
ODC: Open Data Commons
no shar: data openly accessible, but no redistribution of data or derived products is allowed,
private: data are not publicly accessible, but the derived numbers are available

"Link" – URL linking to raw public datasets (websites were active until at least 2019)
"Reference" – reference to original data source. Reference list is found in Table 'References'

Table 'PlotData'

This table provides information about the location of sampling at the highest level, the plot. This includes information about the geospatial location, but also about the sampling and the environmental conditions according to the GIS layers used.

"Plot_ID" – Unique identifier for each site
"DataSource_ID" – unique identifier linking to table 'DataSources'
"PlotName" – descriptive name of the plot within the data source. Used for easy reference by the authors
"Location" – Grouping variable in case of groupings of plots within the data source
"DetailsPlots" – Descriptive names, locations, or details (e.g. plant species sampled), used in the original data source
"ExperimentalTreatment" – In case of experimental setups the experimental treatment of each plot (e.g. polluted, logged, control, number of plant species in plot, etc)
"Latitude" – Latitude (northing) of the site in decimal degrees (WGS84), as precise as provided in the original data source (numerical)
"Longitude" – Longitude (easting) of the site in decimal degrees (WGS84), as precise as provided in the original data source
"Elevation" – Elevation of the site in meters above sea level, if provided
"SourceGeogrData" – source of the geographic data ('Google maps' indicates that the locality was manually found on the digital map of Google maps)
"StartYear" – First year of sampling
"EndYear" – Last year of sampling
"Duration" – Time between first and last sample
"WWFecoRegion" – Original ecoregion according to WWF ecoregions of the world (https://www.worldwildlife.org/biome-categories/terrestrial-ecoregions)
"ClimaticZone" – Climatic zone by grouping of ecoregions: (factor with 4 levels: Boreal/Alpine, Terrestrial, Drylands, Tropical)
"ProtectedArea" – protection status of the site (yes - protected, or no – not protected)
"frcCrop_start" – Fraction of surrounding landscape (~25*25km) covered by crop land in the first year of sampling following the LUH2 database
"frcCrop_end" – Fraction of surrounding landscape (~25*25km) covered by crop land in the last year of sampling following the LUH2 database
"frcUrban_start" – Fraction of surrounding landscape (~25*25km) covered by urban land-use in the first year of sampling following the LUH2 database
"frcUrban_end" – Fraction of surrounding landscape (~25*25km) covered by urban land-use in the last year of sampling following the LUH2 database
"frcForest_start" – Fraction of surrounding landscape (~25*25km) covered by forest in the first year of sampling following the LUH2 database
"frcForest_end" – Fraction of surrounding landscape (~25*25km) covered by forest in the last year of sampling following the LUH2 database

"Urbanization" – Difference in fraction urban land cover between the first and last year of sampling (LUH2 Database)

"Cropification" – Difference in fraction crop land cover between the first and last year of sampling (LUH2 Database)

"frcCrop900m" – Fraction of the local landscape (900*900m) classified as crop land at the end of the sampling period following the ESA-CCI database. Land use codes classified as cropland were: 10, 11, 12, 30 and 40. Because 30 and 40 represent only partial crop cover, the number of cells with code 30 were multiplied by 0.75 and cells with 40 were multiplied by 0.25. Available only for sites where sampling ended in 1992 or later (n = 1567).

"frcUrban900m" – Fraction of the local landscape (900*900m) classified as urban (land use code 190) at the end of the sampling period following the ESA-CCI database. Available only for sites where sampling ended in 1992 or later (n = 1567).

"CRUmnC" - Mean temperature (Celsius) at the landscape scale (0.5° * 0.5°) over the sampled period, calculated from the CRU database for the full period

"CRUmnK" - Mean temperature (Kelvin) at the landscape scale (0.5° * 0.5°) over the sampled period (= CRUmnC + 273.16), calculated from the CRU database for the full period

"CRUdeltaTmean" – Modeled change in temperature per decade. We used a generalized additive model with a spline on month to derive the slope of temperature change for each site. The model estimate is based only on temperature data within the sampling period.

