Clustering and selection of hurricane wind records using a machine learning approach

Xinlong Du¹, Jerome F. Hajjar², Robert Bailey Bond³, Pu Ren⁴ and Hao Sun⁵

¹Graduate Research Assistant, Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, USA. Email: du.xinl@northeastern.edu

²CDM Smith Professor and Chair, Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, USA. Email: jf.hajjar@northeastern.edu

³Graduate Research Assistant, Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, USA. Email: bond.rob@northeastern.edu

⁴Graduate Research Assistant, Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, USA. Email: ren.pu@northeastern.edu

⁵Associate Professor, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China. Email: haosun@ruc.edu.cn

ABSTRACT

In wind engineering, to accurately estimate the nonlinear dynamic response of structures while considering uncertainties of hurricanes, a suite of wind records representing the hurricane hazards of a given location is of great interest. Such a suite generally consists of a large number of hurricane wind records, which may lead to highly computational cost for structural analysis. To reduce the computational demand while still preserving the accuracy of the uncertainty quantification process, this paper proposes a machine learning approach to select a representative subset of all collected hurricane wind records for a location. First, hurricane wind records, which are expressed as time series with information that includes both wind speed and direction, are collected from a synthetic hurricane catalog. The high-dimensional hurricane wind records
records are then compressed into a set of low-dimensional latent feature vectors using an artificial neural network, designated as an autoencoder. The latent feature vectors represent the important patterns of wind records such as duration, magnitude and the changing of wind speeds and directions over time. The wind records are then clustered by applying the k-means algorithm on the latent features, and a subset of records is selected from each cluster. The wind records selected from each cluster are those whose latent feature points are closest to the centroid of all latent feature points in that cluster. In order to do regional analysis while taking into account that the hurricane wind records are site-specific, this paper suggests that a region can be discretized into a set of grids, with the proposed hurricane selection approach applied to each grid. This procedure is demonstrated using Massachusetts as a testbed.

Keywords: hurricane selection, time series clustering, autoencoder, k-means, uncertainty quantification, regional analysis, wind direction

Introduction
Nonlinear dynamic analysis is increasingly being considered in wind design of buildings and other structures as performance-based design becomes an increasingly popular option (ASCE 2019), where controlled inelastic deformations are allowed under strong winds (Wang and Wu 2022). In the fully probabilistic performance-based hurricane engineering framework (Barbato et al. 2013), fragility curves of structures are commonly adopted to do probabilistic damage assessment. Strength limit states of structures usually involve nonlinear behavior that is then integrated into the predictions of likelihood of damage that is offered through fragility analysis. Fragility functions are defined as the failure probability of a structure conditional on the intensity measure of hazards, including hurricanes in this work. If only the failure probability or fragility is of interest for a hurricane event (i.e., at the end of the loading time history) instead of for a certain time interval within the hurricane duration, the uncertainties in the loading time histories can be accounted for through running a series of nonlinear dynamic analysis with a suite of hurricane wind records. The hurricane wind records should include time histories that incorporate the wind directions as well as wind speeds, because the changing of wind directions during hurricanes has significant effects on
the structural response. Consequently, the wind records in this research are time series of both wind speed and direction. To develop accurate fragility curves, the structures should be analyzed with a suite of hurricane wind records that can cover the record-to-record uncertainties in the changing of wind speeds and directions within the hurricane durations. A large amount of hurricane wind records can be collected for a location considering the existing historical and synthetic hurricanes (ASCE 2016b; Vickery et al. 2010; Vickery et al. 2009b; Vickery et al. 2009c). However, it is challenging to run nonlinear dynamic analysis for all of the collected hurricane records due to the high computational demand of finite element analysis of structures; thus, a minimum number of hurricane records should be selected to represent the uncertainties in all of the collected hurricane records.

