ERDC/CHL CHETN-VIII-10
May 2018

US Army Corps of Engineers®

Representative Storm Selection Tool: An Automated Procedure for the Selection of Representative Storm Events from a Probabilistic Database

by Dylan Sanderson and Mark Gravens

PURPOSE: This Coastal and Hydraulics Engineering Technical Note (CHETN) presents and documents the Representative Storm Selection Tool (RSST), a utility that automates much of the analysis necessary to perform the identification and selection of representative storms outlined in Gravens and Sanderson (2018). The ability to easily select representative storms and calculate the relative probabilities will significantly speed up workflow for users developing representative storm suites for use in probabilistic life-cycle analysis (PLCA) models, such as Beach-*fx* (Gravens et al. 2007) and Generation Two Coastal Risk Model (G2CRM) (currently under development at the U.S. Army Engineer Research and Development Center, Coastal and Hydraulics Laboratory [ERDC-CHL] and Institute for Water Resources [IWR]).

BACKGROUND: As summarized in Gravens and Sanderson (2018), the need for a robust representative storm suite is critical for the development of environmental forcing input to PLCA models. PLCA models such as Beach-*fx* and G2CRM are data-driven models that rely on relational databases that are accessed from within the model's computational kernel. These relational databases contain storm responses required by the PLCA models (e.g., in Beach-*fx*, the beach profile's morphological response to storm events). Because PLCA models do not compute responses as a part of their run time, these relational databases must contain all of the response data that are expected to occur over a project's simulated life cycle. For example, in Beach-*fx* this often results in a single beach profile containing on the order of 200 unique pre-storm upper beach configurations. Considering that each storm event is combined with 12 astronomical tides to cover the full variation in tidal range and surge-tide phasing, it is not computationally effective to populate these relational databases using the full probabilistic storm suite available from the Coastal Hazards System (CHS; *https://chs.erdc.dren.mil/default.aspx*), which often contains from 500 to 1000 synthetic or historically based storm surge hydrographs at each save point. To alleviate this computational burden, a representative storm suite is developed that effectively characterizes the full probabilistic space.

METHOD: The RSST performs the four procedures outlined in Gravens and Sanderson (2018).

- Group storms into clusters based on the magnitude of the peak surge generated.
- Further sub-divide storm clusters if appropriate, based on duration of storm surge hydrograph (sub-clusters).
- Select representative storm events within each storm cluster or sub-cluster.
- Assign appropriate relative probability to each selected representative storm.

The RSST outlined in this CHETN has the capability to perform the above steps automatically but allows the user to define or adjust the lower and upper limits of each cluster that will determine how the storms are grouped, to select whether a given cluster should be divided into two sub-clusters, and to replace the representative storm automatically selected with a different storm in the storm cluster. This tool is ultimately aimed at assisting the user with the development of a representative storm suite and increasing workflow efficiency and is not intended to be used without user engagement. User supervision of the selected storms should be employed and the results overridden when deemed appropriate.

When using this tool to develop a representative storm suite, if applicable, the tropical and extra-tropical storms must be analyzed separately. The steps outlined above remain the same for each set of storms, although the cluster grouping and calculation of relative probabilities differ. Because the storms are analyzed separately, it is not necessary, nor recommended, to use the same clustering upper and lower surge limits for both tropical and extra-tropical storms. If the storms are defined from a record of historical data, then they are assumed to span the entire probabilistic space and are thus weighted equally. Conversely, synthetically defined storms are often assigned a relative probability. Relative probabilities of storm intensity developed for the USACE North Atlantic Coast Comprehensive Study (Cialone et al. 2015) and the Sabine to Galveston Study[1] are embedded within the RSST.

USING THE RSST: The RSST is a Windows executable. When launched, the window shown in Figure 1 opens. This window is divided into four regions based on each feature's functionality:

1. Input/output files and directories.
2. Grouping of storms into clusters.
3. Implement hydrograph time series smoothing or apply unit conversions.
4. Run the RSST or clear input values.

The first region defines the input/output files and locations. The button "Browse Time Series (*.h5)" allows the user to select the HDF5 file on the computer that contains the storm surge hydrograph time series. HDF5 files can be obtained by navigating to the CHS website and downloading the file associated with the desired save point. If the user wishes to consider only a subset of the storms in the HDF5 file, for example those that fall within a 200 kilometer (km) radius of the project site location, then the button "Browse Select Storms (*.csv)" allows the user to import a comma separated values (CSV) file containing the full name associated with the selected storms. The CSV is created by the user and needs to be composed of one column and contain one header row. An example CSV for import can be seen in Figure 2. The button "Location for Output Files" allows the user to select the folder where the output files from the storm analysis will be written.

