

# Data Dictionary for /data Directory

This document provides detailed descriptions of the files included in the /data directory of the GitHub repository <https://github.com/likun-stat/XVAE>, to support reproducibility of the analyses presented in the manuscript. All files are in R's .rda binary format and are intended to be loaded using R.

## **example\_X.rda**

**Type:** Simulated (artificial) data

Generated by: ./Simulation Study/Autodiff\_w\_torch\_alpha\_fixed\_half (optim), corresponding to Model III in the manuscript.

**Contents:**

- `x`:  $2000 \times 100$  numeric matrix. Rows represent spatial locations; columns represent simulated time replicates.
- `stations`:  $2000 \times 2$  data frame. Geographic coordinates (longitude, latitude in degrees) for each location in `x`.

## **copulas.rda**

**Type:** Simulated (artificial) data for emulation comparison

**Contents:**

- `U_sim_grid`:  $480 \times 500$  data frame of simulated data (uniform scale).
- `U_xvae_grid`:  $480 \times 500$  data frame of XVAE-emulated data (uniform scale).
- `U_gan_grid`:  $480 \times 500$  data frame of extGAN-emulated data (uniform scale).
- `stations_grid`:  $2000 \times 2$  data frame. Geographic coordinates (longitude, latitude) for each location.

## **mon\_max\_allsites.rda**

**Type:** Real observed data

**Contents:**

- `mon_max_allsites`:  $16703 \times 372$  data frame. Rows = locations, columns = monthly max temperatures ( $^{\circ}\text{C}$ ) over 31 years.

## **new\_loc.rda**

**Type:** Real observed data (location metadata)

**Contents:**

- `new_loc`:  $16703 \times 2$  data frame. Longitude and latitude (degrees) for locations in `mon_max_allsites`.

## **fitted\_gev\_par.rda**

**Type:** Real observed data (fitted model parameters)

**Contents:**

- `fitted_gev_par`:  $16703 \times 4$  data frame. Each row corresponds to a location, with columns:
  1. Location: GEV location parameter ( $^{\circ}\text{C}$ )
  2. Scale: GEV scale parameter ( $^{\circ}\text{C}$ )
  3. Shape: GEV shape parameter (unitless)
  4. P-value: Chi-squared goodness-of-fit test p-value (unitless, between 0 and 1)

**Notes:**

- All .rda files are in R's native binary format and can be loaded using the `load()` function.
- Artificial data sets were generated independently and used for validation or emulation method comparison.
- Real datasets were used in the manuscript's application and model validation sections.