In prior work, Li (2005) and Li and Ellingwood (2006) developed hurricane fragility curves for wood-frame residential construction with a simplified limit state function, where the nonlinear and dynamic effects are neglected. Cui and Caracoglia (2015) carried out fragility analysis on tall buildings only for the serviceability limit state, so frequency domain analysis is adopted, and duration and nonstationary effects of hurricanes cannot be considered. In order to avoid performing structural analysis for long durations of windstorms, the dynamic shakedown method was utilized by researchers to model the inelastic behavior of buildings (Chuang and Spence 2019, 2020; Tabbuso et al. 2016). Other researchers tried to develop hurricane fragility curves using nonlinear dynamic analysis only for a fixed time interval with a constant wind direction. For example, Hallowell et al. (2018) used wind records with 1-hour time intervals, while Ma et al. (2021) used wind records with 2-minute time intervals. The fragility developed for this certain time interval cannot represent the fragility for a whole hurricane because of dynamic effects, yielding and changes in wind speeds and directions. Of course, one can discretize the hurricane duration into a series of short time intervals and apply the developed fragility curves to each short time interval; however, the failure probabilities within those short time intervals are correlated (Der Kiureghian 2005; Kim et al. 2019; Straub et al. 2020). This correlation is difficult to quantify from the view of time-variant reliability and is not considered by the above authors. Given the limitations of the previous research, this paper considers the
failure probability for a hurricane event instead of a certain time interval during a hurricane and tries to
select hurricane wind records that can account for the record-to-record uncertainties in hurricanes. The
selected wind records can be used to estimate failure probabilities of structures with nonlinear time history
analysis. Through this way it is no longer needed to estimate the correlations of failure probabilities in the
short time intervals within a hurricane.

In performance-based earthquake engineering (Moehle and Deierlein 2004), a probabilistic framework has
been proposed to integrate seismic hazard analysis and structural damage analysis, where a suite of ground
motions are adopted to represent the uncertainties in earthquake ground motions. Ground motion selection
has been widely studied in the literature (Baker and Lee 2018; Bojórquez et al. 2013; Du and Padgett 2021;
Jayaram et al. 2011; Naeim et al. 2004). Some generally used ground motions suites are the SAC
(Somerville et al. 1997) records, LMSR (Krawinkler et al. 2003) records, and FEMA-P695 records (FEMA
2009). Recently, machine learning approaches have also been introduced to ground motion selection, where
a reduced number of ground motions are obtained through clustering of a large number of ground motions
(Bond et al. 2022; Kim et al. 2021; Zhang et al. 2020). However, there is no similar research in the literature
for selection of hurricane wind records. There are two instances of prior research that consider uncertainties
in hurricane wind hazards using a set of wind records, but they do not use a selection procedure. Vickery
et al. (2006) studied hurricane fragility curves for building envelope components that were developed in
the Hazus-MH software by comparing the wind pressure demand and the capacity of the envelope
components. The record-to-record uncertainties of hurricane wind speeds were accounted for through the
use of a 20,000-year simulation of hurricanes created by employing the hurricane model described by
Vickery et al. (2000a) and Vickery et al. (2000b). The simulated hurricanes inherently incorporated many
of the duration effects associated with the changes in wind speed and direction which accompany hurricane
winds. Joyner and Sasani (2018) developed fragility curves for the windborne debris damage of building
glazing, where eight hurricanes that made landfall in the U.S. in the last 10 years were adopted.
Uncertainties in the record-to-record variability for different hurricanes were accounted for by employing
the eight hurricane records in the damage analysis. Vickery et al. (2006) used all hurricanes in the 20,000-year simulation, which may address the uncertainties in hurricanes, but is not suitable for nonlinear dynamic analysis considering the computational demand. On the contrary, Joyner and Sasani (2018) only used eight hurricanes without an analysis of the hazard uncertainties, which may not be able to represent the uncertainties in hurricanes for a specific location.

