U. S. Army Engineer Research and Development Center
Coastal and Hydraulics Laboratory

StormSim-SST:

Stochastic Storm Simulation System’s Stochastic Simulation Technique

Quick Start Guide

Version 0.5 Alpha
August 9, 2021

Table of Contents

1 Content of the StormSim-SST Tool package

- MATLAB script Interface_StormSim_SST_Tool.m
    - Principal script that runs the StormSim-SST Tool. Allows users to define inputs and view outputs. Includes an example of the input arguments to evaluate a time series dataset from a NOAA gauge station.
- File StormSim_SST_Tool.p is the StormSim-SST Tool.
- Folder “INPUT” contains a .mat file with the example input data:
    - InputData.mat – time series dataset
- Folder “SST_output” has the output generated for the example input file.
- Quick start guide (this document)
- Feedback form

2 Outline of the script Interface_StormSim_SST.m

The script is divided in four sections, as follows. To run the tool with the provided example just click the “Run” button (green triangle in the Editor tab). The expected output can be found in the folder SST_output.

1. **%% HEADER**
- Contains the disclaimer, licensing details, authors, and history of revisions.

2. **%% Input data files**
- This is an example on how to load the .mat files with the input data.

3. **%% Transfer of input data**
- This is an example on how to transfer the input dataset into the required structure array. Refer to Section 6 for more details.

4. **%% Dataset(s) information**
- This is an example on how to enter a description for the dataset. Refer to Section 6 for more details.

5. **%% General settings**
- The general settings control the following:
    - mode of execution
    - type of input data
    - confidence limits to compute
    - evaluation of small samples (< 20 peak events)
    - the Mean Residual life method’s threshold selection criteria to apply
    - hazard definition: frequencies or probabilities
    - enable use of parallelized code
    - path to output folder
    - vector of responses to interpolate hazard from curve
    - removal of flag values

    - parameters for computing the peaks over threshold sample (applicable when the input data is time series)
    - record length (applicable when the input data is peaks over threshold sample)

6. **%% Plot settings**
   - Plot settings to control the view and label of hazard plots.

7. **%% USER: DO NOT CHANGE ANYTHING BELOW THIS LINE**
   - In this section, the StormSim_SST_Tool.p file is called to perform the evaluation with the user-provided input arguments.

3 Description of the StormSim-SST Tool

The StormSim's Stochastic Simulation Technique (StormSim-SST) tool estimates the hazard curve (HC) of one or multiple response datasets. A dataset can be entered in the form of either a Peaks-Over-Threshold (POT) sample or a raw time series dataset (e.g., gauge observations). The underlying framework was built upon the univariate formulation of the Empirical Simulation Technique (Scheffner et al. 1999; Borgman et al. 1992; Scheffner, Borgman, and Mark 1994), with improved methodologies for a robust characterization of the low frequency tail of a HC. The main features of the SST tool are:

- bootstrap-based methodology
- fits a generalized Pareto distribution (GPD) function to complete the distribution's low-frequency tail, while using the empirical distribution for the high-frequency tail
- employs the mean residual life (MRL) automated threshold detection method (Langousis et al. 2016) for the objective selection of the GPD threshold parameter; and
- can express the hazard as either annual exceedance frequency (AEF) or annual exceedance probability (AEP).

4 Modes of execution

The tool offers two modes of execution for the user: 'Regular' and 'Fast'. These modes are specified in ExecMode (Section 6) and have the following implications:

1. 'Regular': under this mode, the tool will return all available outputs and plots for each input dataset. This provides the opportunity to test different setting values and their impact on the HC, so the user can decide which values are best per dataset.
2. 'Fast': this mode reduces the output information to increase execution time and decrease output size for a straight forward application of several datasets. The user is advised to use this mode after confirming the best settings using the 'Regular' mode.

A description of each input and output argument is presented in sections 6 and 7, respectively.

Production of plots depends on the selected execution mode, as follows:

1. 'Regular': when ExecMode = 'Regular', the tool will produce histogram plots of the GPD shape parameter values, plots of the HC for each GPD threshold identified by the MRL method, and a stacked plot of the MRL method results.
2. 'Fast': when ExecMode = 'Fast', no plots will be produced for faster execution.

