This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Part 1: Data

☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

The dataset consists of observed daily precipitation data in millimeters (mm) from 125 monitoring stations in the Danube river basin (Europe) and from 2229 monitoring stations in the Mississippi river basin (North America) over the period from 1965 to 2020. The temperature covariate in Celsius degree used to fit the model in this project was derived from the ERA5-Land reanalysis data for the corresponding region we selected, which is a global land-surface dataset with a high spatial resolution of 9km, and the projected temperature covariate is derived from climate models outputs of the sixth Coupled Model Intercomparison Project (CMIP6), namely AWI, MIROC, and MPI.

Availability

☒ Data **are** publicly available.
☐ Data **cannot be made** publicly available.

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available data* section, below.

Publicly available data

☒ Data are available online at:

- Climate model outputs and EAR5-Land Reanalysis data: https://www.copernicus.eu/en/access-data
- Precipitation observations: https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily
- Watershed Boundary Dataset: https://www.usgs.gov/national-hydrography/watershed-boundary-dataset
- The data files for plots are stored in Github repository: https://github.com/PangChung/ExtremePrecip/

☐ Data are available as part of the paper's supplementary material.

☐ Data are publicly available by request, following the process described here:

☒ Data are or will be made available through some other mechanism, described here:

- We have saved the data used in our marginal modeling and dependence modeling in the format of R data file (.RData), and we host those data on Github[https://github.com/PangChung/ExtremePrecip] under the folder "data' ', which are public accessible. Though the raw data of the EAR5-Land Reanalysis data from which we derived the temperature covariate in Celsius degree can be downloaded from the website https://www.copernicus.eu/en/access-data, this raw data will be provided upon email request as the size of it is hundreds of gigabytes and it will take great efforts for someone to download directly from the Copernicus website. The email address is peng[dot]zhong[at]unsw[dot]edu[dot]au.

Non-publicly available data

Description

File format(s)

☐ CSV or other plain text.
☒ Software-specific binary format (.Rda, Python pickle, etc.): .Rda (.RData)
☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
☐ Other (please specify):

Data dictionary

☒ Provided by authors in the following file(s):

- data/precip.RData: The precipitation data for the eight subregions, which contain R objects:
  - precip: Lists of lists, raw data of the precipitation in millimeters (mm) in 8 subregions. This is the response variable used in the marginal fit.
  - region.name: a vector, names of the 8 subregions
  - region.id: a vecotr, regional ID number that corresponding to the regional number in the shape files.
  - station: a data frame, contains geograhical information about the monitoring stations, where the precipitation data were recorded. "X" column is latitude degrees, "Y" column is the longitude in degrees, "start" column is the start measuring year, "end" is the last measureing year, "elev" is the elevation of the monitoring stations in meters, and "group.id" corresponds to the region.id for identifiying each of the 8 subregions.
  - START.date and END.date: dates, between which the data were used in our analysis.
- data/temperature.RData: The derived temperature covariate over the period 1965–2020.
  - date.df: data frame, contains the date between 1965–2020, and its corresponding season of the year.
  - temperature.covariate: list of vectors, the derived temperature covariate (30 day moving averages) used in marginal modeling and dependence modeling.
  - temperature: list of vectors,daily spatial temperature avarages in Celsius degree, used only during data preprocessing.
  - loc_df: geographical information about the locations of the ERA5-Land temperature data, used only during data preprocessing.
- data/temperature_pred.RData: The derived temperature covariate from the climate models over the period 2015–2100 under different shared socioeconomic pathways (SSP 2-4.5 or SSP 5-8.5).
  -
  -
  -
  -
- data/marginal_fit_quantiles.RData: Transformed margins based on the marginal fit.
- data/dep.fit.boot.results3.RData: Fitted results from the bootstrap scheme for the dependence model.
- data/era5_geoinfo.RData: Shape files for the eight subregions.
- data/transformed_coordinates.RData: Transformed coordinates for the eight subregions, which transformed the latitude/longitude coordinate to the Eculidean coordinate.