This paper proposes a procedure to select a suite of hurricane wind records that can be used for performance-based design and fragility analysis. The wind speed and direction records for a location are collected from a synthetic hurricane catalog (Liu 2014) with some preprocessing, after which the collected wind records have durations that are short enough to make a nonlinear time history analysis feasible. The collected wind records are then compressed into low-dimensional latent feature vectors using a neural network designated as an autoencoder (Aggarwal 2018), so that it is easier to measure similarity of different wind records and apply the standard clustering algorithms such as the k-means algorithm (Aggarwal et al. 2001; Shalev-Shwartz and Ben-David 2014). Autoencoder is an artificial neural network in which the input and output layers have the same number of neurons, while the number of neurons in the middle is constricted. The training algorithm tries to reconstruct the input data in the output layer; however, this reconstruction is not exact because the neurons in the middle only carry a reduced representation of the input data. The data held by the neurons in the middle (i.e., the low dimensional vectors compared to the input and output layer) are called latent features, to which the clustering algorithm is applied. This means that only important information in the wind records is preserved for clustering. The latent features representing hurricane wind records are then clustered into several groups using the conventional k-means algorithm (Shalev-Shwartz and Ben-David 2014). Finally, only a few hurricane wind records are selected from each cluster for fragility development or design checks, which significantly reduces the number of required time history analyses, while still ensuring that the uncertainties of different hurricanes are covered with a limited number of wind records. Since the properties of hurricanes for different locations have significant differences, a hazard map can be developed for hurricane wind records so that users are able to choose appropriate records for their
locations of interest. As an example, the Commonwealth of Massachusetts has been divided into 92 grids and a suite of hurricane wind records has been selected for each grid using the proposed hurricane selection approach. To demonstrate the usefulness and effectiveness of the selected wind records, they have been adopted to develop fragility curves for electrical transmissions towers in Massachusetts, which can be found in Du et al. (2022) and Du and Hajjar (2022).

**Hurricane simulation**

Synthetic hurricanes are widely used for risk analysis and structural design in wind engineering, which evolved from the single site probabilistic model (Russell 1971) to Vickery’s hurricane track model (Vickery et al. 2000a). For example, ASCE 7 has adopted the hurricane track model when generating the wind hazard maps (ASCE 2016b). This research also uses a 10,000 year synthetic hurricane catalog developed by Liu (2014) for the Atlantic basin based on Vickery’s hurricane track model, which consists of a hurricane genesis model, a track model, an central pressure model, a decay model and a boundary layer model. The temporal and spatial evolution of thousands of hurricanes from emergence to dissipation was modeled using the Monte Carlo method. The HURDAT database (Jarvinen et al. 1984) for historical storms was adopted for building the hurricane model through regression and calibrating the simulated results. In this simulation, the state of a hurricane can be determined with 7 parameters: the hurricane eye’s latitude and longitude, storm translation speed, storm heading angle, storm central pressure, radius to maximum winds (describing storm size), and Holland’s radial pressure profile parameter (i.e., the Holland B parameter (Holland 1980)). These parameters are updated at each 6-hour point. As suggested by Vickery et al. (2000a), linear interpolation is performed within each 6-hour interval, which results in 10-min updates of the parameters as used in Vickery et al. (2009c). Examples of the simulated hurricane tracks are shown in Fig. 1. In this research, the gradient wind speeds are calculated by employing Georgiou’s model (Georgiou 1985), which gives the 10-min sustained wind speeds at 500 m to 2000 m above the ground surface (Cui and Caracoglia 2019; Pei et al. 2014, 2018). An example of the calculated gradient wind field is shown in Fig. 2.
The obtained hurricane gradient wind speeds $V_g$ need to be converted to surface wind speeds $V_{10}$ (10 m above the ground or water) for wind force calculation on structures. The reduction factor $V_g/V_{10}$ over water proposed by Batts et al. is used in this research (Batts et al. 1980; Vickery et al. 2009a). A sea-land transition factor obtained from the model given in Simiu and Scanlan (1996) is then utilized to calculate the surface wind over land (open terrain with surface roughness $z_0 = 0.03$ m) from the surface wind over water ($z_0 = 0.0013$ m). In addition, the surface wind speed over land approaches the fully transitioned value asymptotically over a fetch distance as the wind moves from sea to land; therefore, the transition function proposed in Vickery et al. (2009b) is employed here, which defines the percentage of the sea-land transition.
as a function of the fetch distance. With the methods discussed in this section, the time series of the 10-min sustained wind speeds at 10 m height and the corresponding wind directions at a location of interest (assuming open terrain) during a hurricane may be obtained.

**Hurricane wind records collection and preprocess**

ASCE 7 wind hazard maps display wind speeds with a certain mean recurrence interval (MRI) for the entire U.S., including hurricane prone regions (ASCE 2016b). However, much information regarding hurricane winds is omitted in the ASCE 7 wind hazard maps, such as the variation of wind speeds and directions during a hurricane, and the durations of hurricane winds. This kind of information, which are contained in the time series of hurricane wind speeds and directions, are critical for structural response estimation and risk analysis. Thus, in this section, a number of hurricane wind speed and direction records are collected for a location of interest. In order to collect hurricane wind records for a region, the region is first discretized into a series of grids and then hurricane wind records are collected for each grid.