The tool will store all available outputs on a .mat file named StormSim_SST_output.mat. This file will be saved inside the folder directory specified in path_out (Section 6). Some of the results are organized in a structure array with a different set of fields depending on the mode of execution selected by the user, as follows:

1. 'Regular': when ExecMode = 'Regular', the output structure array will contain the station information (staID); the record length (RL); the POT sample (POT); the outputs of the MRL method (MRL_out); the HC as plotted (HC_plt) for the values in HC_plt_x; the summarized HC (HC_tbl) for the values in HC_tlb_x; the summarized HC (HC_tbl_rsp_x) for the values in HC_tbl_rsp_y; the empirical HC (HC_emp); a warning message (Warning); and an error message when the tool failed to evaluate a dataset (ME).
2. 'Fast': when ExecMode = 'Fast', the output structure array will only contain the station information (staID); the HC as plotted (HC_plt) for the values in HC_plt_x; the summarized HC (HC_tbl) for the values in HC_tlb_x; the summarized HC (HC_tbl_rsp_x) for the values in HC_tbl_rsp_y; a warning message (Warning); and an error message when the tool failed to evaluate a dataset (ME).

5 Types of applications

Current design of the StormSim-SST Tool enables the evaluation of general and case-specific applications. A general application considers the use of raw time series datasets or PDS/POT datasets of any type of parameter. Refer to Section 6 for a description of each input. Settings for general applications are as follows:

- A1 - For raw time series dataset with regular-mode setting: DataType = 'Timeseries' and ExecMode = 'Regular'. This is intended for a detailed evaluation of a dataset, when there is no knowledge of the input arguments values.
- A2 - For raw time series dataset with fast-mode setting: DataType = 'Timeseries' and ExecMode = 'Fast'. This is intended for the evaluation of multiple datasets.
- A3 - For POT dataset with regular-mode setting: DataType = 'POT' and ExecMode = 'Regular'. Same as A1 but using a POT dataset.
- A4 - For POT dataset with fast-mode setting, DataType = 'POT' and ExecMode = 'Fast'. Same as A2 but using a POT dataset.

Available case-specific applications are described below:

- A5 – ADCIRC-simulated hydrodynamic response of tropical cyclones
    - The tool can evaluate the peak hydrodynamic response from ADCIRC model output of tropical cyclone (TC) simulations.
    - In this case, the peak responses dataset is treated as a POT sample.
    - ADCIRC simulation of TCs usually incorporates astronomical tides and/or sea level rise.

- The astronomical tide is a stochastic process which is uncorrelated to the TC parameters but can be represented in the form of a skew tide. The skew tide is computed as the difference between the with-tide (e.g., still water level) and without-tide hydrodynamic response (e.g., surge) as simulated by ADCIRC (Figure 1).

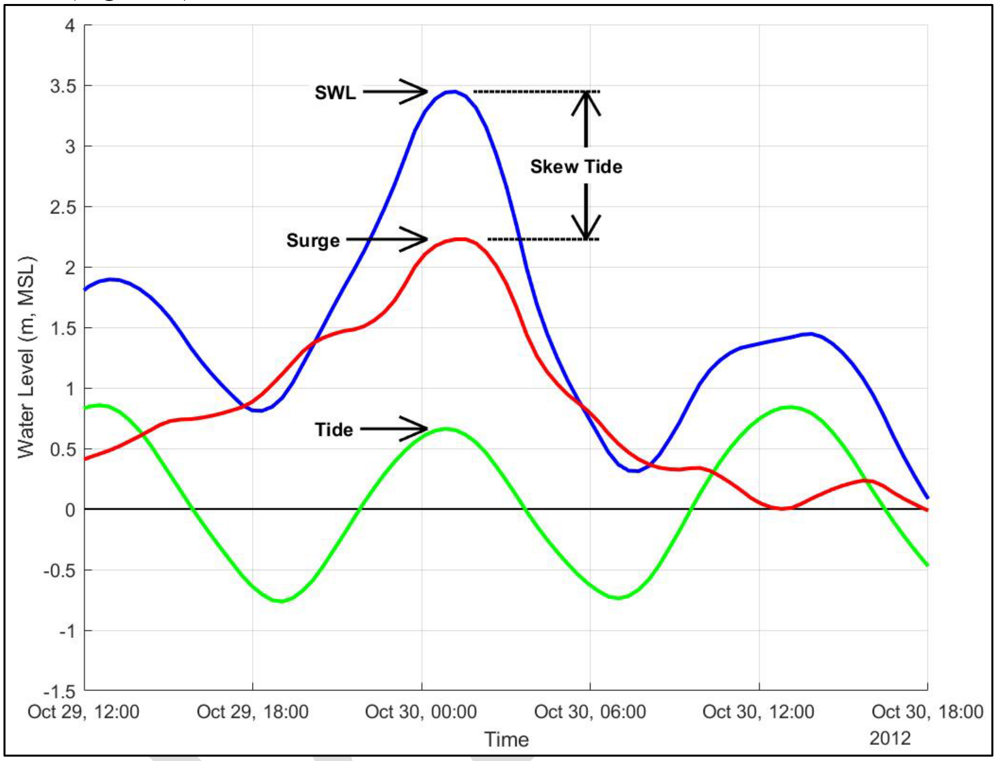

Figure 1: Idealized example of water level components.

