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

Xinlong Du[1], Jerome F. Hajjar[2], Robert Bailey Bond[3], Pu Ren[4] and Hao Sun[5]

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

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

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

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

[5]Associate 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

23   records are then compressed into a set of low-dimensional latent feature vectors using an artificial neural

24   network, designated as an autoencoder. The latent feature vectors represent the important patterns of wind

25   records such as duration, magnitude and the changing of wind speeds and directions over time. The wind

26   records are then clustered by applying the k-means algorithm on the latent features, and a subset of records

27   is selected from each cluster. The wind records selected from each cluster are those whose latent feature

28   points are closest to the centroid of all latent feature points in that cluster. In order to do regional analysis

29   while taking into account that the hurricane wind records are site-specific, this paper suggests that a region

30   can be discretized into a set of grids, with the proposed hurricane selection approach applied to each grid.

31   This procedure is demonstrated using Massachusetts as a testbed.

32   **Keywords:** hurricane selection, time series clustering, autoencoder, k-means, uncertainty quantification,

33   regional analysis, wind direction

Introduction

35   Nonlinear dynamic analysis is increasingly being considered in wind design of buildings and other

36   structures as performance-based design becomes an increasingly popular option (ASCE 2019), where

37   controlled inelastic deformations are allowed under strong winds (Wang and Wu 2022). In the fully

38   probabilistic performance-based hurricane engineering framework (Barbato et al. 2013), fragility curves of

39   structures are commonly adopted to do probabilistic damage assessment. Strength limit states of structures

40   usually involve nonlinear behavior that is then integrated into the predictions of likelihood of damage that

41   is offered through fragility analysis. Fragility functions are defined as the failure probability of a structure

42   conditional on the intensity measure of hazards, including hurricanes in this work. If only the failure

43   probability or fragility is of interest for a hurricane event (i.e., at the end of the loading time history) instead

44   of for a certain time interval within the hurricane duration, the uncertainties in the loading time histories

45   can be accounted for through running a series of nonlinear dynamic analysis with a suite of hurricane wind

46   records. The hurricane wind records should include time histories that incorporate the wind directions as

47   well as wind speeds, because the changing of wind directions during hurricanes has significant effects on

48 the structural response. Consequently, the wind records in this research are time series of both wind speed

49 and direction. To develop accurate fragility curves, the structures should be analyzed with a suite of

50 hurricane wind records that can cover the record-to-record uncertainties in the changing of wind speeds and

51 directions within the hurricane durations. A large amount of hurricane wind records can be collected for a

52 location considering the existing historical and synthetic hurricanes (ASCE 2016b; Vickery et al. 2010;

53 Vickery et al. 2009b; Vickery et al. 2009c). However, it is challenging to run nonlinear dynamic analysis

54 for all of the collected hurricane records due to the high computational demand of finite element analysis

55 of structures; thus, a minimum number of hurricane records should be selected to represent the uncertainties

56 in all of the collected hurricane records.

57 In prior work, Li (2005) and Li and Ellingwood (2006) developed hurricane fragility curves for wood-frame

58 residential construction with a simplified limit state function, where the nonlinear and dynamic effects are

59 neglected. Cui and Caracoglia (2015) carried out fragility analysis on tall buildings only for the

60 serviceability limit state, so frequency domain analysis is adopted, and duration and nonstationary effects

61 of hurricanes cannot be considered. In order to avoid performing structural analysis for long durations of

62 windstorms, the dynamic shakedown method was utilized by researchers to model the inelastic behavior of

63 buildings (Chuang and Spence 2019, 2020; Tabbuso et al. 2016). Other researchers tried to develop

64 hurricane fragility curves using nonlinear dynamic analysis only for a fixed time interval with a constant

65 wind direction. For example, Hallowell et al. (2018) used wind records with 1-hour time intervals, while

66 Ma et al. (2021) used wind records with 2-minute time intervals. The fragility developed for this certain

67 time interval cannot represent the fragility for a whole hurricane because of dynamic effects, yielding and

68 changes in wind speeds and directions. Of course, one can discretize the hurricane duration into a series of

69 short time intervals and apply the developed fragility curves to each short time interval; however, the failure

70 probabilities within those short time intervals are correlated (Der Kiureghian 2005; Kim et al. 2019; Straub

71 et al. 2020). This correlation is difficult to quantify from the view of time-variant reliability and is not

72 considered by the above authors. Given the limitations of the previous research, this paper considers the

73 failure probability for a hurricane event instead of a certain time interval during a hurricane and tries to

74 select hurricane wind records that can account for the record-to-record uncertainties in hurricanes. The

75 selected wind records can be used to estimate failure probabilities of structures with nonlinear time history

76 analysis. Through this way it is no longer needed to estimate the correlations of failure probabilities in the

77 short time intervals within a hurricane.

78 In performance-based earthquake engineering (Moehle and Deierlein 2004), a probabilistic framework has

79 been proposed to integrate seismic hazard analysis and structural damage analysis, where a suite of ground

80 motions are adopted to represent the uncertainties in earthquake ground motions. Ground motion selection

81 has been widely studied in the literature (Baker and Lee 2018; Bojórquez et al. 2013; Du and Padgett 2021;

82 Jayaram et al. 2011; Naeim et al. 2004). Some generally used ground motions suites are the SAC

83 (Somerville et al. 1997) records, LMSR (Krawinkler et al. 2003) records, and FEMA-P695 records (FEMA

84 2009). Recently, machine learning approaches have also been introduced to ground motion selection, where

85 a reduced number of ground motions are obtained through clustering of a large number of ground motions

86 (Bond et al. 2022; Kim et al. 2021; Zhang et al. 2020). However, there is no similar research in the literature

87 for selection of hurricane wind records. There are two instances of prior research that consider uncertainties

88 in hurricane wind hazards using a set of wind records, but they do not use a selection procedure. Vickery

89 et al. (2006) studied hurricane fragility curves for building envelope components that were developed in

90 the Hazus-MH software by comparing the wind pressure demand and the capacity of the envelope

91 components. The record-to-record uncertainties of hurricane wind speeds were accounted for through the

92 use of a 20,000-year simulation of hurricanes created by employing the hurricane model described by

93 Vickery et al. (2000a) and Vickery et al. (2000b). The simulated hurricanes inherently incorporated many

94 of the duration effects associated with the changes in wind speed and direction which accompany hurricane

95 winds. Joyner and Sasani (2018) developed fragility curves for the windborne debris damage of building

96 glazing, where eight hurricanes that made landfall in the U.S. in the last 10 years were adopted.

97 Uncertainties in the record-to-record variability for different hurricanes were accounted for by employing

4

98   the eight hurricane records in the damage analysis. Vickery et al. (2006) used all hurricanes in the 20,000-

99   year simulation, which may address the uncertainties in hurricanes, but is not suitable for nonlinear dynamic

100   analysis considering the computational demand. On the contrary, Joyner and Sasani (2018) only used eight

101   hurricanes without an analysis of the hazard uncertainties, which may not be able to represent the

102   uncertainties in hurricanes for a specific location.

103   This paper proposes a procedure to select a suite of hurricane wind records that can be used for performance-

104   based design and fragility analysis. The wind speed and direction records for a location are collected from

105   a synthetic hurricane catalog (Liu 2014) with some preprocessing, after which the collected wind records

106   have durations that are short enough to make a nonlinear time history analysis feasible. The collected wind

107   records are then compressed into low-dimensional latent feature vectors using a neural network designated

108   as an autoencoder (Aggarwal 2018), so that it is easier to measure similarity of different wind records and

109   apply the standard clustering algorithms such as the k-means algorithm (Aggarwal et al. 2001; Shalev-

110   Shwartz and Ben-David 2014). Autoencoder is an artificial neural network in which the input and output

111   layers have the same number of neurons, while the number of neurons in the middle is constricted. The

112   training algorithm tries to reconstruct the input data in the output layer; however, this reconstruction is not

113   exact because the neurons in the middle only carry a reduced representation of the input data. The data held

114   by the neurons in the middle (i.e., the low dimensional vectors compared to the input and output layer) are

115   called latent features, to which the clustering algorithm is applied. This means that only important

116   information in the wind records is preserved for clustering. The latent features representing hurricane wind

117   records are then clustered into several groups using the conventional k-means algorithm (Shalev-Shwartz

118   and Ben-David 2014). Finally, only a few hurricane wind records are selected from each cluster for fragility

119   development or design checks, which significantly reduces the number of required time history analyses,

120   while still ensuring that the uncertainties of different hurricanes are covered with a limited number of wind

121   records. Since the properties of hurricanes for different locations have significant differences, a hazard map

122   can be developed for hurricane wind records so that users are able to choose appropriate records for their

123   locations of interest. As an example, the Commonwealth of Massachusetts has been divided into 92 grids

124   and a suite of hurricane wind records has been selected for each grid using the proposed hurricane selection

125   approach. To demonstrate the usefulness and effectiveness of the selected wind records, they have been

126   adopted to develop fragility curves for electrical transmissions towers in Massachusetts, which can be found

127   in Du et al. (2022) and Du and Hajjar (2022).

Hurricane simulation

129   Synthetic hurricanes are widely used for risk analysis and structural design in wind engineering, which

130   evolved from the single site probabilistic model (Russell 1971) to Vickery's hurricane track model (Vickery

131   et al. 2000a). For example, ASCE 7 has adopted the hurricane track model when generating the wind hazard

132   maps (ASCE 2016b). This research also uses a 10,000 year synthetic hurricane catalog developed by Liu

133   (2014) for the Atlantic basin based on Vickery's hurricane track model, which consists of a hurricane

134   genesis model, a track model, an central pressure model, a decay model and a boundary layer model. The

135   temporal and spatial evolution of thousands of hurricanes from emergence to dissipation was modeled using

136   the Monte Carlo method. The HURDAT database (Jarvinen et al. 1984) for historical storms was adopted

137   for building the hurricane model through regression and calibrating the simulated results. In this simulation,

138   the state of a hurricane can be determined with 7 parameters: the hurricane eye's latitude and longitude,

139   storm translation speed, storm heading angle, storm central pressure, radius to maximum winds (describing

140   storm size), and Holland's radial pressure profile parameter (i.e., the Holland B parameter (Holland 1980)).