**Wind records for a location of interest**

A location in Massachusetts with latitude 41.7 and longitude -70.1 is used as an example in this section. Wind records are collected for this specific location from 10,000-year synthetic hurricanes developed by Liu (2014). Examples of the collected 10-min sustained wind speed and wind direction records at the location of interest are shown in Fig. 3 to Fig. 5 with the corresponding hurricane tracks. It is seen in Fig. 3(a), Fig. 4(a) and Fig. 5(a) that the hurricane eye usually moves thousands of miles from a hurricane’s genesis to dissipation. It is reasonable to assume that the wind speed induced by a hurricane that is very far away is relatively small and can be neglected. Therefore, as suggested by Vickery et al. (2009c), hurricane winds are considered only when the location of interest is within 250 km of the hurricane eye (see the blue circles in Fig. 3 to Fig. 5). This limit on distance also provides a limit for the durations of the hurricane wind records. Figure 3(c), Fig. 4(c) and Fig. 5(c) illustrate the absolute values of the wind speeds and the wind directions in a polar coordinate system, while Fig. 3(d), Fig. 4(d) and Fig. 5(d) illustrate the hurricane wind speeds in the North and East directions in a Cartesian coordinate system. Note that the wind direction
in the polar coordinate system is clockwise positive from the North direction. It is seen that the pattern of wind speed and direction records are different for different hurricanes, which depends on a number of factors, including the 7 parameters defining the hurricane eye tracks and wind fields. While the impacts of the hurricane wind field as shown in Fig. 2 on the wind records is complex, a qualitative analysis of the impact of hurricane eye tracks on the wind records provides examples of the range of loading developed during hurricanes. Specifically, when the location of interest is very close to the hurricane eye track, the record of the absolute values of the wind speeds usually has two peaks and the drop of the wind speed in the middle is due to the near zero wind speed in the hurricane eye (see Fig. 3(c)). On the contrary, if the location of interest is further from the hurricane eye track, the record of the absolute values of the wind speeds will typically only have one peak (see Fig. 4(c) and Fig. 5(c)). The difference between Fig. 4 and Fig. 5 is that the hurricane eye passes by the West or East side of the location of interest, which dominates the variation of the wind directions as presented in these two figures.

(a) The whole hurricane track (the blue circle represents the 250 km limit)
(b) The hurricane track within the 250 km limit (blue circle) of the location of interest (blue dot)

(c) Wind speed and direction records

(d) Wind speed records in the North and East directions

Fig. 3. An example of hurricanes going through the location of interest
(a) The whole hurricane track (the blue circle represents the 250 km limit)

(b) The hurricane track within the 250 km limit (blue circle) of the location of interest (blue dot)

(c) Wind speed and direction records
(d) Wind speed records in the North and East directions

Fig. 4. An example of hurricanes passing by the West side of the location of interest

(a) The whole hurricane track (the blue circle represents the 250 km limit)