[1] Melby, J. A., N. C. Norberto, J. J. Ratcliff, T. C. Massey, and R. E. Jensen. Draft. *Sabine Pass to Galveston Bay Wave and Water Level Modeling.* ERDC/CHL Technical Report. Vicksburg, MS: U.S. Army Engineer Research and Development Center.

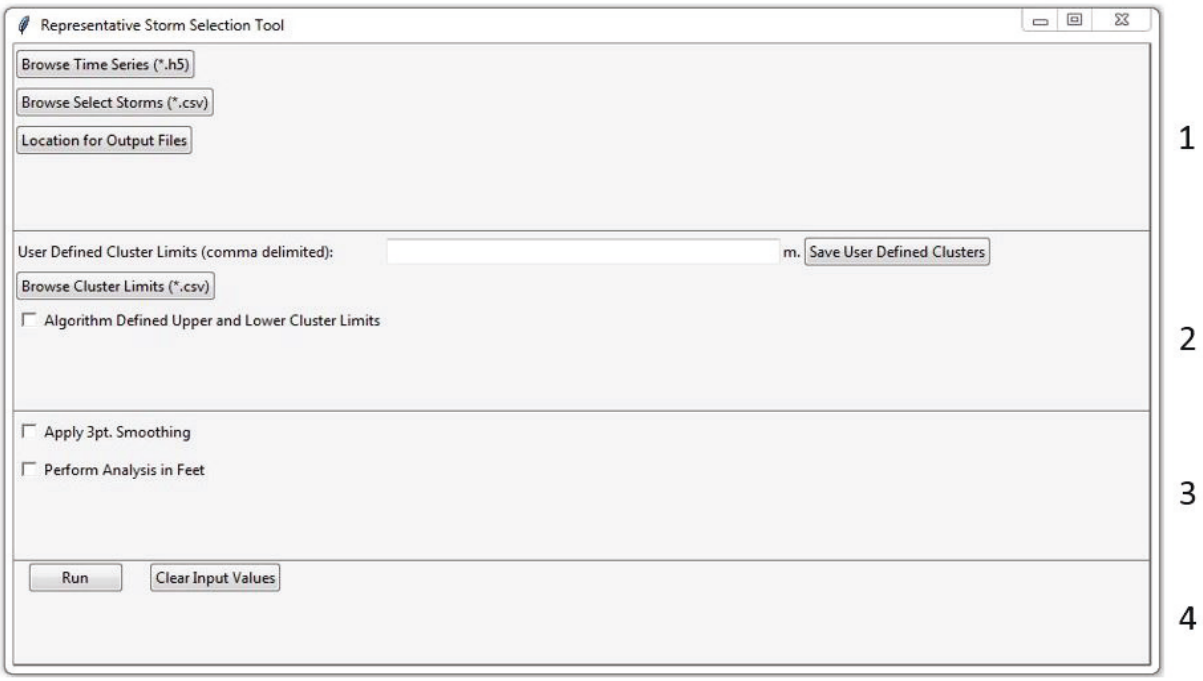

Figure 1. RSST graphical user interface (GUI) with regions labeled on right.

	A
1	header
2	Synthetic_001 - 1
3	Synthetic_002 - 2
4	Synthetic_003 - 3
5	Synthetic_004 - 4
6	Synthetic_005 - 5
7	Synthetic_006 - 6
8	Synthetic_007 - 7
9	Synthetic_008 - 8
10	Synthetic_009 - 9
11	Synthetic_046 - 46

Figure 2. Example of CSV containing subset of storms to be considered.

The second region allows the user to specify how the RSST will group the storm surge hydrographs into clusters. As outlined in Gravens and Sanderson (2018), each hydrograph is placed into a cluster based on the peak surge elevation in relation to the lower and upper cluster surge limits. The user has three options for defining the cluster limits:

1. Manually input lower and upper cluster limits separated by commas.
2. Select a CSV file from the user's computer that contains the lower and upper cluster limits. This CSV is generated from a previous analysis.
3. Allow the RSST to automatically define the lower and upper cluster limits.

The first option allows the user to manually define the cluster limits. An input of

0.5, 1.0, 1.5, 2.0, 3.0, 5.0

will result in the upper and lower limits of each cluster shown in Table 1. Guidance for manually defining the lower and upper cluster limits is outlined in Gravens and Sanderson (2018).

Table 1. Upper and lower surge limits of each cluster.

Lower Surge Limit (meters)	Cluster Number	Upper Surge Limit (meters)
0.5	1	1.0
1.0	2	1.5
1.5	3	2.0
2.0	4	3.0
3.0	5	5.0

The second option of selecting a file to be imported allows the user to use the cluster limits developed from, or used in, a previous analysis. After an analysis is complete, a CSV output file is generated that contains the lower and upper limits (Figure 3). This CSV file can then be used to re-analyze previously selected storms. If the cluster ranges are defined manually or a CSV file is selected, the cluster limits will be printed to the screen.