- For a proper characterization of the HC with the SST tool, however, neither the tides nor the sea level rise (SLC) can be implicit in the peak response dataset. The SST workflow removes the SLC contribution before applying the bootstrap process to the without-tides responses. Afterwards, both the SLC and the skew tide are added to the resulting 1000 synthetic empirical samples. The skew tides are randomly added via Gaussian Process Regression (GPR) modeling. This requires the use of MATLAB's GPR model (The Mathworks 2021), a probabilistic model that can be trained to replicate the relationship between skew tides and a response parameter (e.g., surge) from available ADCIRC model output. This approach offers the opportunity to predict the skew tides for a new set of response values at a virtual gage and randomly incorporate them in the SST analysis. However, training of GPR models requires a separate analysis not performed by the SST tool.
- For this type of application in the SST tool, the following must be provided:
    - Peak response dataset without the contribution of tides (without-tide ADCIRC model output).
    - Second peak response dataset with the contribution of tides (with-tide ADCIRC model output). Needed to replace the empirical distribution

        - (without-tide peaks response) after the bootstrap process is done. Also used to compute the GPD threshold and allow an apples-to-apples comparison in the hazard plots. Must be entered in the input argument input_data.data_values2.
    - Value of SLC (**without steric adjustment**) used in the ADCIRC simulation. Set input argument SLC = 0 when no SLC applies.
    - GPR model trained on skew tide vs. response per virtual gage
        - Must be entered in the input argument gprMdl.mdl.

6 Description of input arguments

The following is a description of the input arguments and available settings of the StormSim-SST tool.

6.1 Input data options

- input_data: raw time series datasets or POT samples to be evaluated; specified as a structure array with fields time_values, data_values and data_values2. Each record of input_data must correspond to an input dataset. Use as follows:
    - When DataType = 'Timeseries':
        - input_data.time_values: must have the timestamps as serial date numbers (see 'datenum' in the MATLAB Documentation)
        - input_data.data_values: must have the time series data values
    - When DataType = 'POT':
        - input_data.time_values can be empty [ ]
        - input_data.data_values must have the POT values
        - input_data.data_values2 can be empty [ ]
    - When ind_Skew = 1: (evaluating POT storm surge with skew tides)
        - DataType = 'POT'
        - input_data.time_values can be empty [ ]
        - input_data.data_values must have the POT storm surge values without tides
        - input_data.data_values2 must have the POT storm surge values with tides

- staID: gauge station information for each input dataset; specified as a cell array with format:
    - Col(01): gauge station ID number; as a character vector. Example: '123456'
    - Col(02): station name (OPTIONAL); as a character vector. Example: 'Dataset_1'

6.2 General settings

- ExecMode: mode or type of execution. Available options are 'Regular' or 'Fast'; specified as a character vector. Example: ExecMode = 'Regular'.

- DataType: indicator for specifying the type of input dataset. Current options are:
    - 'POT' for a Peaks-Over-Threshold sample
    - 'Timeseries' for a raw time series data set

- use_AEP: indicator for expressing the hazard as AEF or AEP. Use 1 for AEP, 0 for AEF. Example: use_AEP = 1;

- GPD_TH_crit: indicator for specifying the GPD threshold option of the Mean Residual Life (MRL) selection process; specified as a scalar. Use as follows:
    - GPD_TH_crit = 0: to evaluate all thresholds identified by the MRL method (one per criterion)
    - GPD_TH_crit = 1: to only evaluate the MRL threshold selected by the lambda criterion
    - GPD_TH_crit = 2: to only evaluate the MRL threshold selected by the minimum WMSE criterion

- prc: percentage values for computing the percentiles; specified as a scalar or vector of positive values. Leave empty [ ] to apply default values 2.28%, 15.87%, 84.13%, 97.72%. User can enter 1 to 4 values. Example: prc = [2 16 84 98];

- HC_tbl_rsp_y: response values used to summarize the HC; specified as a numerical vector of positive values. Set as empty [ ] to apply default values (0 to 20). Example: HC_tbl_rsp_y = [0:05:10];
    - Corresponding probabilities interpolated from the HC will be stored in HC_tbl_rsp_x.