(b) The hurricane track within the 250 km limit (blue circle) of the location of interest (blue dot)
Wind records are first collected through applying the 250 km distance limit between the hurricane eye and the location of interest. Hurricanes with very low wind speeds are then filtered out through a strategy that only considers hurricanes whose maximum wind speeds at the location of interest are greater than the 50-year MRI wind speed at the same location. The 50-year MRI wind speed obtained from the ASCE 7 Hazard Tool (ASCE 2016a, b) is a 3-second gust wind speed at 10 m above ground (47 m/s for this location), which is then converted to 10-min sustained wind speed at 10 m above ground (32.4 m/s for this location) following the approach proposed by Simiu and Scanlan (Simiu and Scanlan 1996). This 50-year MRI 10-min sustained wind speed is used as the threshold for comparison with the collected hurricane wind records to get rid of those with small maximum wind speeds. Thus, a total of 162 hurricane wind records are collected from the 10,000-year synthetic hurricanes, of which 160 records are shown in Fig. 6 within a Cartesian coordinate system and are used in the following sections for clustering. Only 160 records are included because 162 cannot be divided by the batch size (i.e., 16) employed in the training process of the
autoencoder, as will be introduced in the following sections. In addition, Fig. 7 presents the histogram of the durations of all of the collected hurricanes with a mean duration of 12.4 hours. To avoid the impulse effects, a 1-hour linear ramp-up and a 1-hour linear ramp-down are attached to the beginning and the end of the collected wind records, respectively, as recommended in the Prestandard for Performance-Based Wind Design (ASCE 2019). To be consistent with the hurricane wind records with 10-min intervals, the ramp is added as six 10-min steps with a constant wind direction. Note that the ramps are not included in Fig. 3 to Fig. 5 but included in Fig. 6 and Fig. 7. Moreover, as will be discussed in the following sections, the collected wind records may have different durations, but the autoencoder needs the same size for the input data of each record. Therefore, to facilitate training the autoencoder, zero paddings are added to the beginning and the end of the records that are shorter than the longest one. Consequently, all records after preprocess have the same length as the longest one. For each record, zero paddings at the beginning and the end have the same length, which means all records after preprocessing have a midpoint that is usually recorded when the hurricane eye is closest to the location of interest.

Fig. 6. The 160 collected hurricane wind records resolved in two directions.
Wind records for a region

Hurricane wind records for different locations have different patterns. Consequently, it is appropriate to select site-specific wind records instead of generic wind records for all locations. Compared to generic wind records, site-specific wind records have lower uncertainties and thus can be used to predict responses of structures at a given location more accurately. To collect and select site-specific hurricane wind records for a region of interest, this research proposes that this region can be discretized into a set of grids and the centroid of each grid is used to represent the whole grid for recording wind speeds and directions. Thus, hurricane wind records can be collected for all centroids of the grids. To demonstrate this idea, Fig. 8 shows Massachusetts as a testbed, which is divided into 0.2° by 0.2° grids. In Fig. 8, the red dots represent the centroids of the grids that are not associated with Massachusetts, while the 92 blue dots represent the centroids of the grids that are associated with Massachusetts. The hurricane wind records collection procedure proposed in the previous sections is then run for all 92 grids. Note that when generating the wind records, the percentage of the sea-land transition is calculated for the centroid of each grid based on its fetch distance. In addition, the 50-year MRI wind speeds for the centroids of some grids cannot be obtained from the ASCE 7 Hazard Tool because these centroids are over the ocean (see Fig. 8); therefore, for these cases,
locations within the same grids but on the land are used to find the 50-year MRI wind speeds. Figure 9 presents the histogram of the number of hurricanes collected for all 92 grids, with a mean value of 202.

Wind records clustering and selection

The approximately 200 collected hurricane wind records for each grid are still too many for design checks and fragility development, especially considering the long durations of the wind records. Incremental dynamic analysis (IDA) may be used to estimate collapse probability of structures under hurricanes (Du et al. 2022; Vamvatsikos and Cornell 2002). This approach is computationally intensive because direct integration of the nonlinear dynamic governing equations is required over the entire duration of the hurricane wind records and this nonlinear time history analysis needs to be run multiple times with scaled
wind records. As such, it is important to limit the number of records used. Therefore, in this research, the collected wind records for each grid are first clustered using a machine learning approach and then approximately 1/10 of the wind records in each cluster are selected, which are combined together to create approximately 20 selected wind records for each grid. This significantly reduces the number of nonlinear time history analyses required, while still preserving the uncertainties in the collected records. This procedure is similar to stratified sampling in statistics. Sampling is the process of selecting a subset from a population so that the characteristics of the whole population can be estimated using this subset, while stratified sampling is used by dividing the population into subpopulations (i.e., clusters in this paper), where the elements within each subpopulation are similar, and performing sampling on each subpopulation. Stratified sampling may improve the precision of the sample because sampling variability within each subpopulation is smaller than the sampling variability on the entire population (Botev and Ridder 2017; Parsons 2014). Specifically, the selected wind records can cover a spread of properties such as durations, patterns of wind speed records, and patterns of wind direction records, because the collected wind records are divided into clusters based on these properties.