	A	B
1	Units:	m.
2	Cluster Lower Values	Cluster Upper Values
3	0.5183	0.6512
4	0.6512	0.7695

Figure 3. Example of cluster limits file.

If the third option, “Algorithm Defined Upper and Lower Cluster Limits” is selected, then a slider will appear that allows the user to adjust whether there will be few or many clusters. Internally, the algorithm will determine the number of clusters and the lower and upper limits of each cluster by performing a kernel density estimate (KDE) on the peak surge generated by each storm surge hydrograph. The number of automatically defined clusters depends on the storm set and can range from approximately 4 to 30 clusters. The KDE used within the tool is a part of the Python library, Scikit-Learn (Pedregosa et al. 2011).

A KDE attempts to estimate the probability density function of a set of data. The development of a KDE curve is similar to that of creating a histogram, except instead of setting up bins, the data points are treated independently, and a Gaussian curve is *stacked* at each point. The slider that allows the user to specify “More Clusters” or “Less Clusters” controls the width, or bandwidth, of the Gaussian curves.

The internal procedure of selecting lower and upper cluster limits by performing a KDE on the peak surge generated by each storm is illustrated in Figure 4. The peak surge values are shown by the marks along the *x*-axis, ranging from approximately 0.2 to 4.1 meters (m), and the resulting KDE is shown by the black line. The KDE curve returns maximum values (green points) at areas of high density and minimum values (red points) at areas of low density. The cluster limits are specified as the elevations where the minimum values are (in this case resulting in 8 original clusters). If a wider bandwidth were chosen (slider towards “Less Clusters”), the

original KDE would have resulted in a smoothed version of the one shown in Figure 4. To ensure that the RSST returns a similar number of storms in each cluster, the algorithm will automatically reduce the bandwidth in the clusters where the KDE curve is noticeably larger than the other portions of the curve. In Figure 4, it can be seen that the maximum values of the KDE in the first, second, and third clusters are significantly larger than the remainder of the KDE; thus, the bandwidth is reduced in these regions. The blue line shows the resulting, reduced bandwidth KDE curve. In this case, the first, second, and third clusters are divided into six new clusters (0–0.51 m, 0.51–0.65 m, 0.65–1.1 m, 1.1–1.35 m, 1.35–1.51 m, 1.51–1.8 m) while the subsequent clusters remain the same (1.8–2.3 m, 2.3–2.75 m, 2.75–3.5 m, 3.5–4.0 m, >4.0 m).

Furthermore, the cluster limits are adjusted such that any cluster with greater than 50 storms or fewer than 5 storms is either reduced or increased respectively[1]. Figure 5 shows an example distribution of storms resulting from this method of clustering. The storms were sorted sequentially based on the peak surge generated, with each consecutive color representing a cluster.

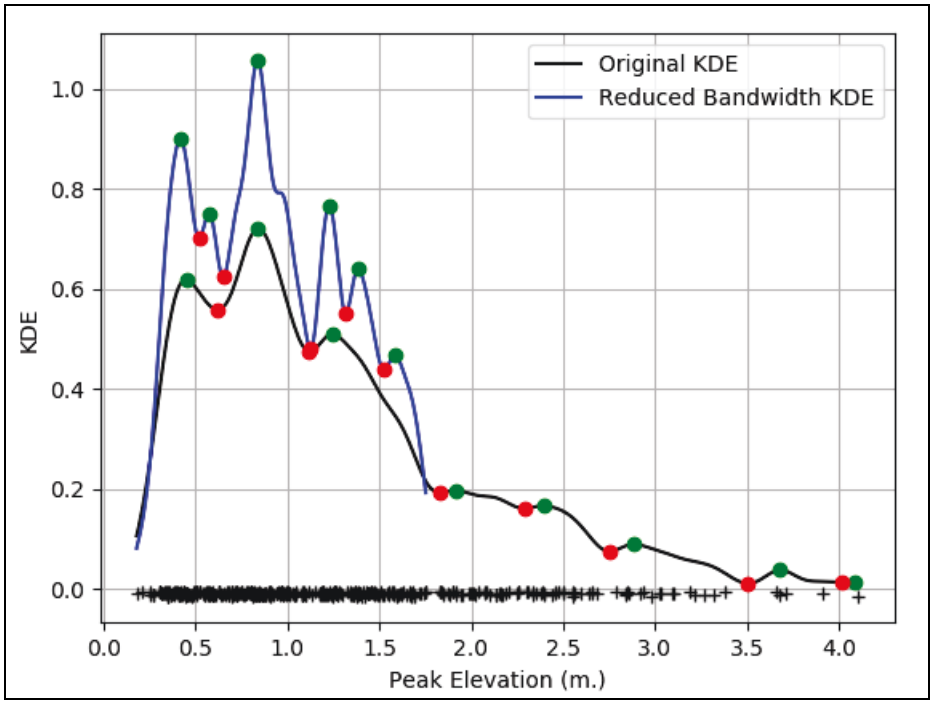

Figure 4. Original and reduced bandwidth KDE.