141   These parameters are updated at each 6-hour point. As suggested by Vickery et al. (2000a), linear

142   interpolation is performed within each 6-hour interval, which results in 10-min updates of the parameters

143   as used in Vickery et al. (2009c). Examples of the simulated hurricane tracks are shown in Fig. 1. In this

144   research, the gradient wind speeds are calculated by employing Georgiou's model (Georgiou 1985), which

145   gives the 10-min sustained wind speeds at 500 m to 2000 m above the ground surface (Cui and Caracoglia

146   2019; Pei et al. 2014, 2018). An example of the calculated gradient wind field is shown in Fig. 2.
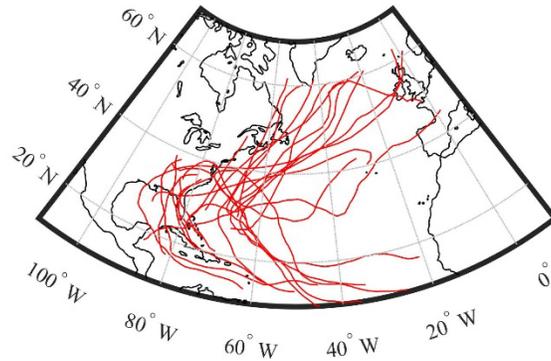
Fig. 1. Examples of the simulated hurricane tracks
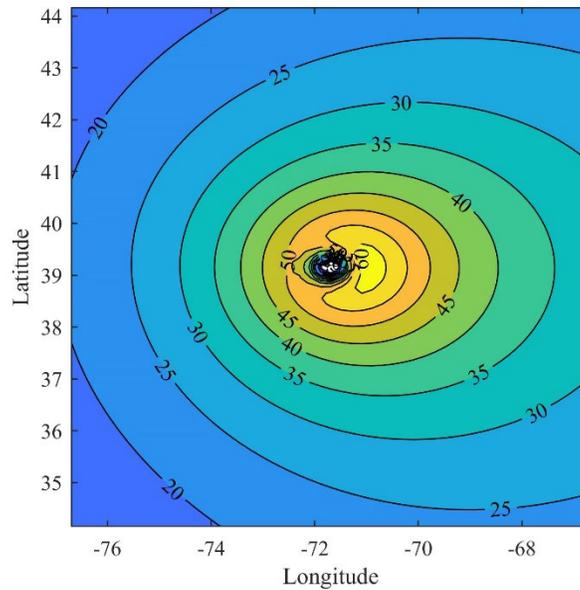
Fig. 2. Example of hurricane gradient wind field (m/s)

151 The obtained hurricane gradient wind speeds V_g need to be converted to surface wind speeds V_{10} (10 m

152 above the ground or water) for wind force calculation on structures. The reduction factor V_g/V_{10} over water

153 proposed by Batts et al. is used in this research (Batts et al. 1980; Vickery et al. 2009a). A sea-land transition

154 factor obtained from the model given in Simiu and Scanlan (1996) is then utilized to calculate the surface

155 wind over land (open terrain with surface roughness $z_0 = 0.03\ m$) from the surface wind over water ($z_0 =$

156 $0.0013\ m$). In addition, the surface wind speed over land approaches the fully transitioned value

157 asymptotically over a fetch distance as the wind moves from sea to land; therefore, the transition function

158 proposed in Vickery et al. (2009b) is employed here, which defines the percentage of the sea-land transition

159 as a function of the fetch distance. With the methods discussed in this section, the time series of the 10-min

160 sustained wind speeds at 10 m height and the corresponding wind directions at a location of interest

161 (assuming open terrain) during a hurricane may be obtained.

Hurricane wind records collection and preprocess

163 ASCE 7 wind hazard maps display wind speeds with a certain mean recurrence interval (MRI) for the entire

164 U.S., including hurricane prone regions (ASCE 2016b). However, much information regarding hurricane

165 winds is omitted in the ASCE 7 wind hazard maps, such as the variation of wind speeds and directions

166 during a hurricane, and the durations of hurricane winds. This kind of information, which are contained in

167 the time series of hurricane wind speeds and directions, are critical for structural response estimation and

168 risk analysis. Thus, in this section, a number of hurricane wind speed and direction records are collected

169 for a location of interest. In order to collect hurricane wind records for a region, the region is first discretized

170 into a series of grids and then hurricane wind records are collected for each grid.

Wind records for a location of interest

172 A location in Massachusetts with latitude 41.7 and longitude -70.1 is used as an example in this section.

173 Wind records are collected for this specific location from 10,000-year synthetic hurricanes developed by

174 Liu (2014). Examples of the collected 10-min sustained wind speed and wind direction records at the

175 location of interest are shown in Fig. 3 to Fig. 5 with the corresponding hurricane tracks. It is seen in Fig.

176 3(a), Fig. 4(a) and Fig. 5(a) that the hurricane eye usually moves thousands of miles from a hurricane's

177 genesis to dissipation. It is reasonable to assume that the wind speed induced by a hurricane that is very far

178 away is relatively small and can be neglected. Therefore, as suggested by Vickery et al. (2009c), hurricane

179 winds are considered only when the location of interest is within 250 km of the hurricane eye (see the blue

180 circles in Fig. 3 to Fig. 5). This limit on distance also provides a limit for the durations of the hurricane

181 wind records. Figure 3(c), Fig. 4(c) and Fig. 5(c) illustrate the absolute values of the wind speeds and the

182 wind directions in a polar coordinate system, while Fig. 3(d), Fig. 4(d) and Fig. 5(d) illustrate the hurricane

183 wind speeds in the North and East directions in a Cartesian coordinate system. Note that the wind direction

184   in the polar coordinate system is clockwise positive from the North direction. It is seen that the pattern of

185   wind speed and direction records are different for different hurricanes, which depends on a number of

186   factors, including the 7 parameters defining the hurricane eye tracks and wind fields. While the impacts of

187   the hurricane wind field as shown in Fig. 2 on the wind records is complex, a qualitative analysis of the

188   impact of hurricane eye tracks on the wind records provides examples of the range of loading developed

189   during hurricanes. Specifically, when the location of interest is very close to the hurricane eye track, the

190   record of the absolute values of the wind speeds usually has two peaks and the drop of the wind speed in

191   the middle is due to the near zero wind speed in the hurricane eye (see Fig. 3(c)). On the contrary, if the

192   location of interest is further from the hurricane eye track, the record of the absolute values of the wind

193   speeds will typically only have one peak (see Fig. 4(c) and Fig. 5(c)). The difference between Fig. 4 and

194   Fig. 5 is that the hurricane eye passes by the West or East side of the location of interest, which dominates

195   the variation of the wind directions as presented in these two figures.
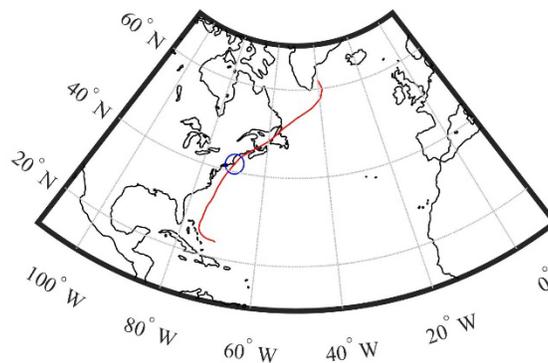
196
197   (a) The whole hurricane track (the blue circle represents the 250 km limit)

9
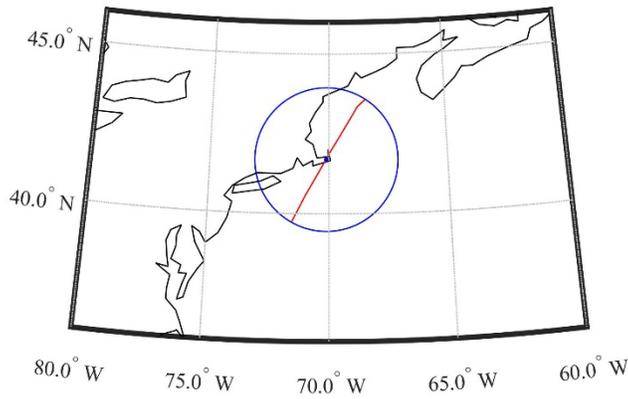
198
199   (b) The hurricane track within the 250 km limit (blue circle) of the location of interest (blue dot)
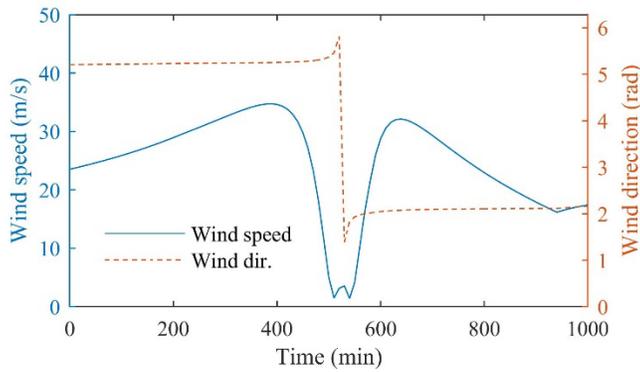
200
201   (c) Wind speed and direction records
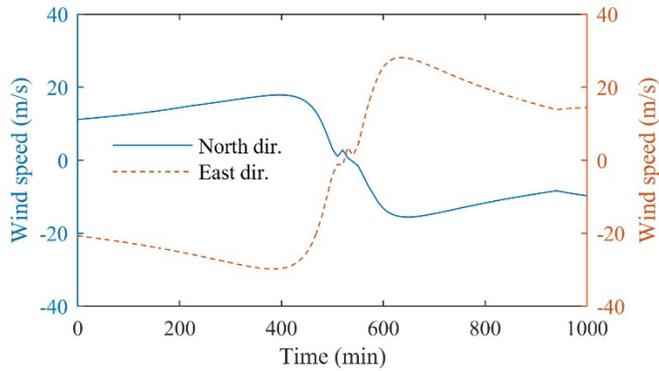
202
203   (d) Wind speed records in the North and East directions
204   Fig. 3. An example of hurricanes going through the location of interest