**Fully connected autoencoder**

Since the collected hurricane wind records are time series of both wind speed and direction with different durations, it is challenging to cluster the records directly. To facilitate the clustering process, the high dimensional wind records are first transformed into low dimensional latent features using an artificial neural network named autoencoder (Aggarwal 2018; Bond et al. 2022; Tavakoli et al. 2020). The architecture of the autoencoder for wind records at the location of interest given in the previous section is presented in Fig. 10. It is seen that the input matrix is the original wind speed records in the Cartesian system, which has two columns with each column representing wind speed time histories in the North and East directions, respectively. The input matrix is first flattened into a vector as the input layer of the fully connected autoencoder and then passed through other hidden layers to reconstruct the data as another vector in the output layer, which is finally reshaped to a matrix as the reconstructed wind speed records in the Cartesian
system. Even though the two columns of the input matrix are correlated time series of wind speeds in two
directions, this “flatten” and “reshape” process is reasonable because the correlations are considered in the
flattened vectors (input and output layers) through the weights of the fully connected layers. In another
word, “flatten” and “reshape” only change the appearance of the data while retaining the relationships and
correlations of the elements within the data. A fully connected autoencoder means that all the neurons in
one layer are connected to all the neurons in the next layer. The autoencoder architecture consists of two
parts: the encoder that compresses the high dimensional input data into the small-size latent feature vector,
and the decoder that utilizes the latent features to reconstruct the input data. In this example, the flattened
wind speeds in the input layer are transformed into 5 latent features through the encoder process, which are
then expanded to form the reconstructed but still flattened wind records in the output layer through the
decoder process. The hidden layers with a nonlinear activation function (Tanh) are included to enhance the
power of this autoencoder so that it can map the input data into much smaller dimensional spaces. Here
Tanh is adopted because it has better performance than other activation functions based on numerical tests
in this research. This autoencoder architecture requires that all input matrices have the same size; thus, the
size of the longest wind record is used as the size of the input matrices and zero padding is added to the
beginning and the end of all other shorter wind records. This strategy retains all information in the wind
records. In this example, the longest record has 156 data points (including the ramp-up and ramp-down)
with 10-min intervals, so the number of rows of the input matrices is 156.
The training of this autoencoder is conducted by minimizing the error between the reconstructed data in the output layer and the input data, which ensures that the latent features can represent the important patterns of the wind records. The Adaptive Moment Estimation (Adam) algorithm is adopted for stochastic optimization and batch normalization is added to some hidden layers as shown in Fig. 10 to address the exploding and vanishing gradient problems (Aggarwal 2018). Since the chosen batch size is 16, only 160 of the collected 162 wind records are used for the training process. In addition, Fig. 11 illustrates the reconstructed 160 wind records in the North and East directions after training the proposed autoencoder neural network. The histogram of the reconstruction error between the original and the reconstructed wind records is shown in Fig. 12, which demonstrates that the reconstructed records match well with the original ones and the latent features hold the most important characteristics of the wind records. It should be noted that since the 312 data points in the input layer is compressed into only 5 latent features, there must be some loss of information in this process and the discrepancies between the original and the reconstructed records are inevitable. However, these discrepancies are usually induced by noise or other nonsignificant factors;
therefore, the low dimensional latent features should be adequate for clustering because the important information has been extracted through the autoencoder.

Fig. 11. The 160 reconstructed wind records in two directions for the location of interest

Fig. 12. Histogram of the reconstruction error
Clustering and selection based on latent features

The location of interest studied in the previous sections is used here as an example. After the training process, all wind speed time series are converted into latent feature vectors, on which the k-means algorithm is applied for clustering. The goal of clustering is to maximize the similarity of data within each cluster and maximize the dissimilarity of data in distinct clusters. Therefore, one can take a subset of the data in a cluster to represent all data in that cluster, the accuracy of which depends on the number of clusters used.

