1	
2	Forecasting and Identifying the Meteorological and Hydrological Conditions Favoring the
3	Occurrence of Severe Hazes in Beijing and Shanghai using Deep Learning
4	
5	
6	Chien Wang
7	
8	Laboratoire d'Aerologie, CNRS and University Paul Sabatier
9	14 Avenue Edouard Belin, 31400 Toulouse, France
10	
11	
12	
13	March 2021
14	
15	
16	
17	
18	
19	
20	
21	
22	Correspondence to: Chien Wang (chien.wang@aero.obs-mip.fr)
23	
24	

25 Abstract. Severe haze or low visibility event caused by abundant atmospheric aerosols has

26 become a serious environmental issue in many countries. A framework based on deep

- 27 convolutional neural networks has been developed to forecast the occurrence of such events in
- 28 two Asian megacities: Beijing and Shanghai. Trained using time sequential regional maps of
- 29 meteorological and hydrological variables alongside surface visibility data over the past 41
- 30 years, the machine has achieved a good overall accuracy in associating the haze events with
- favorite meteorological and hydrological conditions. <u>Certain valuable knwoledge has also</u>
   obtained from the training such as the sensitivity of the machine's performance to the spatial
- scale of feature patterns that could benefit future applications using meteorological and

35 scale of reature patients that could benefit future applications using meteorological and
 34 hydrological data. Furthermore, an unsupervised cluster analysis using features with a greatly

34 <u>invertorogical data.</u> Furthermore, an unsupervised cluster analysis using features with a greatly 35 reduced dimensionality produced by the trained machine has, arguably for the first time,

- successfully categorized typical regional meteorological-hydrological regimes alongside local
- 30 successfully categorized typical regional intercological-hydrological regimes alongside loca
   37 quantities associated with haze and non-haze events in the two targeted cities, providing
- substantial insights to advance our understandings of this environmental extreme.

# 39 1 Introduction

40 Frequent low visibility or haze events caused by elevated abundance of atmospheric aerosols

41 due to fossil fuel and biomass burning <u>has have</u> become a serious environmental issue in many

Asian countries in recent decades, interrupting economic and societal activities and causing
human health issues (*e.g.*, Chan and Yao, 2008; Silva et al., 2013; Lee *et al.*, 2017). For example,

rapid economic development and urbanization in China have caused various pollution-related

45 health issues particularly in populated metropolitans such as Beijing-Tianjin region and <del>Yangtze</del>

46 <u>riverYangtze River</u> delta centered in Shanghai (*e.g.*, Liu *et al.*, 2017). In Singapore, the total

47 economic cost of brought by severe hazes events in 2015 is estimated to be \$510 million (0.17%

of <u>the GDP</u>), or \$643.5 million based on a wiling-to-pay analysis (Lin *et al.*, 2016). To ultimately
 prevent this detrimental environmental extreme from happening requires rigid emission control
 measures in place through significant changes in energy consumption as well as land and

50 measures in place through significant changes in energy consumption as well as land and 51 plantation management. Before all these measures could finally take place, it would be more 52 practical to develop skills to accurately predict its-the occurrence of hazes hence to allow

53 mitigation measures to be implemented ahead of time.

Severe haze events arise from the solar radiation extinction by aerosols in the atmosphere,
 this mechanism can be enhanced with the increase of relative humidity that enlarges the size of
 particles (*e.g.*, Kiehl and Briegleb 1993). Aerosols also need favorite atmospheric transport and
 mixing conditions to reach places away from their immediate source locations, while their
 lifetime in the atmosphere can be significantly reduced by rainfall removal. In addition, soil

58 moisture is also a key to dust emissions. Therefore, meteorological and hydrological conditions

- 60 are critical to the occurrence of haze events besides particulate emissions. To forecast the
- 61 occurrence of such events using existing atmospheric numerical models developed based on fluid
- 62 dynamics and explicit or parameterized representations of physical and chemical processes, the
- 63 actual task is to accurately predict the concentration of aerosols at a given geographic location

and a given time in order to correctly derive surface visibility (*e.g.*, Lee *et al.* 2017 & 2018).

65 However, the propagation of numerical or parameterization errors through the model integration

66 could easily drift the model away from the original track, not mentioning that lack of real-time

emission data alone would simply handicap such an attempt. Therefore, a more fundamental
issue in practice is whether these models could reproduce the *a posteriori* distribution of the

possible outcomes of the targeted low-probability extreme events. Ultimately, lack of knowledge
about the extreme events would, in turn, hinder the effort to improve the forecasting skills.