[1] The process of reducing or increasing cluster sizes is done iteratively. In some cases, the number of storms in a cluster will not converge to either below 50 or above 5, so the algorithm will break out of this iterative process. Due to the failure to converge, this can result in clusters with more than 50 or fewer than 5 storms.

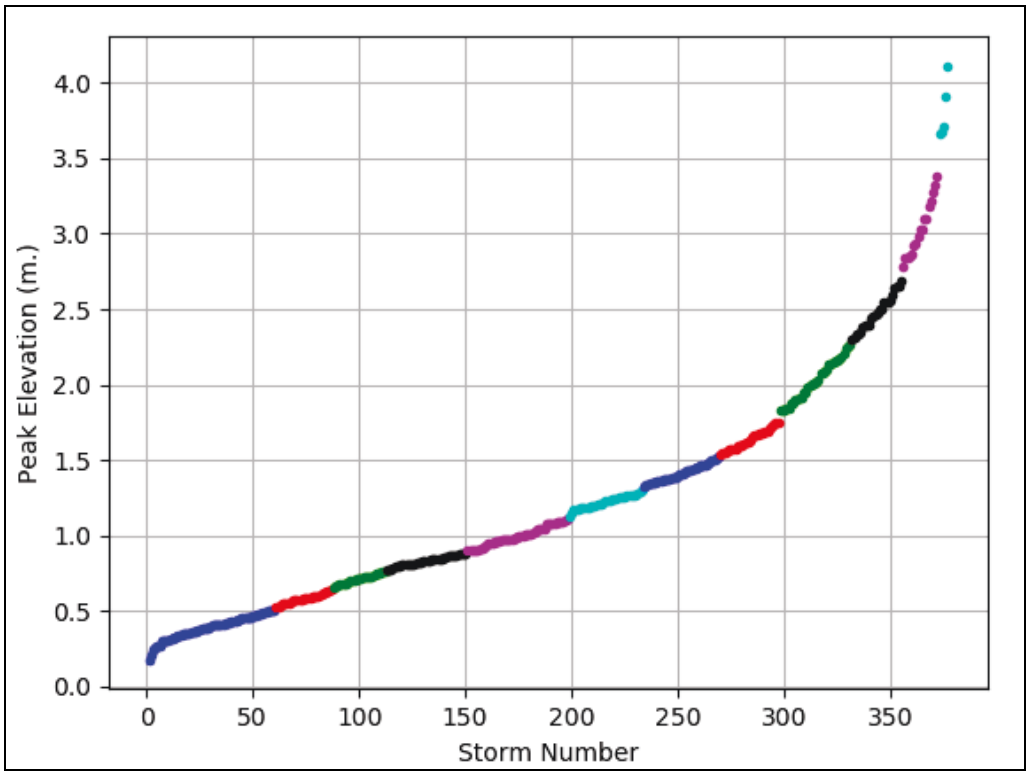

Figure 5. Distribution of storms into clusters.

The third region of the GUI allows the user to specify options, such as applying a three-point moving average to each storm surge hydrograph or converting each hydrograph from meters to feet. The former option is selected if the time series are noisy and the user wishes to smooth each storm surge hydrograph. If the latter option "Perform Analysis in Feet" is selected, the algorithm will convert the storm surge hydrograph time series from meters to feet. The output CSV file that contains the cluster limits will include the units of the analysis in the header row. If this file is imported to reanalyze a storm suite, the algorithm will read the header row and perform unit conversions as necessary.

The fourth and final region of the GUI allows the user to run the RSST with the above specifications or to clear all input values.

When the tool is run, the algorithm will internally load the storm surge hydrograph time series from the HDF5 file, group the hydrographs into clusters as specified in the second region of the initial window, and suggest whether each cluster should be divided into one or two sub-clusters. To distinguish between one or two sub-clusters, the algorithm analyzes the duration of each hydrograph above a threshold elevation.

An average hydrograph for each cluster of storms is computed, from which the midpoint between the lowest elevation and the peak elevation is determined and specified as the midpoint threshold. The duration that each storm is above this midpoint threshold is extracted. Figure 6 shows a cluster of storms, the average storm surge hydrograph, and the midpoint threshold.