"CRUrelDeltaTmean" – Relative change in temperature for each site (= CRUdeltaTmean / CRUmnK)

"CRUmnPrec" – Mean monthly precipitation (mm) at the landscape scale (0.5° * 0.5°) over the sampled period, calculated from the CRU database for the full period

"CRUdeltaPrec" – Modeled change in monthly precipitation per decade. We used a generalized additive model with a spline on month to derive the slope of precipitation change for each site. The model estimate is based only on precipitation data within the sampling period.

"CRUrelDeltaPrec" – Relative change in precipitation for each site (= CRUdeltaPrec / CRUmnP)

"CHELSAmnC" - Mean temperature (Celsius) at the local scale (1 km^2) over the sampled period, calculated from the CHELSA database (= CHELSAmnK – 273.16). Available only for sites where sampling started after 1978 and ended latest in 2013 (n= 669 sites)

"CHELSAmnK" – Mean temperature (Kelvin) at the local scale (1 km^2) over the sampled period, calculated from the CHELSA database. Available only for sites where sampling started after 1978 and ended latest in 2013 (n= 669 sites).

"CHELSAdeltaTmean" – Modeled change in temperature per decade. We used a generalized additive model with a spline on month to derive the slope of temperature change for each site. The model estimate is based only on temperature data within the sampling period, and is available only for sites where sampling started after 1978 and ended latest in 2013 (n= 669 sites).

"CHELSArelDeltaTmean" – Relative change in temperature for each site (= CHELSAdeltaTmean / CHELSAmnK)

"CHELSAmnPrec" – Mean monthly precipitation (mm) at the local scale (1 km^2) over the sampled period, calculated from the CHELSA database. Available only for sites where sampling started after 1978 and ended latest in 2013 (n= 669 sites).

"CHELSAdeltaPrec" – Modeled change in monthly precipitation per decade. We used a generalized additive model with a spline on month to derive the slope of precipitation change for each site. The model estimate is based only on precipitation data within the sampling period, and is available only for sites where sampling started after 1978 and ended latest 2013.

"CHELSArelDeltaPrec" – Relative change in precipitation for each site (= CHELSAdeltaPrec / CHELSAmnP). Available only for sites where sampling started after 1978 and ended latest 2013 (n= 669 sites).

Table 'SampleData'

This table provides information on the data extraction and sampling methods used in the various data sources

Headers:

"Sample_ID" – Unique identifier for each sample description linking to the table 'finalData' (numerical 237 entries)

"DataSource_ID" – Unique identifier linking to table 'DataSources' (numerical 165 entries)

"DataCarrier" – Source of this part of the data (table, figure number)

"DataExtractionMethod" – software used to extract data from graphs or tables (factor with 4 levels: 'ImageJ', 'Metadigitise', 'pdftoexcel.com', 'values from owner')

"SamplingMethod" - Type of sampling method. Categorical with 27 levels:

- Artificial_substrate (insects were collected on artificially placed substrates after a set amount of time)
- Bagged_branches (insects were collected from branches after these were collected in a bag. The insect abundance is corrected for tree biomass sampled.)
- Bait (insects were attracted with group specific bait)
- Colored_pan_traps (flower visit insects attracted by colored bowls)
- Emergence_trap (insects trapped in flight after emergence)
- Hand_sorting (macro-invertebrates collected by hand from soil samples)
- Light_trap (nocturnal insects attracted by light)
- Malaise_trap (flying insects collected in a stationary tent-like net)
- Nest_counts (visual counts of ant nests)
- Pelagic_net (Apstein net - aquatic insects collected in the water column)
- Pitfall_traps (surface dwelling insects collected in containers in the soil with the rim flush with the soil surface)
- Sampling ring (vegetation dwelling insects visually counted after trapping them in a large sampling ring covering a standardized surface area)
- Soil_litter_extraction (Tullgren/Berlese funnel - soil and litter dwelling (micro)arthropods extracted using heat and/or light)
- Sticky_traps (insects collected on sticky cards suspended above the vegetation)
- Stream_bed_sampling (Surber/Hess/kick/Tee samplers - aquatic insects collected from the stream bed by disturbing a standardized area of the bed and letting the dislodged individuals drift into a net)
- Substrate_grab (Petersen/Eckman/Ponar grab – a standardized area of the substrate of a waterbody is taken and invertebrates are sorted out)