Here, the elbow rule is adopted to find an optimal number of clusters (Thorndike 1953). To do so, the k-means algorithm has been run multiple times on the latent features with different number of clusters ranging from 2 to 20. For this example, when the number of clusters k equals 8, the Within-Cluster-Sum of Squared Errors (WSS) curve reaches its elbow as shown in Fig. 13. Therefore, the 160 hurricane wind records are divided into 8 clusters. Since it is difficult to show the 5 latent features on a 2D or 3D figure, principal component analysis is performed on the latent features and the first 3 principal components are plotted in Fig. 14 to demonstrate the results of the k-means clustering. This is acceptable because the first 3 principal components possess 82% of the variation of the 5 latent features and it is believed that the 5 latent features must show better performance than the 3 principal components if they can be plotted in a figure. In Fig. 14, the first 3 principal components are presented using 8 different colors for the 8 clusters, from which it may be seen that the hurricane wind records are clustered well because the principal components of different clusters have rare overlaps and the principal components of each cluster are gathered closely around their centroid.
To demonstrate the effectiveness of the proposed clustering approach, Fig. 15 to Fig. 22 illustrate the hurricane wind speeds and tracks of the 8 clusters. It is seen that the clustering results are successful, because hurricane wind speeds and durations within each cluster have similar patterns. Specifically,
Clusters 2 and 7 have hurricanes whose tracks pass by the East side of the location of interest, while other clusters have hurricanes whose tracks pass by the West side of the location of interest. The number of hurricanes in each cluster from Cluster 1 to Cluster 8 are 31, 17, 31, 13, 20, 12, 19 and 17, respectively. So much more hurricanes pass by the West side of the location of interest than the East side. The main difference between Cluster 2 and Cluster 7 is that the durations of hurricanes in Cluster 7 is longer. For the clusters passing by the West side of the location of interest, Cluster 3 and Cluster 4 has the shortest and the longest durations, respectively, while Clusters 1, 5, 6 and 8 have durations in the middle. Clusters 6 and 8 have very similar durations, but they are divided into two clusters because they have different shapes for the profile of the wind speed time histories. There are outliers in some clusters such as the one with abrupt changing of the storm heading direction as seen in the figure of hurricane tracks of Cluster 2. This can be expected because the k-means algorithm cannot eliminate all outliers, but instead assigns outliers to their closest cluster. Usually, outliers are rare and their latent feature points are far from the centroid of all points in a cluster. Therefore, the outlier commonly will not be included to the final suite of wind records considering the selection strategy within a cluster that will be introduced below. The wind field shown in Fig. 2 also has impacts on the clustering results, which cannot be explained explicitly here because its information is included in the latent features through the operations on the wind records during the training of the autoencoder.

Considering the computational demand of nonlinear time-history analyses that these wind records will be used to perform, approximately 1/10 of the hurricanes in each cluster are selected and combined together as the final suite of hurricane wind records. The number of records selected from each cluster is proportional to the total number of records in each cluster, which results in 3, 2, 3, 1, 2, 1, 2 and 2 records from each cluster, respectively. This strategy is used to make sure the proportions of different patterns of wind records are similar in the selected 16 hurricanes and the original 160 ones. It is also reasonable to make sure the selected records from each cluster are the most representative ones. To achieve this goal, the clustering results of the latent features are used, and for each cluster it is recommended to select those records whose
latent feature points are the closest to the centroid of all latent feature points in that cluster. The selected records for each cluster are highlighted in bold solid lines as shown in Fig. 15(a) to Fig. 22(a), which is a demonstration of the validity of this selection strategy within a cluster. In Fig. 15(a) to Fig. 22(a), all wind records are shown in different colors and curves resolved from the same record are shown in the same color in the upper subplot and lower subplot. It is seen that the selected records are representative, as they are near the middle of all the records. The total of 16 selected hurricanes can be employed to represent uncertainties in wind loading for design check and fragility development for structures at the location of interest. Note that these selected wind records are only time series of 10-min mean wind speed at 10 meters height. If one wants to use them for structural dynamic analysis, the fluctuating wind speeds and the atmospheric boundary layer should be considered.