10
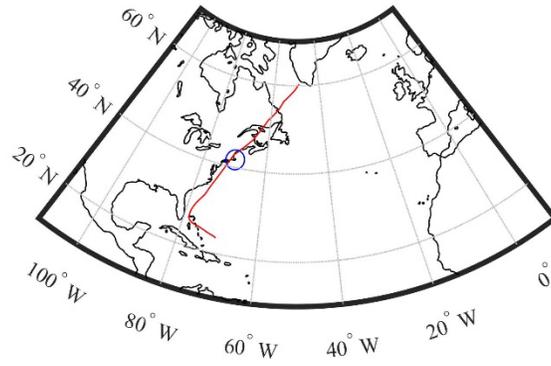
205
206   (a) The whole hurricane track (the blue circle represents the 250 km limit)
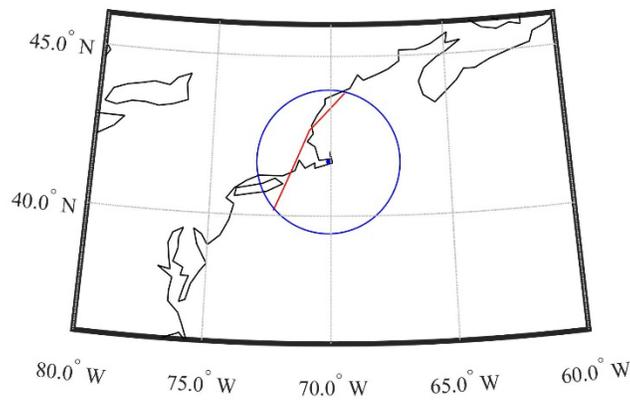
207
208   (b) The hurricane track within the 250 km limit (blue circle) of the location of interest (blue dot)

209
210   (c) Wind speed and direction records
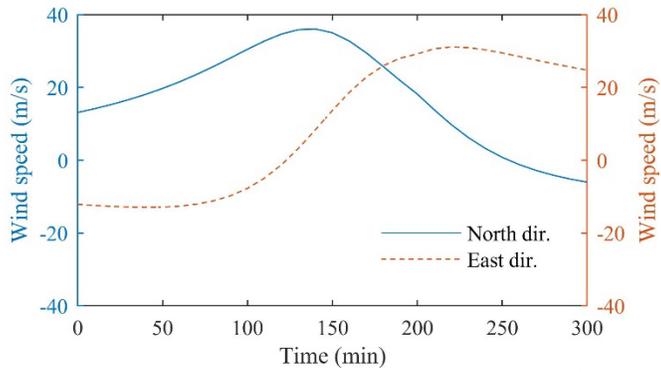
211
212   (d) Wind speed records in the North and East directions
213   Fig. 4. An example of hurricanes passing by the West side of the location of interest
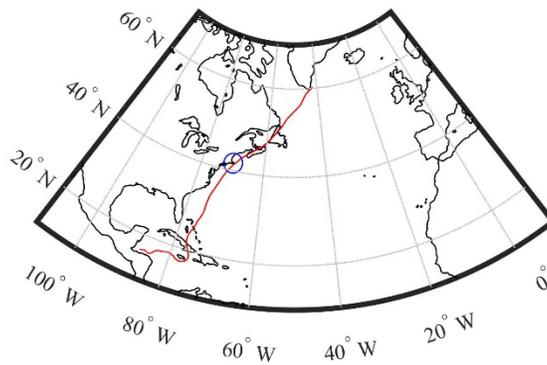
214
215   (a) The whole hurricane track (the blue circle represents the 250 km limit)

216
217   (b) The hurricane track within the 250 km limit (blue circle) of the location of interest (blue dot)

12
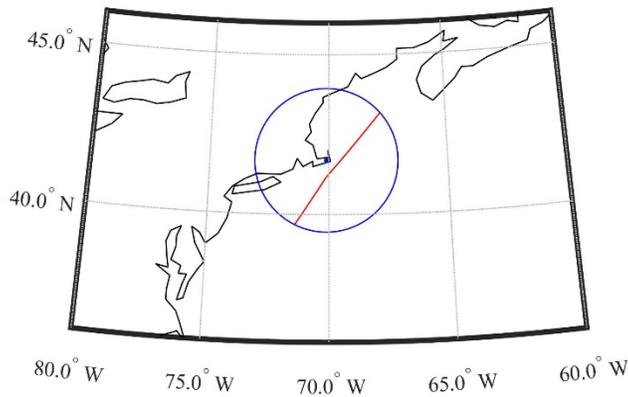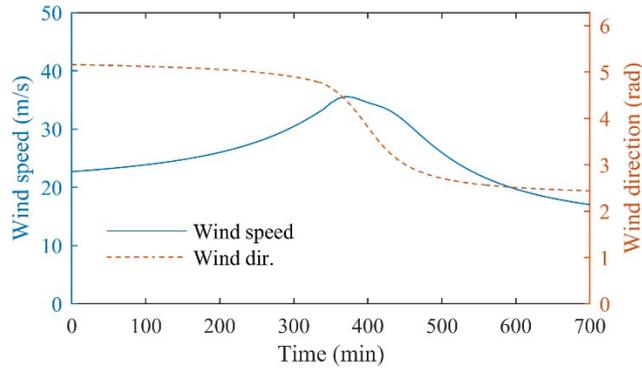
(c) Wind speed and direction records
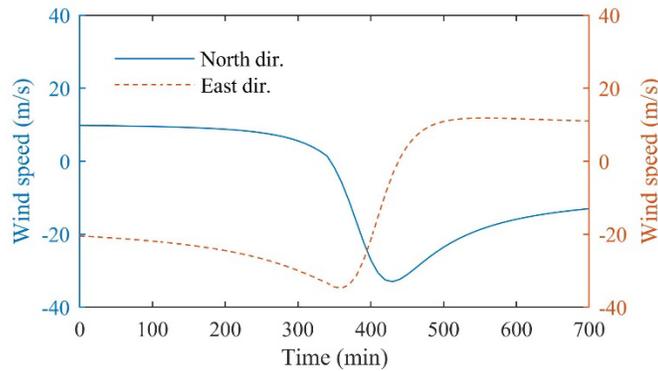
(d) Wind speed records in the North and East directions

Fig. 5. An example of hurricanes passing by the East side of the location of interest

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 hurricanes whose maximum wind speeds at the location of interest are greater than the 50-year MRI wind speed at the same location are considered. 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

13

235   autoencoder, as will be introduced in the following sections. In addition, Fig. 7 presents the histogram of

236   the durations of all of the collected hurricanes with a mean duration of 12.4 hours. To avoid the impulse

237   effects, a 1-hour linear ramp-up and a 1-hour linear ramp-down are attached to the beginning and the end

238   of the collected wind records, respectively, as recommended in the Prestandard for Performance-Based

239   Wind Design (ASCE 2019). To be consistent with the hurricane wind records with 10-min intervals, the

240   ramp is added as six 10-min steps with a constant wind direction. Note that the ramps are not included in

241   Fig. 3 to Fig. 5 but included in Fig. 6 and Fig. 7. Moreover, as will be discussed in the following sections,

242   the collected wind records may have different durations, but the autoencoder needs the same size for the

243   input data of each record. Therefore, to facilitate training the autoencoder, zero paddings are added to the

244   beginning and the end of the records that are shorter than the longest one. Consequently, all records after

245   preprocess have the same length as the longest one. For each record, zero paddings at the beginning and the

246   end have the same length, which means all records after preprocessing have a midpoint that is usually

247   recorded when the hurricane eye is closest to the location of interest.
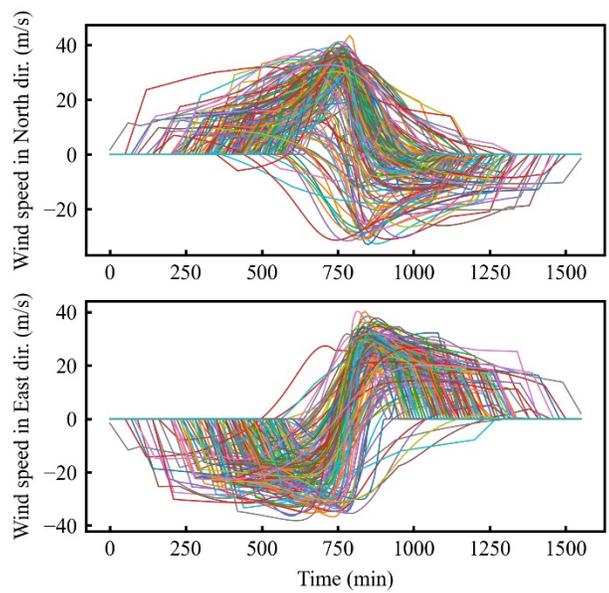
248
249   Fig. 6. The 160 collected hurricane wind records resolved in two directions

14
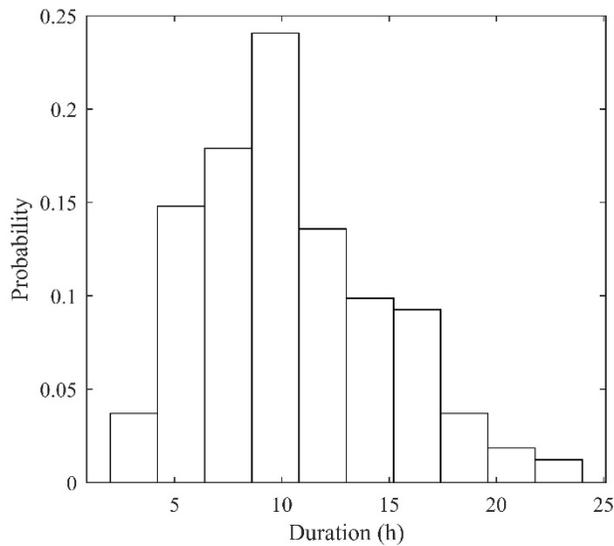
Fig. 7. Histogram of the durations of the collected hurricane wind records

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,

267 locations within the same grids but on the land are used to find the 50-year MRI wind speeds. Figure 9

268 presents the histogram of the number of hurricanes collected for all 92 grids, with a mean value of 202.
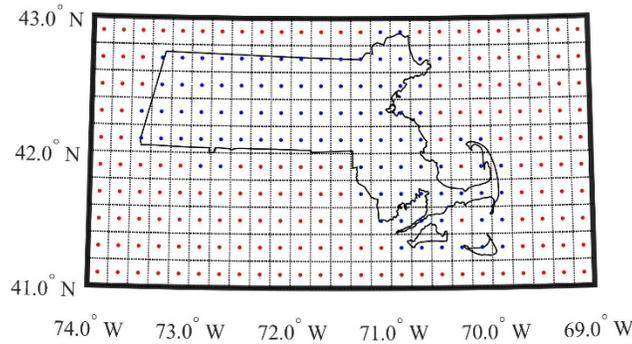
269
270   Fig. 8. Massachusetts is discretized into grids
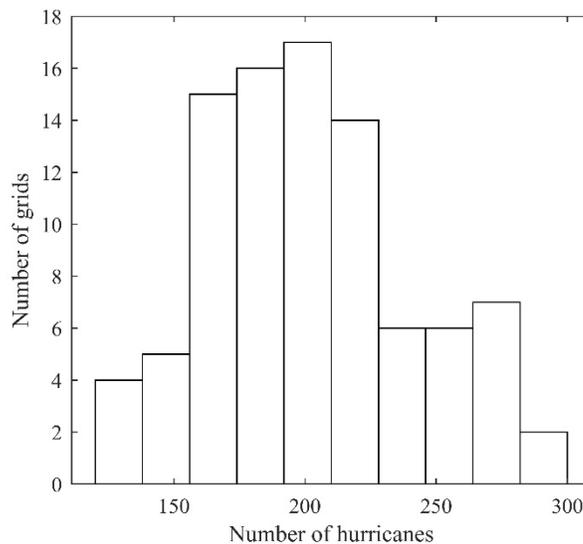
271
272   Fig. 9. Histogram of the number of collected hurricanes for the 92 grids