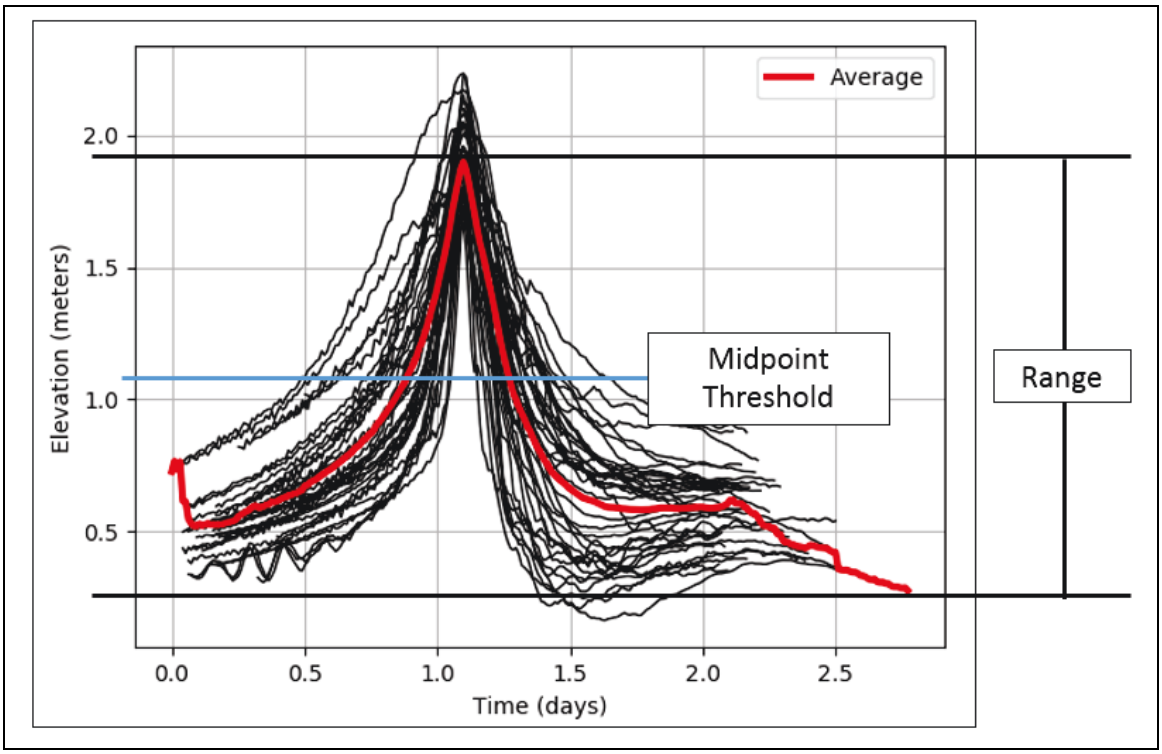

Figure 6. Average storm surge hydrograph and the midpoint threshold.

From the duration of each storm surge hydrograph above the midpoint, a KDE is generated. If there are two noticeable maximum values in the KDE, then two sub-clusters are chosen, indicating that there is a significant population of long- and short-duration storms.

After each cluster is evaluated for sub-clusters, a new window (Figure 7) will appear that allows the user to view each cluster of storm surge hydrographs and review the suggested number of sub-clusters. Note that each cluster can only contain 1 or 2 sub-clusters. The clusters are sorted sequentially based on the peak surge elevation limits, and a number is assigned to each ranging from 1 to the total number of clusters. To view a cluster of storm surge hydrographs, the box in the lower-left corner is selected, and a dropdown menu appears, from which the user can select a cluster number. Upon selection, the plot on the left will be updated with the storm surge hydrographs contained within the selected cluster. The option bar at the top left of the window provides the user with plotting options. From this menu the user can zoom, pan, return to the original window, change the size of the plot, or save the image as a portable network graphics (PNG) file. A table to right of the plot shows a summary of each cluster number, how many sub-clusters are selected, and the total number of storm surge hydrographs in the cluster.

If the user decides to change how many sub-clusters are present, then the user can input “1” or “2” in the input box located in the lower-left corner. The button “Save Selected Number of Sub-Clusters” must be pressed, and the table to the right will be updated with the new number. When the user is satisfied with the number of sub-clusters in each group, the “Continue” button is pressed.