(a) Wind records of Cluster 1
(b) Hurricane eye tracks of Cluster 1
Fig. 15. Hurricanes in Cluster 1

(a) Wind records of Cluster 2

(b) Hurricane eye tracks of Cluster 2
Fig. 16. Hurricanes in Cluster 2
Fig. 17. Hurricanes in Cluster 3
(a) Wind records of Cluster 3
(b) Hurricane eye tracks of Cluster 3
Fig. 18. Hurricanes in Cluster 4

(a) Wind records of Cluster 4

(b) Hurricane eye tracks of Cluster 4

Fig. 18. Hurricanes in Cluster 4
Fig. 19. Hurricanes in Cluster 5

(a) Wind records of Cluster 5

(b) Hurricane eye tracks of Cluster 5
Fig. 20. Hurricanes in Cluster 6

(a) Wind records of Cluster 6

(b) Hurricane eye tracks of Cluster 6

Fig. 20. Hurricanes in Cluster 6
Fig. 21. Hurricanes in Cluster 7

(a) Wind records of Cluster 7

(b) Hurricane eye tracks of Cluster 7
As a case study for regional analysis, Massachusetts has been divided into 92 grids and wind records have been collected for each grid. Here the procedures for wind records clustering and selection introduced in the previous sections are applied to all 92 grids. The same autoencoder architecture is used for all grids except for the slightly different sizes of the input vectors for different grids, which is because the maximum duration of the collected records for different grids may be different. The same k-means algorithm is also adopted for clustering on the latent features; however, the number of clusters may vary for different grids because it is dynamically determined using the elbow rule. The histogram of the number of clusters for all...
grids is presented in Fig. 23 with a mean value of 5.65. Since approximately 200 hurricane wind records are collected for each grid, then approximately 20 records are selected for each grid according to the method introduced in the previous sections. Finally, a wind map is generated so that a suite of hurricane wind speed and direction records can be provided for any locations in Massachusetts. For example, Fig. 24(a) gives 16 wind records selected from 8 clusters for a grid whose centroid has a latitude of 41.7 and a longitude of -70.1 (this location is used in the previous sections), while Fig. 24(b) gives 19 wind records selected from 4 clusters for a grid whose centroid has a latitude of 42.1 and a longitude of -72.5. Here the wind records selected from the same cluster are shown in the same color, and it is seen that wind records within the same cluster have similar characteristics in terms of wind speeds, directions, and durations. This approach provides an alternative to the ASCE 7 wind map. The ASCE 7 wind map can only provide a wind speed without any information of variation of the wind speed and direction during a hurricane. This methodology can be generalized to any other regions besides Massachusetts.

Fig. 23. Histogram of the number of clusters for all grids
(a) For a grid whose centroid has a latitude of 41.7 and a longitude of -70.1

(b) For a grid whose centroid has a latitude of 42.1 and a longitude of -72.5

Fig. 24. Examples of selected hurricane wind records

Conclusions
This paper presents a machine learning approach for collecting and selecting hurricane wind speed and direction records for a location and a region, which can be used for efficiently developing fragility curves or assessing probabilistic behaviors of structures considering uncertainties in hurricanes. The selected hurricane wind records are supposed to address the uncertainties in hurricanes because 1) they are selected
from 10,000-year synthetic hurricanes and 2) the collected records with similar properties are first divided into clusters and then the most representative ones are selected from each cluster. The preprocess of the wind records is also important since it can remove hurricanes with very small wind speeds and limit the durations of the records to a relatively short time. The proposed autoencoder architecture is shown to be able to reconstruct the wind speed time series and compress them into low dimensional latent features. The clustering results based on the latent features using the k-means algorithm are successful, because the points in the latent space are divided clearly into several clusters, and the wind records in the same cluster exhibit similar properties in duration, hurricane track, and changing of wind speed and directions. A method is also proposed to select the most representative records from each cluster based on the clustering results of latent features. This hurricane selection procedure is demonstrated using wind records from both a location and a region. For the regional hurricane selection, Massachusetts is used as a testbed, and it is discretized into a set of grids with performing the proposed hurricane selection procedure on each grid. Usually, approximate 20 wind records are selected for a location, which make the nonlinear structural analysis feasible for uncertainty propagation simulation under hurricanes.

**Data availability**

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. Some or all data, models, or code used during the study were provided by a third party. Direct requests for these materials may be made to the provider as indicated in the Acknowledgements.

**Acknowledgement**

The authors wish to thank Dr. Weichiang Pang of Clemson University for providing the synthetic hurricane catalog. The material presented in this paper is based upon work supported by National Science Foundation under Grant No. CRISP-1638234 and Northeastern University. This support is gratefully acknowledged.
References