Wind records clustering and selection

274 The approximately 200 collected hurricane wind records for each grid are still too many for design checks

275 and fragility development, especially considering the long durations of the wind records. Incremental

276 dynamic analysis (IDA) may be used to estimate collapse probability of structures under hurricanes (Du et

277 al. 2022; Vamvatsikos and Cornell 2002). This approach is computationally intensive because direct

278 integration of the nonlinear dynamic governing equations is required over the entire duration of the

279 hurricane wind records and this nonlinear time history analysis needs to be run multiple times with scaled

280   wind records. As such, it is important to limit the number of records used. Therefore, in this research, the

281   collected wind records for each grid are first clustered using a machine learning approach and then

282   approximately 1/10 of the wind records in each cluster are selected, which are combined together to create

283   approximately 20 selected wind records for each grid. This significantly reduces the number of nonlinear

284   time history analyses required, while still preserving the uncertainties in the collected records. This

285   procedure is similar to stratified sampling in statistics. Sampling is the process of selecting a subset from a

286   population so that the characteristics of the whole population can be estimated using this subset, while

287   stratified sampling is used by dividing the population into subpopulations (i.e., clusters in this paper), where

288   the elements within each subpopulation are similar, and performing sampling on each subpopulation.

289   Stratified sampling may improve the precision of the sample because sampling variability within each

290   subpopulation is smaller than the sampling variability on the entire population (Botev and Ridder 2017;

291   Parsons 2014). Specifically, the selected wind records can cover a spread of properties such as durations,

292   patterns of wind speed records, and patterns of wind direction records, because the collected wind records

293   are divided into clusters based on these properties.

294   *Fully connected autoencoder*

295   Since the collected hurricane wind records are time series of both wind speed and direction with different

296   durations, it is challenging to cluster the records directly. To facilitate the clustering process, the high

297   dimensional wind records are first transformed into low dimensional latent features using an artificial neural

298   network named autoencoder (Aggarwal 2018; Bond et al. 2022; Tavakoli et al. 2020). The architecture of

299   the autoencoder for wind records at the location of interest given in the previous section is presented in Fig.

300   10. It is seen that the input matrix is the original wind speed records in the Cartesian system, which has two

301   columns with each column representing wind speed time histories in the North and East directions,

302   respectively. The input matrix is first flattened into a vector as the input layer of the fully connected

303   autoencoder and then passed through other hidden layers to reconstruct the data as another vector in the

304   output layer, which is finally reshaped to a matrix as the reconstructed wind speed records in the Cartesian

305 system. Even though the two columns of the input matrix are correlated time series of wind speeds in two

306 directions, this "flatten" and "reshape" process is reasonable because the correlations are considered in the

307 flattened vectors (input and output layers) through the weights of the fully connected layers. In another

308 word, "flatten" and "reshape" only change the appearance of the data while retaining the relationships and

309 correlations of the elements within the data. A fully connected autoencoder means that all the neurons in

310 one layer are connected to all the neurons in the next layer. The autoencoder architecture consists of two

311 parts: the encoder that compresses the high dimensional input data into the small-size latent feature vector,

312 and the decoder that utilizes the latent features to reconstruct the input data. In this example, the flattened

313 wind speeds in the input layer are transformed into 5 latent features through the encoder process, which are

314 then expanded to form the reconstructed but still flattened wind records in the output layer through the

315 decoder process. The hidden layers with a nonlinear activation function (Tanh) are included to enhance the

316 power of this autoencoder so that it can map the input data into much smaller dimensional spaces. Here

317 Tanh is adopted because it has better performance than other activation functions based on numerical tests

318 in this research. This autoencoder architecture requires that all input matrices have the same size; thus, the

319 size of the longest wind record is used as the size of the input matrices and zero padding is added to the

320 beginning and the end of all other shorter wind records. This strategy retains all information in the wind

321 records. In this example, the longest record has 156 data points (including the ramp-up and ramp-down)

322 with 10-min intervals, so the number of rows of the input matrices is 156.
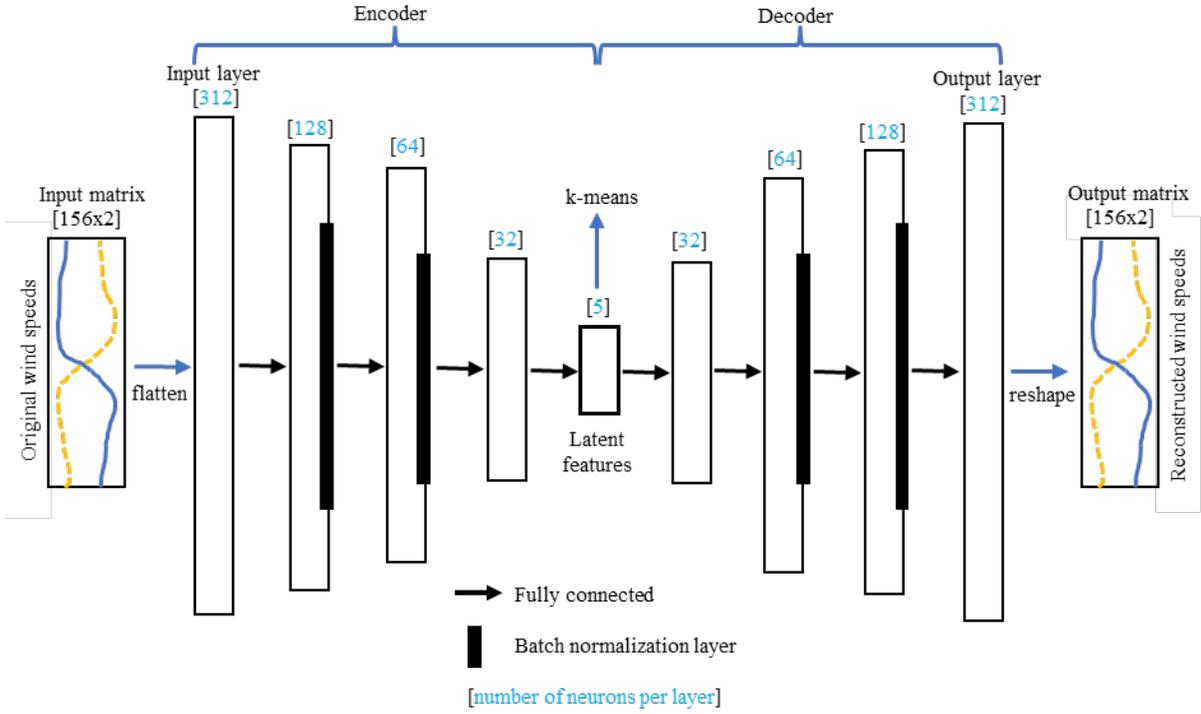
Fig. 10. The proposed autoencoder architecture

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;

338   therefore, the low dimensional latent features should be adequate for clustering because the important

339   information has been extracted through the autoencoder.
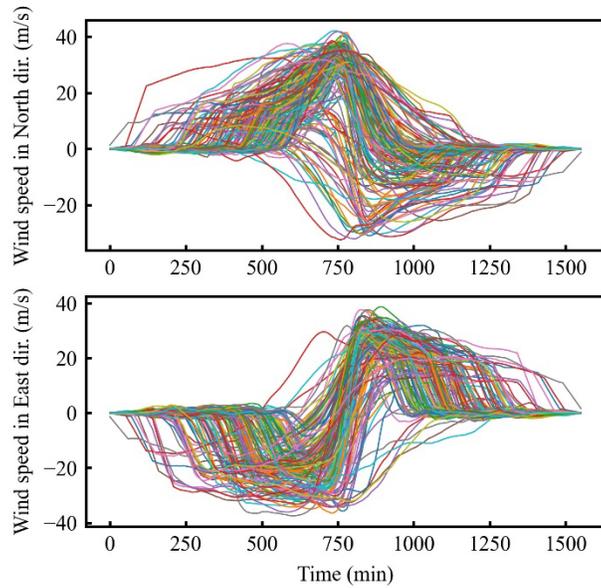
340
341   Fig. 11. The 160 reconstructed wind records in two directions for the location of interest
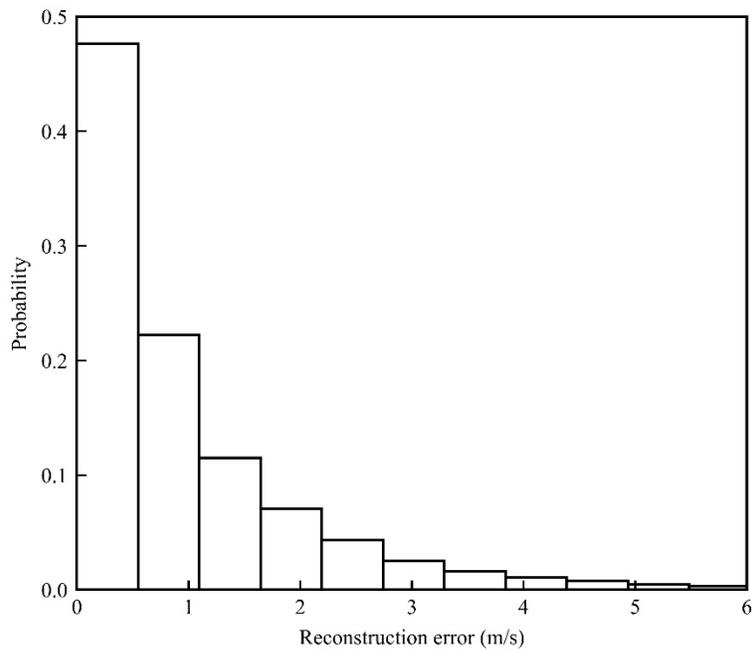
342
343   Fig. 12. Histogram of the reconstruction error

344

345

Clustering and selection based on latent features

346 The location of interest studied in the previous sections is used here as an example. After the training

347

348 process, all wind speed time series are converted into latent feature vectors, on which the k-means algorithm

349 is applied for clustering. The goal of clustering is to maximize the similarity of data within each cluster and

350 maximize the dissimilarity of data in distinct clusters. Therefore, one can take a subset of the data in a

351 cluster to represent all data in that cluster, the accuracy of which depends on the number of clusters used.