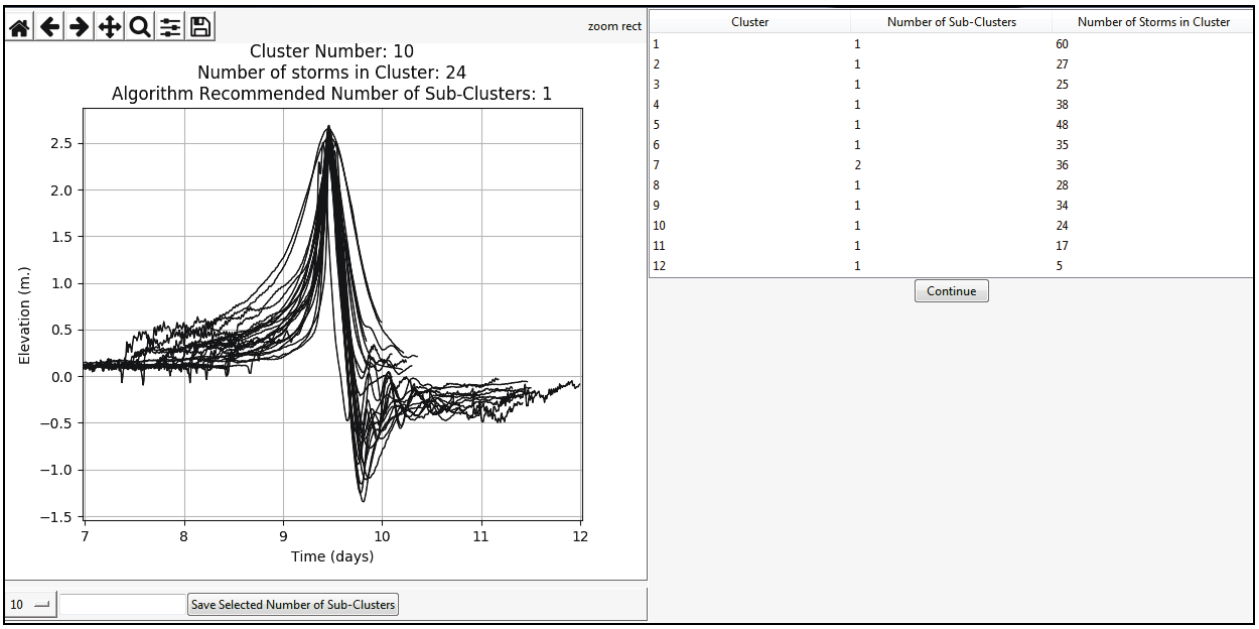

Figure 7. Window allowing user to select one or two sub-clusters.

The RSST will then select a single representative storm for each sub-cluster. Internally, each storm surge hydrograph is reduced to a series of significant points that characterize key features of the hydrograph, such as duration and elevation. The duration that each hydrograph is above three elevation thresholds (left plot of Figure 8) and the elevation of each storm surge hydrograph at five time-steps (right plot of Figure 8) are extracted. In the right plot of Figure 8, it can be seen that one of the time-steps where the elevations are extracted is located where the storm surge hydrographs are at a peak elevation. This elevation is extracted twice to be weighted more heavily than the other points.

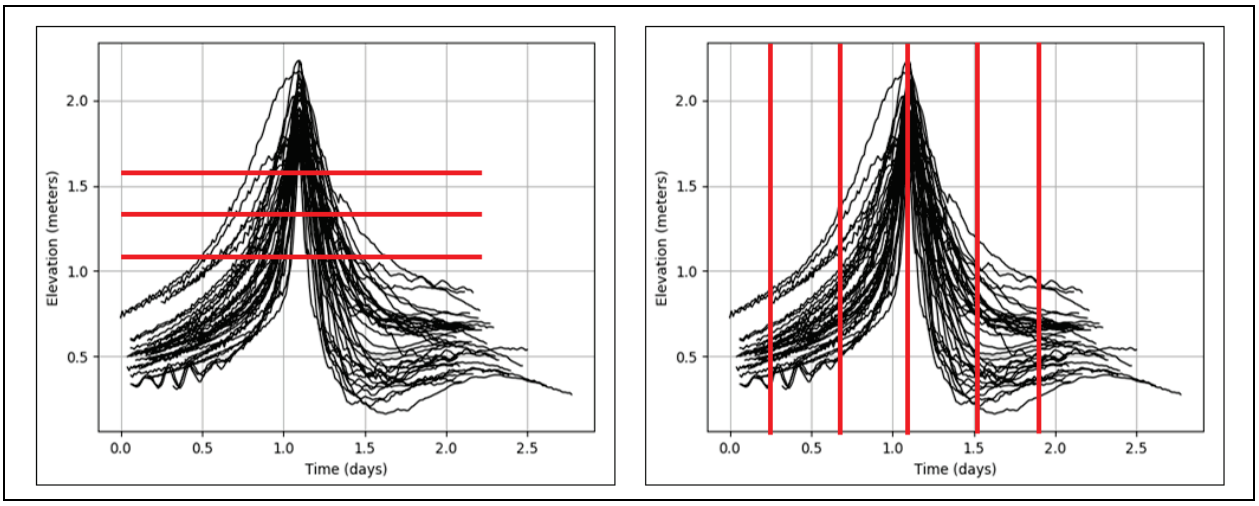

Figure 8. Duration of storm hydrographs above three elevation thresholds.