☐ Data file(s) is(are) self-describing (e.g., netCDF files)
☒ Available at the following URL:

- Github: https://github.com/PangChung/ExtremePrecip/

Additional Information (optional)

Part 2: Code

Abstract

The code contains all the code to process the raw data (upon request) and generate all the figures and tables in the main manuscript and the supplemental material from the .RData files provided. The R script file code/bootstrap.R is used to fit the marginal model as well as the dependence model when the variable bootstrap.ind = 0, otherwise, it will fit the marginal model together with the dependence model to the bootstrap data when the variable bootstrap.ind ! = 0. One can use the bash script code/bootstrap.sh to fit model in parallel on a computer clusters or HPC system running PBS (a HPC management software). To generate all the figures, one can use and follow the R script code/plots.R.

Description

Code format(s)

- ☒ Script files
  - ☒ R
  - ☐ Python
  - ☐ Matlab
  - ☐ Other:
- ☒ Package
  - ☒ R
  - ☐ Python
  - ☐ MATLAB toolbox
  - ☐ Other:
- ☐ Reproducible report
  - ☐ R Markdown
  - ☐ Jupyter notebook
  - ☐ Other:
- ☒ Shell script
- ☐ Other (please specify):

Supporting software requirements

- de Fondeville R, Belzile L (2023). *mvPot: Multivariate Peaks-over-Threshold Modelling for Spatial Extreme Events*. R package version 0.1.6,https://CRAN.R-project.org/package=mvPot.
- Youngman BD (2022). "evgam: An R Package for Generalized Additive Extreme Value Models." *Journal of Statistical Software*, *103*(3), 1-26. doi:10.18637/jss.v103.i03 https://doi.org/10.18637/jss.v103.i03.

Version of primary software used

- R version 4.3.2
- OpenPBS 23.06.06

Libraries and dependencies used by the code

- R packages:
  - mvPotST (a spatial-temporal version of mvPot version 0.1.6, located at code/archived/mvPotST/mvPotST_0.0.0.900 in the Github repository.)
  - evgam version 1.0.0
  - mgcv version 1.9-0
  - evd version 2.3-6.1

  - nlme version 3.1-163
  - lubridate version 1.9.3
  - ggplot2 version 3.5.0
  - ggpubr version 0.6.0
  - gridExtra version 2.3

Supporting system/hardware requirements (optional)

Access to HPC with OpenPBS management software.

Parallelization used

- ☐ No parallel code used
- ☐ Multi-core parallelization on a single machine/node
  - Number of cores used:
- ☒ Multi-machine/multi-node parallelization
  - Number of nodes and cores used: 56 nodes with 16 cores on each nodes

License

- ☒ MIT License (default)
- ☐ BSD
- ☐ GPL v3.0
- ☐ Creative Commons
- ☐ Other: (please specify)

Additional information (optional)

Part 3: Reproducibility workflow

Scope

The provided workflow reproduces:

- ☐ Any numbers provided in text in the paper
- ☒ The computational method(s) presented in the paper (i.e., code is provided that implements the method(s))
- ☒ All tables and figures in the paper
- ☐ Selected tables and figures in the paper, as explained and justified below:

Workflow

Location

The workflow is available:

- ☐ As part of the paper's supplementary material.
- ☒ In this Git repository: https://github.com/PangChung/ExtremePrecip
- ☐ Other (please specify):

Format(s)

- ☐ Single master code file
- ☒ Wrapper (shell) script(s)
- ☐ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach
- ☐ Text file (e.g., a readme-style file) that documents workflow
- ☐ Makefile

☐ Other (more detail in *Instructions* below)

Instructions

- First, one need to specify the working directory, which should be *~/ExtremePrecip/* in our case
- Then, one need to install the packages mvPotST as well as other packages listed in the R script files.
- Run the marginal fit and the dependence model fit via the following code, *Rscript code/bootstrap.R "idx.region=1;bootstrap.ind=0;computer="local""*. Here we fit the model for the region 1 (Danube) by setting *idx.region=1*.
- Generate the plots using the code in *code/plots.R*, in some cases, one can directly load the plots into memory by load the R Data file data/plot_<name of the plot>.RData
- Other R scripts are used to pre-process the raw data (provided upon request).

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

☐ < 1 minute
☐ 1-10 minutes
☐ 10-60 minutes
☐ 1-8 hours
☐ > 8 hours
☒ Not feasible to run on a desktop machine, as described here: We use a computer cluster with 56 nodes to do the bootstrap. However, one can still use a multicore workstation to fit the model for individual subregions within several hours.

Additional information (optional)

Notes (optional)