352 Here, the elbow rule is adopted to find an optimal number of clusters (Thorndike 1953). To do so, the k-

353 means algorithm has been run multiple times on the latent features with different number of clusters ranging

354 from 2 to 20. For this example, when the number of clusters k equals 8, the Within-Cluster-Sum of Squared

355 Errors (WSS) curve reaches its elbow as shown in Fig. 13. Therefore, the 160 hurricane wind records are

356 divided into 8 clusters. Since it is difficult to show the 5 latent features on a 2D or 3D figure, principal

357 component analysis is performed on the latent features and the first 3 principal components are plotted in

358 Fig. 14 to demonstrate the results of the k-means clustering. This is acceptable because the first 3 principal

359 components possess 82% of the variation of the 5 latent features and it is believed that the 5 latent features

360 must show better performance than the 3 principal components if they can be plotted in a figure. In Fig. 14,

361 the first 3 principal components are presented using 8 different colors for the 8 clusters, from which it may

362 be seen that the hurricane wind records are clustered well because the principal components of different

363 clusters have rare overlaps and the principal components of each cluster are gathered closely around their

364 centroid.
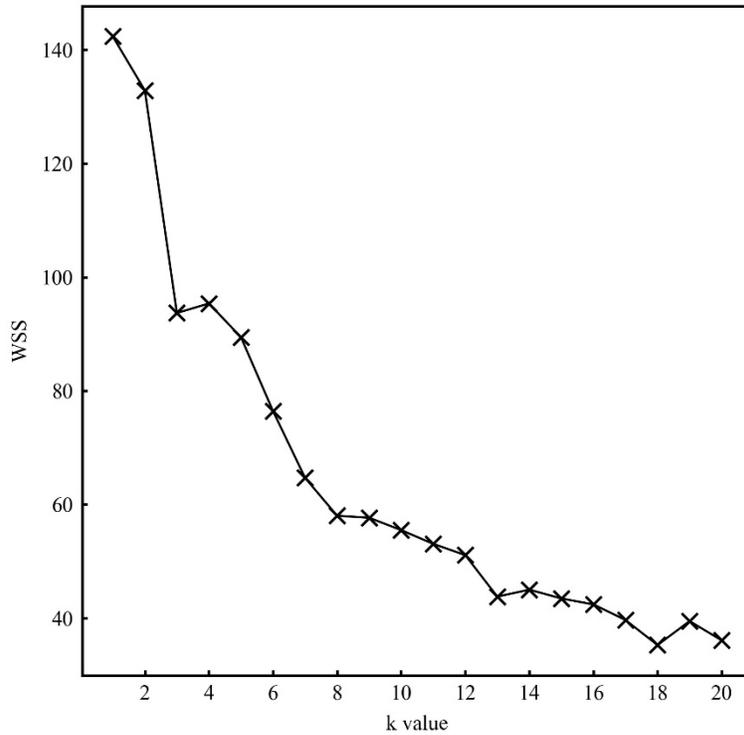
365
366 Fig. 13. The WSS for different number of clusters
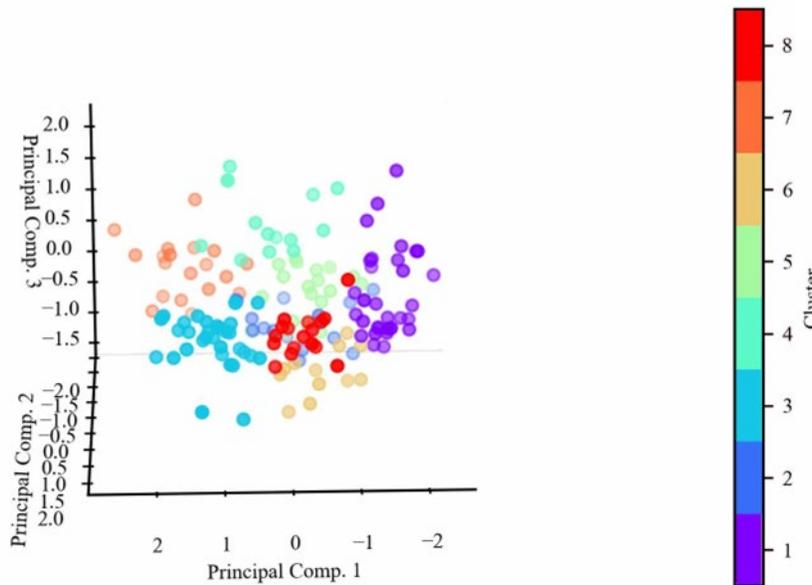
367
368 Fig. 14. Principal components of the latent features for the 8 clusters

369 To demonstrate the effectiveness of the proposed clustering approach, Fig. 15 to Fig. 22 illustrate the

370 hurricane wind speeds and tracks of the 8 clusters. It is seen that the clustering results are successful,

371 because hurricane wind speeds and durations within each cluster have similar patterns. Specifically,

372 Clusters 2 and 7 have hurricanes whose tracks pass by the East side of the location of interest, while other

373 clusters have hurricanes whose tracks pass by the West side of the location of interest. The number of

374 hurricanes in each cluster from Cluster 1 to Cluster 8 are 31, 17, 31, 13, 20, 12, 19 and 17, respectively. So

375 much more hurricanes pass by the West side of the location of interest than the East side. The main

376 difference between Cluster 2 and Cluster 7 is that the durations of hurricanes in Cluster 7 is longer. For the

377 clusters passing by the West side of the location of interest, Cluster 3 and Cluster 4 has the shortest and the

378 longest durations, respectively, while Clusters 1, 5, 6 and 8 have durations in the middle. Clusters 6 and 8

379 have very similar durations, but they are divided into two clusters because they have different shapes for

380 the profile of the wind speed time histories. There are outliers in some clusters such as the one with abrupt

381 changing of the storm heading direction as seen in the figure of hurricane tracks of Cluster 2. This can be

382 expected because the k-means algorithm cannot eliminate all outliers, but instead assigns outliers to their

383 closest cluster. Usually, outliers are rare and their latent feature points are far from the centroid of all points

384 in a cluster. Therefore, the outlier commonly will not be included to the final suite of wind records

385 considering the selection strategy within a cluster that will be introduced below. The wind field shown in

386 Fig. 2 also has impacts on the clustering results, which cannot be explained explicitly here because its

387 information is included in the latent features through the operations on the wind records during the training

388 of the autoencoder.

389 Considering the computational demand of nonlinear time-history analyses that these wind records will be

390 used to perform, approximately 1/10 of the hurricanes in each cluster are selected and combined together

391 as the final suite of hurricane wind records. The number of records selected from each cluster is proportional

392 to the total number of records in each cluster, which results in 3, 2, 3, 1, 2, 1, 2 and 2 records from each

393 cluster, respectively. This strategy is used to make sure the proportions of different patterns of wind records

394 are similar in the selected 16 hurricanes and the original 160 ones. It is also reasonable to make sure the

395 selected records from each cluster are the most representative ones. To achieve this goal, the clustering

396 results of the latent features are used, and for each cluster it is recommended to select those records whose

23

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.
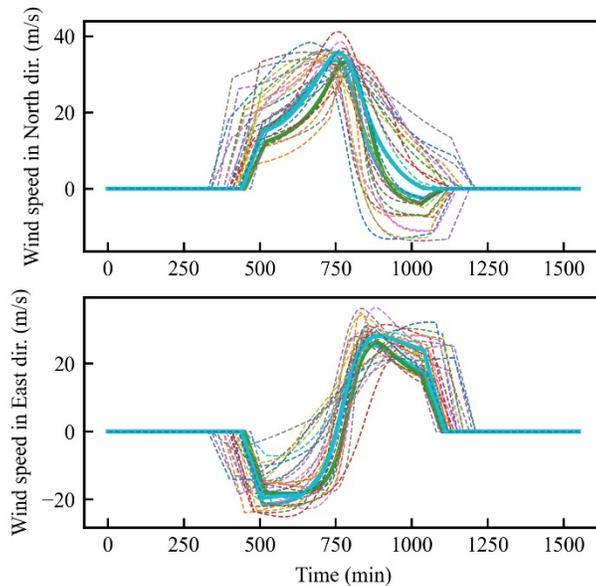
(a) Wind records of Cluster 1

24
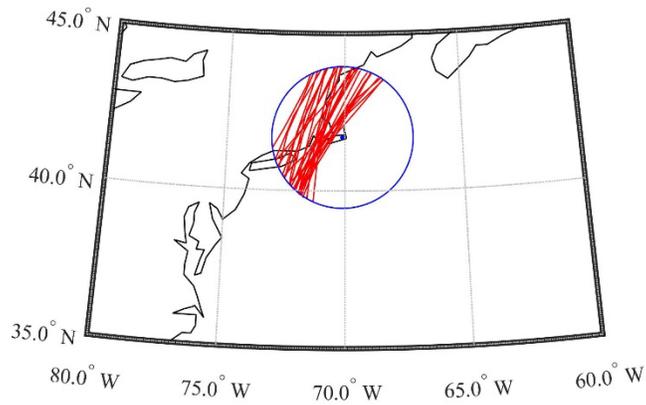
(b) Hurricane eye tracks of Cluster 1