The centroid, or mean, of each of the extracted nine points is calculated ($c_1 - c_9$) for every sub-cluster. The Euclidian distance, d_n, from each of the hydrograph's nine points ($s_{n1} - s_{n9}$) within a sub-cluster to the centroid of the sub-cluster is computed as follows:

$$d_n = \sqrt{(c_1 - s_{n1})^2 + (c_2 - s_{n2})^2 + \cdots + (c_9 - s_{n9})^2}$$

Where n corresponds to each storm surge hydrograph contained within the sub-cluster. The storm that minimizes the Euclidian distance is selected as the representative storm for the sub-cluster.

After representative storms are identified, a new window (Figure 9) allows the user to examine the results and the storm surge hydrographs in each cluster. If two sub-clusters are chosen, then the cluster will be divided into an *A* and *B* group, corresponding to short- and long-duration storms. The user can view and change the selected representative storm for each sub-cluster by selecting either the *A* or *B* sub-cluster from the menu. Both the long and short sub-clusters can be viewed simultaneously, but the selected representative storms cannot be updated from this plot. The table to the right of the plot indicates the selected representative storm in each cluster.

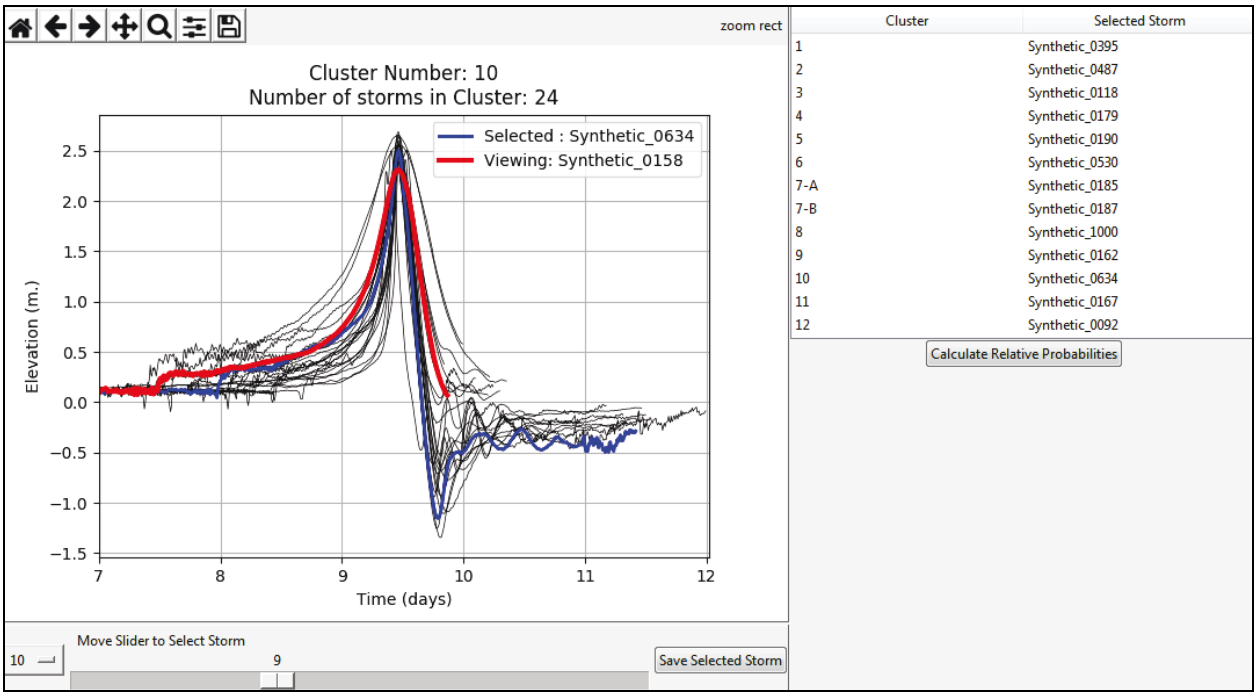

Figure 9. Viewing recommended representative storm surge hydrograph and comparing plausible alternatives.

The slider at the bottom of the screen allows the user to select and view other possible representative storm surge hydrographs in the cluster by highlighting a single hydrograph in red while the selected representative storm will remain highlighted in blue. The use of the slider to view other possible representative storms is illustrated in Figure 9. If the user wishes to replace the automatically selected storm surge hydrograph with another storm surge hydrograph, then the user highlights the desired storm with the slider and the "Save Selected Storm" button must be pressed. The table to the right will be updated with the new selected representative storm. When the figure is reloaded, the new selected representative storm will be represented by the blue hydrograph.

