The Journal of Open Source Software

geslaR: An R package to deal with the GESLA dataset

Fernando P. Mayer [1*¶] **and Niamh Cahill** [1*]

1 Department of Mathematics and Statistics, Maynooth University, Ireland ¶ Corresponding author *
These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

The GESLA (Global Extreme Sea Level Analysis) project aims to provide a global database of higher-frequency sea-level records for researchers to study tides, storm surges, extreme sea levels, and related processes. Three versions of the GESLA dataset are available for download, including a zip file containing the entire dataset, a CSV file containing metadata, and a KML file for plotting the location of all station records in Google Earth. The web application developed here facilitates the access of the GESLA dataset by providing basic filters to select the data of interest, generating informative plots, and showing the selected sites all over the world. Users can download the selected subset of data in CSV or Parquet file formats, with the latter being recommended due to its smaller size and the ability to handle it in many programming languages through the Apache Arrow language for in-memory analytics. The web interface was developed using the Shiny R package, with the CSV files from the GESLA dataset converted to the Parquet format and stored in an Amazon AWS bucket.

Statement of need

The first version of the GESLA dataset, denoted **GESLA-1** was made available in 2009 (Menéndez & Woodworth, 2010). An updated second version, the **GESLA-2**, was assembled in 2015 and 2016, and released in 2016 (Woodworth et al., 2017). The latest version, the **GESLA-3**, was compiled in 2020 and 2021, and it is the one made available at the official GESLA website for download, since November 2021 (Haigh et al., 2021).

There are three files available for download, which, following the website, are:

- A **zip** file containing the entire GESLA-3 dataset. The size of the zip file is 5.15 GB and when unzipped the data files are a total of 38.72 GB
- A **csv** file containing the meta-data (e.g., station names, code, latitudes and longitudes, etc)
- A **KML** file for plotting the location of all the station records in google earth

Although the GESLA-3 dataset is freely available, the way it was distributed could be a hurdle for many researchers. Anyone interested in using the data for one (or more) specific locations, should have to download all the CSV files, and then search for the desired data, and possibly even merging several of them. The GESLA project also provides two auxiliary scripts (in Python and Matlab), to help researcher to handle and merge the desired subset of data. However, it is still needed to download all the dataset and to know in advance the files needed.

The goal of this application is to facilitate the access to the GESLA dataset, without the need to download and handle all the files, which may be computationally infeasible due to it's large size. By accessing the interface, the user has available three basic filters to select the data of interest: country, year and site. The user can select one or more options in each of these filters, and the resulting subset will be made available through a sample of the subset (the

"Data preview" tab) (Figure 1), some basic informative plots (the "Plots" tab), and a map showing the selected sites all over the world (the "Map" tab). Upon selection, the user can download only that subset of data. The application is available at this link. The GESLA-3 dataset files were converted to the Parquet file format, for easy of storage and manipulation, and is hosted in an Amazon AWS bucket. However, the user can download any subset of data in CSV or Parquet.

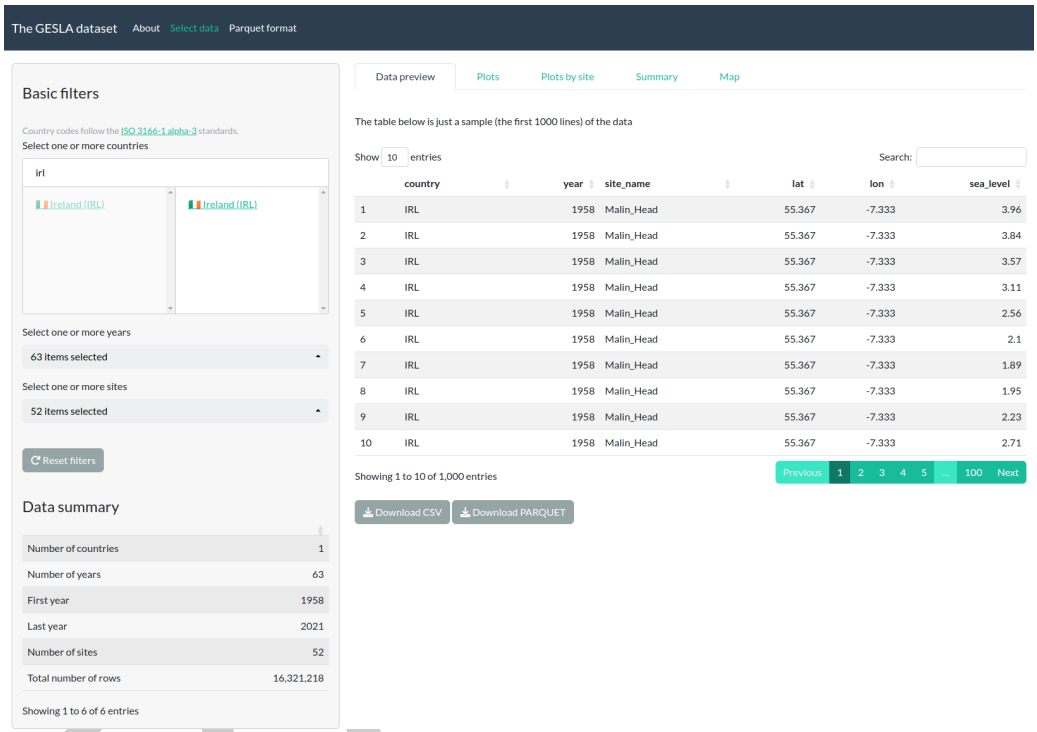

Figure 1: The fisrt page of the geslaR application, showing the "Data preview" tab after basic filtering.

Acknowledgements

This work has emanated from research conducted with the financial support of Science Foundation Ireland and co-funded by GSI under Grant number 20/FFP-P/8610. We also would like to thank the Hamilton Institute at Maynooth University for hosting the application.

References

Haigh, D., Marcos, M., Talke, S. A., Woodworth, P. L., Hunter, J. R., Hague, B. S., Bradshaw, E., & Thompson, P. (2021). GESLA version 3: A major update to the global higher-frequency sea-level dataset. *EarthArXiv*. https://doi.org/https://doi.org/10.1002/gdj3.174

Menéndez, M., & Woodworth, P. L. (2010). Changes in extreme high water levels based on a quasi-global tide-gauge data set. *Journal of Geophysical Research: Oceans*, *115*. https://doi.org/10.1029/2009JC005997

Woodworth, P. L., Hunter, J. R., Marcos, M., Caldwell, P., Menéndez, M., & Haigh, I. (2017). Towards a global higher-frequency sea level dataset. *Geoscience Data Journal*, *3*, 50–59. https://doi.org/10.1002/gdj3.42