- apply_GPD_to_SS: indicates if the hazard of a small POT sample should be evaluated with either the empirical or the GPD. Here, a small sample is that with either a sample size < 20 events or a record length < 20 years.
    - Specify GPD_to_SS = 1 to apply GPD
    - Specify GPD_to_SS = 0 to use the empirical distribution.

- apply_Parallel: Enter one (1) to perform the evaluation in parallel; enter zero (0) otherwise.

- path_out: path to output folder; specified as a character vector. Leave empty [ ] and the tool will create a folder named SST_output in the current directory.
    - The tool will search for the default SST_output output folder and will not create another one if it already exists.

- flag_value: any flag value to search and remove from the input dataset; specified as a scalar. Leave empty [ ] otherwise. Example: flag_value = -999;

- tLag: inter-event time in hours; specified as a positive scalar or vector. Use as follows:
    - When DataType = 'POT': leave empty [ ].
    - When DataType = 'Timeseries': Enter one value per dataset or one value for all datasets. Examples: tLag = [48 24]; tLag = 48;

- lambda: mean annual rate of events (events/year); specified as a positive scalar or vector. Use as follows:
  - When DataType = 'POT': leave empty [ ].
  - When DataType = 'Timeseries': Enter one value per dataset or one value for all datasets. Examples: lambda = [12 6 3]; lambda = 12;

- Nyrs: record length in years; specified as a positive scalar or vector. Use as follows:
  - When DataType = 'POT': enter one value per dataset or one value for all datasets. Examples: Nyrs = [50 15 30]; Nyrs = 50;
  - When DataType = 'Timeseries': leave empty [ ] since the tool will compute the effective duration.

6.3 Plot settings

- yaxis_Label: parameter name, units and vertical datum for label of the plot y-axis; specified as a cell array of character vectors. Available options for this input are as follows:
  - Enter one label that applies to all input datasets. Example: yaxis_Label = {'Still Water Level (m, MSL)'};
  - Enter one label per input dataset. Example for three input datasets: yaxis_Label = [{'Still Water Level (m, MSL)'},{'Significant wave height (m)'},{'Peak wave period (s)'}];
  - When ExecMode = 'Fast', leave empty [ ].
  - When ind_Skew = 1 and the input data is storm surge, set yaxis_Label = 'Still water level' with associated units and datum. Example: 'Still Water Level (m, MSL)'

- yaxis_Limits: lower and upper limits for the plot y-axis; specified as a two-column numerical array. Available options for this input are as follows:
  - Enter a two-value vector that applies to all input datasets. Example: yaxis_Limits = [0 10];
  - Enter as a two-column matrix with one row per input dataset. Example for three input datasets: yaxis_Limits = [[0 10];[2 10];[0 20]];
  - Leave empty [ ] to apply default values.
  - When ExecMode = 'Fast', leave empty [ ].

6.4 Skew tide settings

- ind_Skew: indicator for computing/adding skew tides to the response dataset. Use as follows:
  - ind_Skew = 1 for the tool to compute and add the skew tides
  - ind_Skew = 0 otherwise

- SLC: magnitude of the sea level change implicit in the simulated response dataset; specified as a positive scalar.
  - When ind_Skew = 1: must have same units as the input POT storm surge dataset. Example: SLC = 1.8;

- When ind_Skew = 0: leave empty [].

- gprMdl: GPR model created with the MATLAB function 'fitrgp'; specified as a structure array with field 'mdl'.
  - When ind_Skew = 1: Field 'mdl' must contain one GPR object per input dataset included in input_data. Refer to MATLAB Documentation for 'fitrgp'. Train the GPR with the response data (e.g., storm surge) as a predictor and skew tides as the response.
  - When ind_Skew = 0: Leave empty [ ].

7 Description of output arguments

The following is a description of the output arguments provided by the StormSim-SST tool.

- HC_plt_x: predefined vector of probabilities used to plot the HC. The type is:
  - AEP when use_AEP = 1
  - AEF when use_AEP = 0

- HC_tbl_x: predefined vector of probabilities used to summarize the HC. The type is:
  - AEP when use_AEP = 1
  - AEF when use_AEP = 0

- HC_tbl_rsp_y: predefined vector of response values used to summarize the HC.