Fig. 15. Hurricanes in Cluster 1
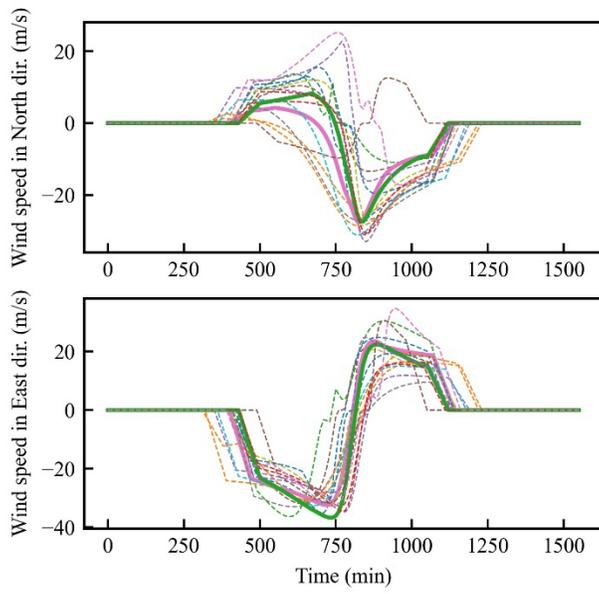
(a) Wind records of Cluster 2
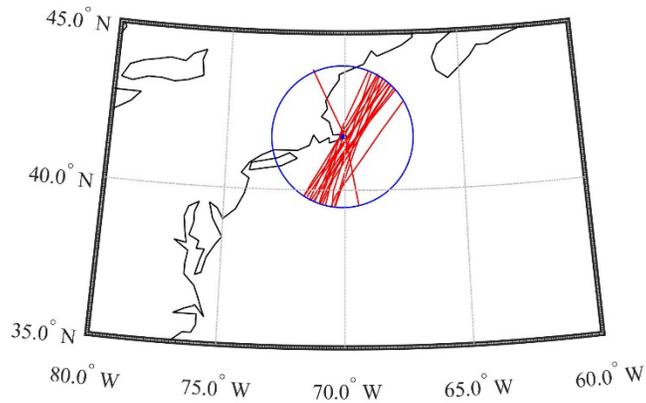
(b) Hurricane eye tracks of Cluster 2

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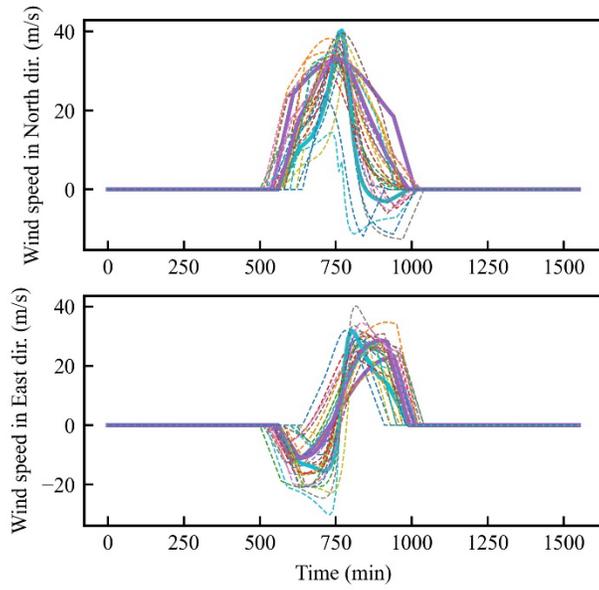
(a) Wind records of Cluster 3
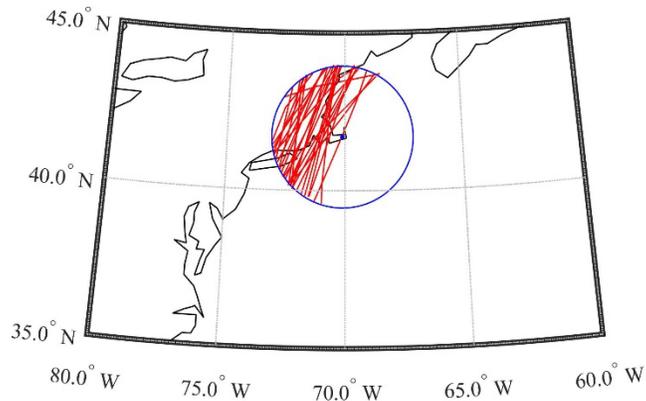
(b) Hurricane eye tracks of Cluster 3

Fig. 17. Hurricanes in Cluster 3
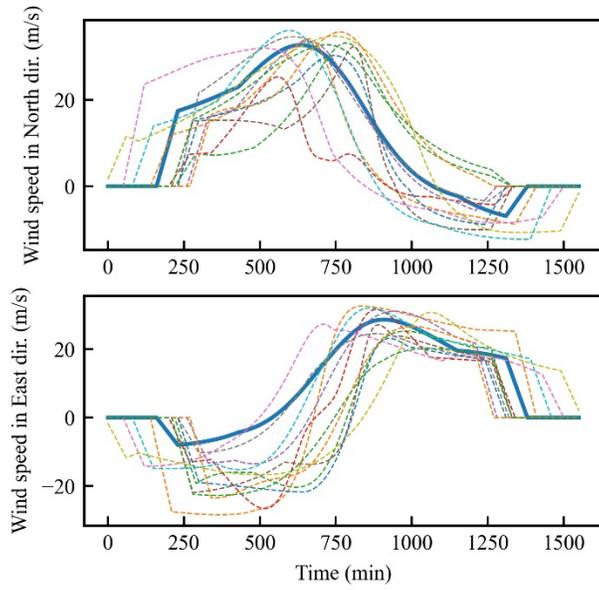
422
423   (a) Wind records of Cluster 4
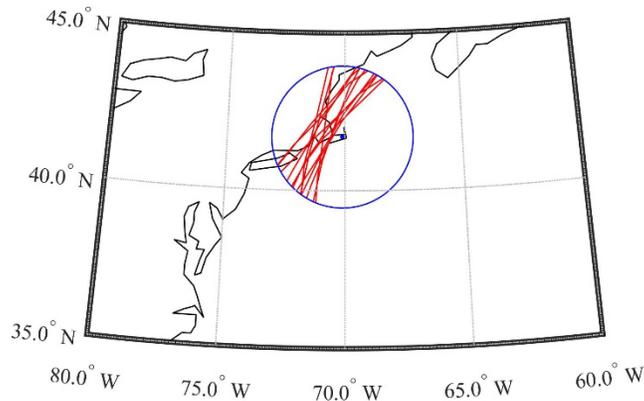
424
425   (b) Hurricane eye tracks of Cluster 4
426   Fig. 18. Hurricanes in Cluster 4
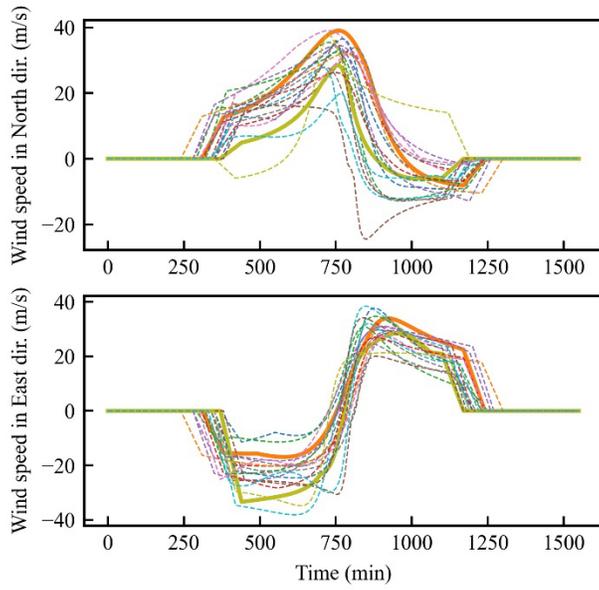
(a) Wind records of Cluster 5
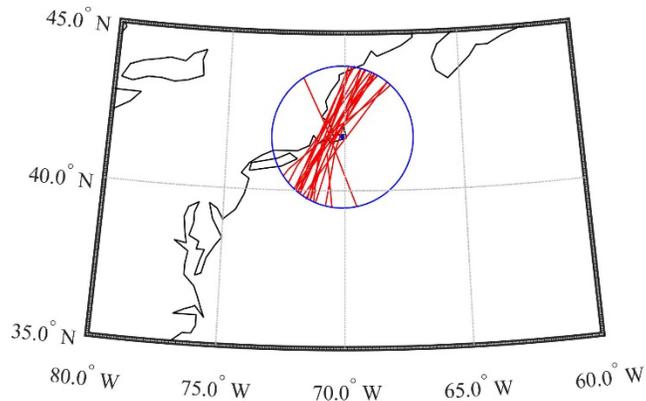
(b) Hurricane eye tracks of Cluster 5

Fig. 19. Hurricanes in Cluster 5

432
433
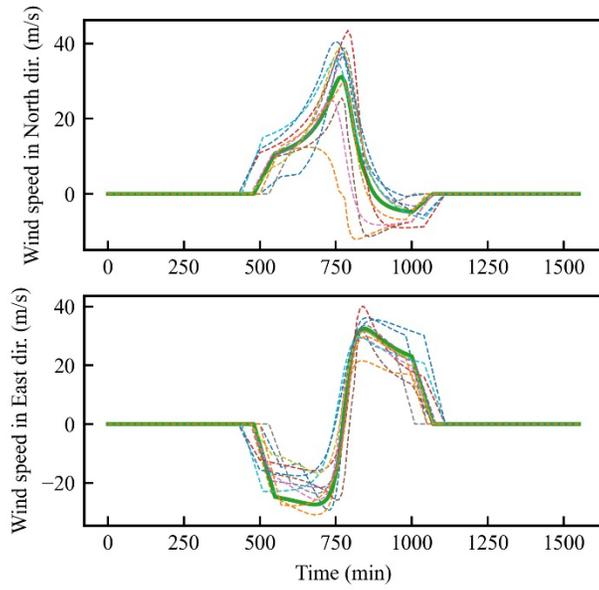
(a) Wind records of Cluster 6
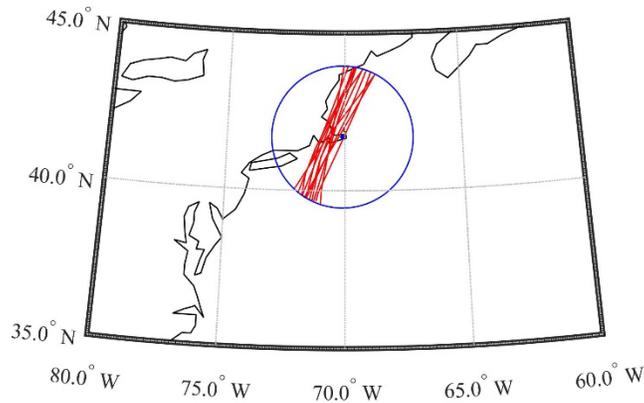
434
435 (b) Hurricane eye tracks of Cluster 6
436 Fig. 20. Hurricanes in Cluster 6
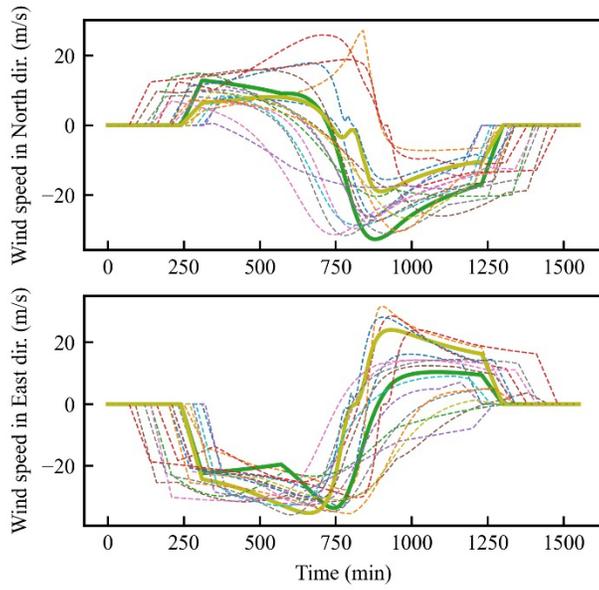
(a) Wind records of Cluster 7