When the user is satisfied with the representative storms in each cluster, the "Calculate Relative Probabilities" button is pressed, from which a new window emerges with a table summarizing the results of the analysis (Figure 10). The relative probability computed for each representative storm is defined as the sum of the relative probabilities of the storms, which that specific storm represents. The summary of the analysis shows each cluster with the selected representative storm, the number of storms that the selected storm represents, and the computed relative probability.

Cluster	Selected Storm	Number of Storms Represented	Relative Probability
1	Synthetic_0395	60	0.06249
2	Synthetic_0487	27	0.025318
3	Synthetic_0118	25	0.020442
4	Synthetic_0179	38	0.041278
5	Synthetic_0190	48	0.051388
6	Synthetic_0530	35	0.029931
7-A	Synthetic_0185	9	0.001154
7-B	Synthetic_0187	27	0.025087
8	Synthetic_1000	28	0.014756
9	Synthetic_0162	34	0.014752
10	Synthetic_0634	24	0.009921
11	Synthetic_0167	17	0.00727
12	Synthetic_0092	5	0.000617

Figure 10. Table with summary of analysis.

After the analysis is complete, the summary shown in Figure 10 and the cluster limits for each bin are written to separate output CSV files (Storm_Selection_Summary.csv and Storm_Selection_Clusters.csv). The output files will be succeeded by either "Tropical" or "Extra-Tropical," depending on the type of storms analyzed. The file "Storm_Selection_Clusters.csv" can be directly imported into the initial GUI window shown in Figure 1, allowing the user to easily repeat the analysis if necessary using the same clusters.

CONCLUSION: This CHETN describes the usage of the Representative Storm Selection Tool developed for identifying and selecting representative storms from a probabilistic storm database. The tool developed follows the procedure outlined by Gravens and Sanderson (2018). Storm surge hydrograph time series are obtained from the CHS in HDF5 file format, which are used as input to the RSST. A CSV file containing a sub-set of storms to analyze from the HDF5 files can be imported, identifying, for example, storms whose storm tracks pass within a 200 km radius of the project site. The RSST has the ability to automatically define cluster limits and place the storms into clusters by determining areas of low density on the KDE curve resulting from the peak value of each storm surge hydrograph. Alternatively, the user can manually define the lower and upper cluster limits. The algorithm will suggest whether each cluster should contain one or two sub-clusters and identify a possible representative storm for each cluster, both of which can be reviewed and updated by the user. After the representative storms are selected, the relative probabilities of the selected storms are computed.

The tool described in this CHETN is intended to increase workflow efficiency of representative storm suite development for use in PLCA models. The RSST provides recommended representative storms; however, user supervision of the selected representative storms is recommended and should be employed. The user is provided the capability to override any of the recommended representative storms and should exercise that capability when appropriate.

ADDITIONAL INFORMATION: This Coastal and Hydraulics Engineering Technical Note was prepared by Dylan R. Sanderson (*Dylan.R.Sanderson@usace.army.mil*) and Mark B. Gravens, U.S. Army Engineer Research and Development Center. The study is funded by the USACE Flood and Coastal Systems R&D Program. This technical note should be cited as follows:

> Sanderson, D. R., and M. B. Gravens. 2018. *Representative Storm Selection Tool: An Automated Procedure for the Selection of Representative Storm Events from a Probabilistic Database*. ERDC/CHL CHETN-VIII-10. Vicksburg, MS: U.S. Army Engineer Research and Development Center. *http://dx.doi.org/10.21079/11681/26829*.

The RSST can be obtained by contacting Dylan Sanderson at *Dylan.R.Sanderson@usace.army.mil*.

REFERENCES

Cialone, M. A., T. C. Massey, M. E. Anderson, A. S. Grzegorzewski, R. E. Jensen, A. Cialone, D. J. Mark, K. C. Pevey, B. L. Gunkel, T. O. McAlpin, N. C. Nadal-Caraballo, J. A. Melby, and J. J. Ratcliff. 2015. *North Atlantic Coast Comprehensive Study (NACCS) Coastal Storm Model Simulations: Waves and Water Levels.* ERDC/CHL TR-15-12. Vicksburg, MS: U.S. Army Engineer Research and Development Center. *http://hdl.handle.net/11681/7339*.

Gravens, M. B., R. M. Males, and D. A. Moser. 2007. "Beach-fx: Monte Carlo Life-Cycle Simulation Model for Estimating Shore Protection Project Evolution and Cost Benefit Analyses." *Shore and Beach* 75(1): 12–19.

Gravens, M. B., and D. R. Sanderson. 2018. *Identification and Selection of Representative Storm Events from a Probabilistic Storm Data Base*. ERDC/CHL CHETN-VIII-9. Vicksburg, MS: U.S. Army Engineer Research and Development Center.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12(October): 2825–2830.