- Removed_datasets: Message in the form of a cell array, with the list of datasets not evaluated when:
  - Dataset resulted empty after removing flag, zeros, NaN and/or Inf values.
  - Dataset has less than 3 unique values.

- Check_datasets:
  - When ind_Skew = 1: Message with a list of the input datasets to which the skew tides were not applied. For example: 'These stations were evaluated without applying skew tides since ind_Skew = 1, but entered invalid values for either the second POT dataset (with tides) or the GPR model'.
  - When ind_Skew = 0: will be empty.

- SST_output: structure array containing the output data of the StormSim-SST Tool. The number of fields in this array will change depending on the value of input ExecMode.
  - When ExecMode = ‘Regular’, the tool stores the full array of details inside argument SST_output, resulting in 9 fields (staID, RL, POT, MRL_output, HC_plt, HC_tbl, HC_tbl_rsp_x, Warning, ME)
  - However, when GPD_TH_crit = ‘Fast’, argument SST_output will only contain 5 of the fields (staID, HC_plt, HC_tbl, HC_tbl_rsp_x, Warning, ME).

  The following is a description of each available field:
  - staID: identifier of the gauge station as specified by the user

- RL: record length in years
- POT:
  - When DataType = 'Timeseries', this is the POT sample computed with the tool; as a matrix array with format:
    - col(01): time of POT values in serial date number format
    - col(02): POT values
    - col(03): data time range used for selection of POT value: lower bound
    - col(04): data time range used for selection of POT value: upper bound
  - When DataType = 'POT', this is the same input POT dataset.

- MRL_output: output of the Mean Residual Life (MRL) threshold selection process. The information inside this array will change depending on the input POT sample size. When the sample has less than 10 values, the MRL process is not applied and a default GPD threshold is computed. This will be indicated in the field eMsg. Otherwise, a successful application of the MRL will return the output MRL_output as a structure array with the following fields:
  - Summary = summary of the threshold selection results, as a table array with format:
    - Threshold: list of threshold values evaluated
    - MeanExcess: mean excess
    - Weight: weights
    - WMSE: weighted mean square error (MSE)
    - GPD_Shape: GPD shape parameter
    - GPD_Scale: GPD scale parameter
    - Events: number of events above each threshold
    - Rate: sample intensity (mean annual rate of events using sample of events above threshold)

    Note: When sample size < 20, array Summary will consist of a row of NaN values.
  - Selection = selected threshold with other parameters, as a table array with format:
    - Criterion: criterion applied by MRL method for selecting the threshold
    - Threshold: selected threshold value
    - id_Summary: location (row ID) of selected threshold in the Summary field
    - Events: number of events above the selected threshold
    - Rate: sample intensity (mean annual rate of events using sample of events above threshold)

    Note: When sample size < 10, array Selection will only have the following information:
      - Criterion = ‘Default’

- ❖ Threshold = the actual threshold used in the tool will be computed for each bootstrap iteration as 0.99 times the minimum value of the random sample. The sample stored here will be the average of those thresholds.

  - pd_TH_wOut = GPD threshold parameter values used in the bootstrap process
  - pd_k_wOut = initial GPD shape parameter values obtained in the bootstrap process
  - pd_sigma = GPD scale parameter values used in the bootstrap process
  - pd_k_filled = modified GPD shape parameter values used in the bootstrap process
  - eMsg = Status message indicating when the MRL methodology was not able to objectively determine a threshold for the GPD. In this case, the GPD threshold is set to 0.99 times the minimum value of the bootstrap sample. Otherwise, eMsg will be empty.

- HC_plt: full HC. The format of this field depends on the value of ind_TH_mrl. When ind_TH_mrl = 1 or 2, HC_plt is a numerical array with the 5 rows mentioned below. However, specifying ind_TH_mrl = 0, will return HC_plt as a structure array with two fields: 'out' and 'MRL_Crit'. It may contain up to two records (one per MRL GPD threshold) with the following format:
  - out = numerical array with the following 5 rows
    - row(01): mean values
    - row(02): values of 2% percentile or 1st percentage of input prc
    - row(03): values of 16% percentile or 2nd percentage of input prc
    - row(04): values of 84% percentile or 3rd percentage of input prc
    - row(05): values of 98% percentile or 4th percentage of input prc
  - MRL_Crit = character vector indicating the MRL threshold selection criterion used for the HC_plt.out record.