437
438
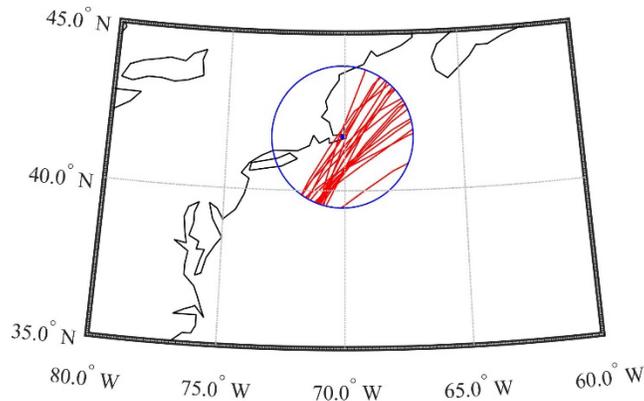
439
440
441

(b) Hurricane eye tracks of Cluster 7
Fig. 21. Hurricanes in Cluster 7

442
443
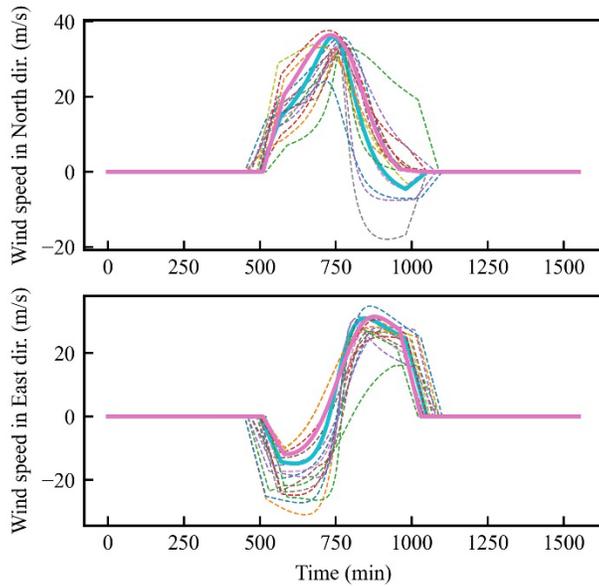
(a) Wind records of Cluster 8
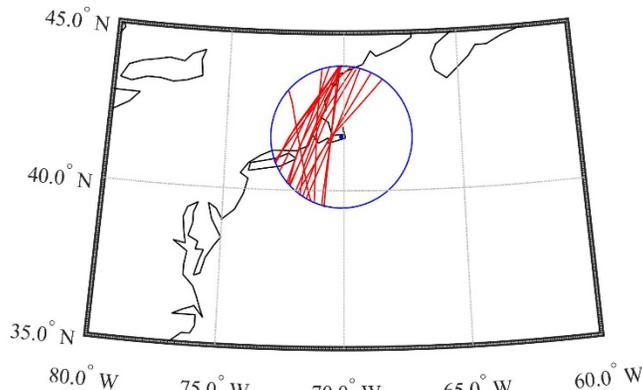
444
445 (b) Hurricane eye tracks of Cluster 8
446 Fig. 22. Hurricanes in Cluster 8

Wind records selection for a region

447

448 As a case study for regional analysis, Massachusetts has been divided into 92 girds and wind records has

449 been collected for each grid. Here the procedures for wind records clustering and selection introduced in

450 the previous sections are applied to all 92 grids. The same autoencoder architecture is used for all grids

451 except for the slightly different sizes of the input vectors for different grids, which is because the maximum

452 duration of the collected records for different grids may be different. The same k-means algorithm is also

453 adopted for clustering on the latent features; however, the number of clusters may vary for different grids

454 because it is dynamically determined using the elbow rule. The histogram of the number of clusters for all

31

455   grids is presented in Fig. 23 with a mean value of 5.65. Since approximately 200 hurricane wind records

456   are collected for each grid, then approximately 20 records are selected for each grid according to the method

457   introduced in the previous sections. Finally, a wind map is generated so that a suite of hurricane wind speed

458   and direction records can be provided for any locations in Massachusetts. For example, Fig. 24(a) gives 16

459   wind records selected from 8 clusters for a grid whose centroid has a latitude of 41.7 and a longitude of -

460   70.1 (this location is used in the previous sections), while Fig. 24(b) gives 19 wind records selected from 4

461   clusters for a grid whose centroid has a latitude of 42.1 and a longitude of -72.5. Here the wind records

462   selected from the same cluster are shown in the same color, and it is seen that wind records within the same

463   cluster have similar characteristics in terms of wind speeds, directions, and durations. This approach

464   provides an alternative to the ASCE 7 wind map. The ASCE 7 wind map can only provide a wind speed

465   without any information of variation of the wind speed and direction during a hurricane. This methodology

466   can be generalized to any other regions besides Massachusetts.
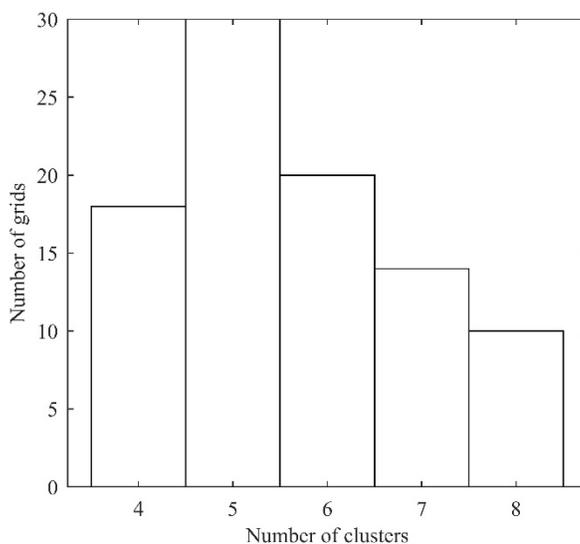
467
468   Fig. 23. Histogram of the number of clusters for all grids
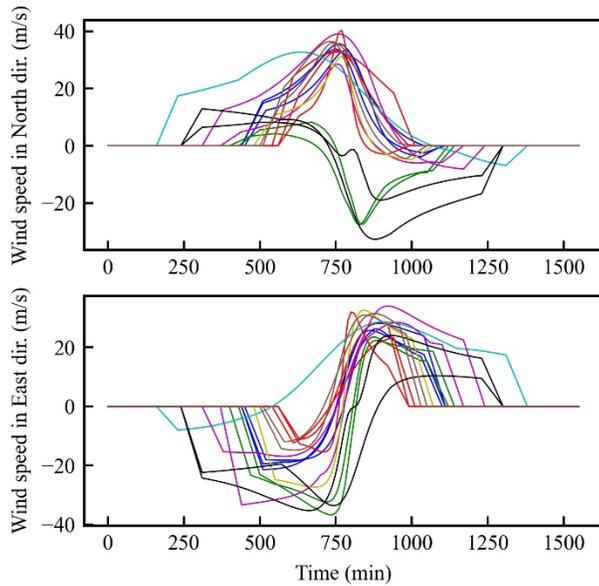
(a) For a grid whose centroid has a latitude of 41.7 and a longitude of -70.1
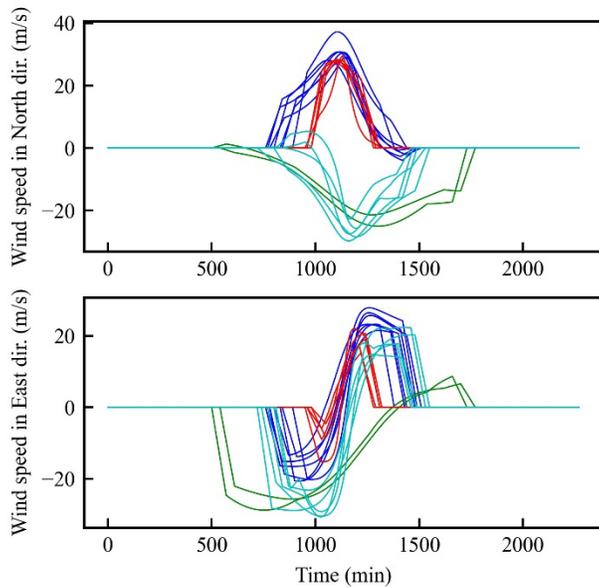
(b) For a grid whose centroid has a latitude of 42.1 and a longitude of -72.5

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

33

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

References

Aggarwal, C. C. 2018. *Neural networks and deep learning*. Yorktown Heights, NY: Springer International Publishing AG, part of Springer Nature.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. 2001. "On the surprising behavior of distance metrics in high dimensional space." In *Proc., International conference on database theory*, 420-434. Springer, Berlin, Heidelberg.

ASCE. 2016a. "ASCE 7 Hazard Tool." Accessed October 17, 2021. https://asce7hazardtool.online/.

ASCE. 2016b. *Minimum Design Loads and associated criteria for Buildings and Other Structures*. ASCE 7-16. Reston, VA: ASCE.

ASCE. 2019. *Prestandard for Performance-Based Wind Design*. Reston, VA: ASCE.

Baker, J. W., and Lee, C. 2018. "An improved algorithm for selecting ground motions to match a conditional spectrum." *Journal of Earthquake Engineering*, 22(4), 708-723.