- Substrate_grab_and_stream_bed_sampling (both stream bed sampling and substrate grabing methods used)
- Substrate_scraping (aquatic invertebrates are scraped from a standardized surface area of substrate)
- Suction_pipe (flying insects are collected by sucking air into a stationary, upwards directed pipe)
- Suction sampling (vegetation dwelling insects are collected by sucking them into a using motorized suction machine of a standardized surface area)
- Sweep_net (vegetation-dwelling or aquatic insects are collected in a sweep net using a standardized number of sweeps or sampling area)
- Timed_counts (insects are visually counted for a standardized amount of time)
- Transect_counts (e.g. Pollard walks – insects are counted along standardized transects)
- unknown_standardized_methods (the trapping method is unclear, but is standardized over time
- Vertical_radar (upward directed radar to count the number of flying insects
- Visual_counts (insects visually counted in a standardized area)
- Window_trap (flying insects collected as they fly against a window-like structure with a collection vessel underneath it)

"SamplingMethodDetailed" – Method the invertebrates were sampled (as described in the original publication)

"Stratum" – Place in which the insects were sampled: (factor with 6 levels:

- 'Air' (transect counts, suction pipes, light traps, malaise traps, window traps and pan traps)
- 'Trees' (arboreal window traps, visual counts, sticky traps)
- 'Herb layer' (sweep-net transects, suction sampling, sampling rings, visual counts)
- 'Soil surface' (pitfall traps, and ant nest counts)
- 'Underground' (soil cores)
- 'Water' (e.g. kicksamplers, Surber samplers, aquatic emergence traps).

"SampleArea" – Surface area of sample, where applicable. 'NA' indicates that the size of the sampling area is unknown, as is the case for activity dependent methods (such as pitfall traps, light traps and malaise traps)

"NumberOfReplicates" – Number of replicates which constitute one sample. Often not clearly described in the original publication.

"AggregationOfReplicates" – How these replicates are merged to produce the reported value. Often not clearly described in the original publication.

"GroupInData" – Invertebrate group represented in the data carrier (factor with 95 levels)

"OriginalMetric" – Metric of assemblage size as reported in the original data carrier

"Calculations" – when applicable any calculations done by us to standardize the data. For example, standardization of sampling effort or inverse log-transformations.

"Metric" – Standardized metric of assemblage size (factor with 3 levels: 'biomass', 'abundance', 'density' = abundance per unit area)

Table 'InsectAbundanceBiomassData'

This table contains the measured insect abundances at each of the sites over time.

"DataSource_ID" – Unique identifier linking to table 'DataSources', and to table 'PlotData'

"Plot_ID" – Unique identifier for site, linking to table 'PlotData'
"MetricAB" – Metric of assemblage size: abundance or biomass
"Period" – Period in the year of sampling. This can be month, season, etc. The finest grain is 'month'. This variable was used as random effect in the analysis to account for seasonality.
"Stratum" - Place in which the insects were sampled: factor with 6 levels, see SampleData for details. (categorical)
"Year" – year in which measurement was taken
"Number" – Value for insect abundance or biomass as measured at a given time and place. NA's are retained in place, as this was required for our INLA analysis. These are easily removed.

Table 'References'
This table contains all references referred to in any of the tables