- HC_tbl: summarized HC. The format of this field depends on the value of ind_TH_mrl. When ind_TH_mrl = 1 or 2, HC_tbl is a numerical array with the 5 rows mentioned below. However, specifying ind_TH_mrl = 0, will return HC_tbl as a structure array with two fields: 'out' and 'MRL_Crit'. It may contain up to two records (one per MRL GPD threshold) with the following format:
  - out = numerical array with the following 5 rows
    - row(01): mean values
    - row(02): values of 2% percentile or 1st percentage of input prc
    - row(03): values of 16% percentile or 2nd percentage of input prc
    - row(04): values of 84% percentile or 3rd percentage of input prc
    - row(05): values of 98% percentile or 4th percentage of input prc
  - MRL_Crit = character vector indicating the MRL threshold selection criterion used for the HC_tbl.out record.

- HC_tbl_rsp_x: hazard values interpolated from the HC that correspond to the responses in HC_tbl_rsp_y. The format of this field depends on the value of ind_TH_mrl. When ind_TH_mrl = 1 or 2, HC_tbl_rsp_x is a numerical array with the 5 rows mentioned below. However, specifying ind_TH_mrl = 0, will return HC_tbl_rsp_x as a structure array with two fields: 'out' and 'MRL_Crit'. It may contain up to two records (one per MRL GPD threshold) with the following format:
  - out = numerical array with the following 5 rows
    - row(01): mean values
    - row(02): values of 2% percentile or 1st percentage of input prc
    - row(03): values of 16% percentile or 2nd percentage of input prc
    - row(04): values of 84% percentile or 3rd percentage of input prc
    - row(05): values of 98% percentile or 4th percentage of input prc
  - MRL_Crit = character vector indicating the MRL threshold selection criterion used for the HC_tbl_rsp_x.out record.

- HC_emp: empirical HC; as a table array with column headings as follows:
  - Response: Response vector sorted in descending order
  - Rank: Rank or Weibull's plotting position
  - CCDF: Complementary cumulative distribution function (CCDF)
  - Hazard: hazard as AEF or AEP
  - ARI: annual recurrence interval (ARI)

- Warning: message in the form of a cell array, to indicate a mismatch exist between the GPD fit and the empirical distribution. Example:
  - Warning: At 0.1 AEP/AEF, best estimate HC value is greater than 1.75 times the empirical HC value. Manual verification is recommended.

- ME: a MATLAB error message will be stored here when an unexpected error makes the tool fail.

8 History of revisions

- 20201015 – released version 0.1 Alpha
- 20201025 – released version 0.2 Alpha
- 20210311 – released version 0.3 Alpha
- 20210405 – released version 0.4 Alpha
- 20210809 – released version 0.5 Alpha

9 Feedback form

Please use the “Feedback form” to report any errors found and send to the following email addresses.

Efrain Ramos-Santiago	Efrain.Ramos-Santiago@erdc.dren.mil
Victor M. Gonzalez, PE	Victor.M.Gonzalez@erdc.dren.mil
Dr. Norberto C. Nadal-Caraballo	Norberto.C.Nadal-Caraballo@erdc.dren.mil

10 References

Borgman, Leon, Martin Miller, Lee Butler, and Robin Reinhard. 1992. “Empirical Simulation of Future Hurricane Storm Histories as a Tool in Engineering and Economic Analysis.” In , 42–65. ASCE. https://cedb.asce.org/CEDBsearch/record.jsp?dockey=0079320.

Langousis, Andreas, Antonios Mamalakis, Michelangelo Puliga, and Roberto Deidda. 2016. “Threshold Detection for the Generalized Pareto Distribution: Review of Representative Methods and Application to the NOAA NCDC Daily Rainfall Database.” *Water Resources Research* 52 (4): 2659–81. https://doi.org/10.1002/2015WR018502.

Scheffner, Norman W., Leon E. Borgman, and David J. Mark. 1994. “Applications of Large Domain Hydrodynamic Models to Generate Frequency-of-Occurrence Relationships.” In , 264–78. ASCE. https://cedb.asce.org/CEDBsearch/record.jsp?dockey=0087074.

Scheffner, Norman W., James E. Clausner, Adele Militello, Leon E. Borgman, Billy L. Edge, and Peter J. Grace. 1999. “Use and Application of the Empirical Simulation Technique: User’s Guide.” 99–21. Technical Report CHL. U.S. Army Engineer Waterways Experiment Station. http://hdl.handle.net/11681/7478.

The Mathworks. 2021. “Gaussian Process Regression Models.” Statistics and Machine Learning Toolbox. 2021. https://www.mathworks.com/help/stats/gaussian-process-regression-models.html#References.