Barbato, M., Petrini, F., Unnikrishnan, V. U., and Ciampoli, M. 2013. "Performance-based hurricane engineering (PBHE) framework." *Structural Safety*, 45, 24-35.

Batts, M. E., Cordes, M. R., Russell, L. R., Shaver, J. R., and Simiu, E. 1980. "Hurricane wind speeds in the United States." National Bureau of Standards, Report Number BSS-124, US Department of Commerce.

Bojórquez, E., Reyes-Salazar, A., Ruiz, S. E., and Bojórquez, J. 2013. "A New Spectral Shape-Based Record Selection Approach Using and Genetic Algorithms." *Mathematical Problems in Engineering*, 2013, 679026.

Bond, R. B., Ren, P., Hajjar, J. F., and Sun, H. 2022. "An unsupervised machine learning approach for ground motion clustering and selection." *Earthquake Spectra*. (under review).

Botev, Z., and Ridder, A. 2017. "Variance reduction." *Wiley statsRef: Statistics reference online*, 1-6.

Chuang, W.-C., and Spence, S. M. 2019. "An efficient framework for the inelastic performance assessment of structural systems subject to stochastic wind loads." *Engineering Structures*, 179, 92-105.

Chuang, W.-C., and Spence, S. M. 2020. "Probabilistic performance assessment of inelastic wind excited structures within the setting of distributed plasticity." *Structural Safety*, 84, 101923.

Cui, W., and Caracoglia, L. 2015. "Simulation and analysis of intervention costs due to wind-induced damage on tall buildings." *Engineering Structures*, 87, 183-197.

Cui, W., and Caracoglia, L. 2019. "A new stochastic formulation for synthetic hurricane simulation over the North Atlantic Ocean." *Engineering Structures*, 199, 109597.

Der Kiureghian, A. 2005. "First-and second-order reliability methods." Chap. 14 in *Engineering design reliability handbook*, edited by E. Nikolaidis, D. Ghiocel, and S. Singhal. Boca Raton, FL: CRC Press.

Du, A., and Padgett, J. E. 2021. "Refined multivariate return period-based ground motion selection and implications for seismic risk assessment." *Structural Safety*, 91, 102079.

Du, X., and Hajjar, J. F. 2022. "Nonlinear dynamic analysis and fragility development of electrical transmission towers under hurricanes." In *Proc., Annual Stability Conference*. Chicago, IL: Structural Stability Research Council.

Du, X., Hajjar, J. F., Bond, R. B., and Sun, H. 2022. "Collapse fragility development of electrical transmission towers subjected to hurricanes." In *Proc., IABSE Symposium Prague 2022, Challenges for Existing and Oncoming Structures*, 235-242. Zurich, Switzerland: International Association for Bridge and Structural Engineering (IABSE).

FEMA. 2009. *Quantification of building seismic performance factors*. FEMA-P695. Redwood City, CA: Applied Technology Council.

Georgiou, P. N. 1985. "Design wind speeds in tropical cyclone-prone regions." PhD Dissertation, University of Western Ontario, London, Ontario, Canada.

Hallowell, S. T., Myers, A. T., Arwade, S. R., Pang, W., Rawal, P., Hines, E. M., Hajjar, J. F., Qiao, C., Valamanesh, V., and Wei, K. 2018. "Hurricane risk assessment of offshore wind turbines." *Renewable Energy*, 125, 234-249.

553 Holland, G. J. 1980. "An analytic model of the wind and pressure profiles in hurricanes." *Monthly weather*

554   *review*, 108(8), 1212-1218.

555 Jarvinen, B. R., Neumann, C. J., and Davis, M. A. 1984. "A tropical cyclone data tape for the North Atlantic

556   Basin, 1886-1983: Contents, limitations, and uses." NOAA Technical Memorandum No. NWS-

557   NHC-22, US Department of Commerce, Washington, D.C.

558 Jayaram, N., Lin, T., and Baker, J. W. 2011. "A computationally efficient ground-motion selection

559   algorithm for matching a target response spectrum mean and variance." *Earthquake Spectra*, 27(3),

560   797-815.

561 Joyner, M. D., and Sasani, M. 2018. "Multihazard Risk-Based Resilience Analysis of East and West Coast

562   Buildings Designed to Current Codes." *Journal of Structural Engineering*, 144(9), 04018156.

563 Kim, S.-M., Ok, S.-Y., and Song, J. 2019. "Multi-scale dynamic system reliability analysis of actively-

564   controlled structures under random stationary ground motions." *KSCE Journal of Civil Engineering*,

565   23(3), 1259-1270.

566 Kim, T., Kwon, O. S., and Song, J. 2021. "Clustering-based adaptive ground motion selection algorithm

567   for efficient estimation of structural fragilities." *Earthquake Engineering & Structural Dynamics*,

568   50(6), 1755-1776.

569 Krawinkler, H., Medina, R., and Alavi, B. 2003. "Seismic drift and ductility demands and their dependence

570   on ground motions." *Engineering structures*, 25(5), 637-653.

571 Li, Y. 2005. "Fragility methodology for performance-based engineering of wood-frame residential

572   construction." PhD Dissertation, Georgia Institute of Technology, Atlanta, GA.

573 Li, Y., and Ellingwood, B. R. 2006. "Hurricane damage to residential construction in the US: Importance

574   of uncertainty modeling in risk assessment." *Engineering structures*, 28(7), 1009-1018.

575 Liu, F. 2014. "Projections of future US design wind speeds and hurricane losses due to climate change."

576   PhD Dissertation, Clemson University, Clemson, SC.

577   Ma, L., Khazaali, M., and Bocchini, P. 2021. "Component-based fragility analysis of transmission towers
578   subjected to hurricane wind load." *Engineering Structures*, 242, 112586.

579   Moehle, J., and Deierlein, G. G. "A framework methodology for performance-based earthquake
580   engineering." In *Proc., 13th world conference on earthquake engineering*. Vancouver, B.C.,
581   Canada.

582   Naeim, F., Alimoradi, A., and Pezeshk, S. 2004. "Selection and scaling of ground motion time histories for
583   structural design using genetic algorithms." *Earthquake spectra*, 20(2), 413-426.

584   Parsons, V. L. 2014. "Stratified sampling." *Wiley StatsRef: Statistics Reference Online*, 1-11.

585   Pei, B., Pang, W., Testik, F. Y., Ravichandran, N., and Liu, F. 2014. "Mapping joint hurricane wind and
586   surge hazards for Charleston, South Carolina." *Natural hazards*, 74(2), 375-403.

587   Pei, B., Pang, W., Testik, F. Y., Ravichandran, N., and Liu, F. 2018. "Selection of hazard-consistent
588   hurricane scenarios for regional combined hurricane wind and flood loss estimation." *Natural
589   Hazards*, 91(2), 671-696.

590   Russell, L. R. 1971. "Probability distributions for hurricane effects." *Journal of the Waterways, Harbors
591   and Coastal Engineering Division*, 97(1), 139-154.

592   Shalev-Shwartz, S., and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*.
593   New York, NY: Cambridge University Press.

594   Simiu, E., and Scanlan, R. H. 1996. *Wind effects on structures: fundamentals and applications to design
595   (3rd ed.)*. New York, NY: John Wiley & Sons, Inc.

596   Somerville, P., Smith, N., Punyamurthula, S., and Sun, J. 1997. *Development of ground motion time
597   histories for phase 2 of the FEMA/SAC steel project*. Richmond, CA: SAC Joint Venture.

598   Straub, D., Schneider, R., Bismut, E., and Kim, H.-J. 2020. "Reliability analysis of deteriorating structural
599   systems." *Structural safety*, 82, 101877.

Tabbuso, P., Spence, S. M., Palizzolo, L., Pirrotta, A., and Kareem, A. 2016. "An efficient framework for the elasto-plastic reliability assessment of uncertain wind excited systems." *Structural Safety*, 58, 69-78.

Tavakoli, N., Siami-Namini, S., Khanghah, M. A., Soltani, F. M., and Namin, A. S. 2020. "An autoencoder-based deep learning approach for clustering time series data." *SN Applied Sciences*, 2(5), 1-25.

Thorndike, R. L. 1953. "Who belongs in the family?" *Psychometrika*, 18(4), 267-276.

Vamvatsikos, D., and Cornell, C. A. 2002. "Incremental dynamic analysis." *Earthquake Engineering & Structural Dynamics*, 31(3), 491-514.

Vickery, P., Skerlj, P., and Twisdale, L. 2000a. "Simulation of hurricane risk in the US using empirical track model." *Journal of structural engineering*, 126(10), 1222-1237.

Vickery, P. J., Masters, F. J., Powell, M. D., and Wadhera, D. 2009a. "Hurricane hazard modeling: The past, present, and future." *Journal of Wind Engineering and Industrial Aerodynamics*, 97(7-8), 392-405.

Vickery, P. J., Skerlj, P., Steckley, A., and Twisdale, L. 2000b. "Hurricane wind field model for use in hurricane simulations." *Journal of Structural Engineering*, 126(10), 1203-1221.

Vickery, P. J., Skerlj, P. F., Lin, J., Twisdale Jr, L. A., Young, M. A., and Lavelle, F. M. 2006. "HAZUS-MH hurricane model methodology. II: Damage and loss estimation." *Natural Hazards Review*, 7(2), 94-103.

Vickery, P. J., Wadhera, D., Galsworthy, J., Peterka, J. A., Irwin, P. A., and Griffis, L. A. 2010. "Ultimate wind load design gust wind speeds in the United States for use in ASCE-7." *Journal of structural engineering*, 136(5), 613-625.

Vickery, P. J., Wadhera, D., Powell, M. D., and Chen, Y. 2009b. "A hurricane boundary layer and wind field model for use in engineering applications." *Journal of Applied Meteorology and Climatology*, 48(2), 381-405.

624   Vickery, P. J., Wadhera, D., Twisdale Jr, L. A., and Lavelle, F. M. 2009c. "US hurricane wind speed risk

625   and uncertainty." *Journal of structural engineering*, 135(3), 301-320.

626   Wang, H., and Wu, T. 2022. "Statistical investigation of wind duration using a refined hurricane track

627   model." *Journal of Wind Engineering and Industrial Aerodynamics*, 221, 104908.

628   Zhang, R., Hajjar, J., and Sun, H. 2020. "Machine learning approach for sequence clustering with

629   applications to ground-motion selection." *Journal of Engineering Mechanics*, 146(6), 04020040.

630