Differing from the deterministic models, an alternative statistical prediction approach could 71 72 be adopted, should if the predictors of a targeted event could be identified and a statistical correlation between them could be established with confidence. However, this is a rather difficult 73 74 task for the traditional approaches, because it requires an analysis dealing with a very large 75 quantity of high-dimensional data in order toto establish a likely multi-variate and nonlinear 76 correlation of generalization that can be generalized. Nevertheless, such attempts can obviously benefit now from the fast-growing machine learning (ML) and deep learning (DL) algorithm 77 78 development (e.g., LeCun et al., 2015). In addition, technological advancement and continuous 79 investment from governments and other sectors across the world have led to a rapid increase of 80 quantity alongside substantially improved quality of meteorological, oceanic, hydrological, land, and atmospheric composition data. These data might still not be sufficient for evaluating and 81 improving certain detailed aspects of the deterministic forecasting models. However, rich 82 83 information contained in these data about favorite environmental conditions for the occurrence of extreme events such as hazes could already have a great value for developing alternative 84

85 forecasting skills.

86 Many Earth science applications dealing with meteorological or hydrological data need a trained machine to not only forecast values but also recognize patterns or images. However, this 87 88 can easily lead to a curse of dimensionality of many traditional ML algorithms. Fortunately, deep 89 learning that directly links large quantity of raw data with targeted outcomes through deep convolutional neural networks or CNNs (Goodfellow et al., 2016) offers a clear advantage in 90 91 sufficiently training deep networks suitable for solving highly nonlinear issues. In doing so, DL 92 can also eliminate the possible mistakes in data derivation or selection introduced by subjective 93 human opinion regarding a poorly understood phenomenon. Recently, DL algorithms have been 94 explored in various applications in atmospheric, climate, and environmental sciences, ranging 95 from recognizing specific weather patterns (e.g., Liu et al., 2016; Kurth et al., 2018; Lagerquist et al., 2019; Chattopadhyay et al., 2020), weather forecasting including hailstorm detection (e.g., 96 97 Grover et al., 2015; Shi et al., 2015; Gagne et al., 2019), to deriving model parameterizations 98 (e.g., Jiang et al., 2018), and beyond.

99 When In certain applications, the targeted outcomes are the same features as the input but at 100 a different time, *e.g.*, a given weather feature(s) such as temperature or pressure at a given level. patterns associated with targeted outcome are known or irrelevant to the task. The the forecasting 101 102 can thus be normally proceeded to recognize a given pattern by using pattern-to-pattern 103 correlation from a sequential training dataset with spatial-information-preserving full CNNs such as U-net (Ronneberger et al., 2015; Weyn et al., 2020). However, this is certainly not the case 104 for the applications where the environmental conditions associated with targeted outcome are yet 105 known. For such applications, a possible solution is to utilize a large quantity of raw data with 106 107 minimized human intervention in data selection to train a deep CNN in order toto associate targeted outcomes with favorite environmental conditions. This study represents such an attempt, 108 where a DL forecast framework is trained to identify the meteorological and hydrological 109 110 conditions associated with the occurrences of severe hazes. The DL framework has been 111 developed initially with the severe hazes in Singapore (Wang, 2020), and now hazes in two megacities of China, Beijing and Shanghai. In terms of particulate pollutant emissions, all these 112 113 cities share certain sources including fossil fuel combustions from transportation, domestic, and

114 industries. On the other hand, each city also has its own unique sources, for instance, desert and

- 115 perhaps anthropogenic dust for Beijing, and massive biomass burning in Singapore (Chen *et al.*,
- 116 2013; Liu *et al.*, 2017; Lee *et al.*, 2017, 2018, & 2019). It is obvious that besides meteorological
- 117 and hydrological conditions, dynamical patterns of anthropogenic activities leading to the
- 118 emissions of particulate matters are also important factors behind the occurrence of severe hazes.
- 119 Nevertheless, the major purpose of this study is to advance our fundamental knowledge about the
- weather conditions favoring the occurrence of hazes and, through an in-depth analysis on the
- forecasting results to identify the limit of such a machine and thus to provide useful information
- 122 for establishing a more complete forecasting platform for the task.
- In the paper, the architecture alongside method and data for training are firstly described after this <u>i</u>Introduction, followed by a discussion of training and validation results. Then, an unsupervised cluster analysis benefited from the trained machine is introduced along with the results that furthers the understanding of the CNN's performance and summarizes, for the first time, the various typical meteorological and hydrological regimes associated with haze versus
- non-haze situations in the two cities. The last section concludes the effort and major efforts and
- 129 findings.

# 130 2 Network Architecture, Training Methodology and Data

131 The convolutional neural network used in this study, the HazeNet (Wang 2020), has been

developed by adopting the general architecture of the CNN developed by the Oxford

133 University's Visual Geometry Group or VGG-Net (Simonyan and Zisserman, 2015). The actual

134 structure alongside hyper-parameters of HazeNet have been adjusted and fine-tuned based on

numerous test trainings. In addition, certain techniques that were not available when the original

136 VGG net was developed, *e.g.*, batch normalization (Ioffe and Szegedy, 2015), have been

137 included as well. The current version for haze applications of Beijing and Shanghai, though

138 <u>trained separately</u>, contains <u>the same number of parameters of</u> 20,507,161 parameters (11,376

139 non-trainable) owing to the same optimized kernel sizes. Figure 1 shows the general architecture

140 of a HazeNet version with 12 convolutional and 4 dense layers (in total 57 layers).

Conv = ZeroPadding2D + Conv2D + BatchNormalization Output 2 Conv + Dropout Sigmoid or + Maxpool Softmax 2 Conv + Dropout + 2 Conv + Dropout 1, 2, or 3... + Maxpool + Maxpool 2 Conv + Dropout + 2 Conv + Dropout 512 3x3 (1x1) filters + Maxpool + Maxpool 2 Conv + Dropout 192 6x6 filters 2 Dense Layers 384 3x3 filters + Maxpool 4096 Input Layer 92 10x10 or **WxHxN** 20x20 filters



**Figure 1**. Architecture of the 12 convolutional plus 4 dense layer HazeNet. Here "Conv" represents a unit containing a zero-padding then a 2D convolutional layer, followed by a batch normalization layer. There

is a flatten layer before the 2 dense layers. W = width, H = height, and N = number of features of the
input fields, they are 64, 96, and 16 for Beijing, and 64, 64, and 16 for Shanghai case, respectively.

146 The network has been trained in a standard supervised learning procedure for classification-147 In this procedure, where the network takes input features to produce classification output that are 148 then compared with known results or labels based on observations. The coefficients of the network are thereafter optimized in order to minimize the error between the prediction and the 149 150 observation or label. The loss function used in optimization is cross-entropy (e.g., Goodfellow et 151 al., 2017). Such a procedure is repeated until the performance of the network can no longer be improved. In practice, the trainings usually last about 2000 epochs (each epoch is a training cycle 152 that uses up the entire training dataset). This procedure in nature is to train a deep CNN to 153 recognize then associate input features (bundled meteorological and hydrological conditions in 154 this case) with corresponding class, *i.e.*, severe haze events or non-haze events. As a result, the 155 knowledge specifically about the favorite meteorological and hydrological conditions of severe 156 157 hazes could be advanced.

158 The labels for the training are derived using the observed daily surface visibility (*vis.* 

159 thereafter), obtained from the Global Surface Summary Of the Day or GSOD dataset consisting 160 of daily observations of meteorological conditions from tens of thousands of airports around the 161 globe (Smith et al., 2011). In the cases of Beijing and Shanghai, data are from observations in 162 corresponding airports of these two cities during the time period time from 1979 to 2019, 163 containing 14,975 samples. For simplicity, the discussions will be mainly on the 2-class training, where events with vis.  $\leq$  the long-term mean value of the 25<sup>th</sup> percentile or p25 of vis. (6.27 km 164 165 in Beijing, 5.95 km in Shanghai; Fig. 2, right panel; also Fig. S1 in Supplementary) are defined 166 as class 1 or severe hazes, otherwise the class 0 or non-haze cases. The Although p25 values vary 167 interannually, their long-term means actually represent a substantial reduction of vis. due to high particulate pollution (e.g., Lee et al., 2017). Note that unlike in the case of Singapore (Wang 168 169 2020), fog and mist are more common low visibility events in Beijing and Shanghai and thus have been excluded from the labels of severe hazes by following GSOD fog marks. The number 170 171 of severe haze events occurred during 1979-2019 defined in the above procedure is 2999 and





173 Longitude
174 Figure 2. (Left) The input-feature defining domains for Beijing (red box and dot, 99.25 - 123E, 32.25175 48N; 96x64 grids with ERA5 data) and Shanghai (white box and dot, 109.25-125E, 26-41.25N; 64x64
176 grids), made using Basemap library, a matplotlib extension. (Right) Annual means (solid curves), 25<sup>th</sup>
177 percentiles (dash curves), and 25<sup>th</sup> percentile means (solid straight lines) of surface visibility in Beijing
(red) and Shanghai (thus) between 1070 and 2010

179 The training and validation of HazeNet also need the input features with the same sample 180 dimension of the labels. These input data are derived from hourly longitude-latitude maps of meteorological and hydrological variables covering the data collection domain (Fig. 2, Left), 181 182 obtained from ERA5 reanalysis data produced by the European Centre for Medium-range Weather Forecasts or ECMWF (Hersbach *et al.*, 2020). These data are distributed in a grid 183 system with a horizontal spatial interval of 0.25 degree. Up to 16 features are derived from the 184 original hourly data fields covering the analysis domain respectively for Beijing (64x96 grids) 185 and Shanghai (64x64 grids), including: daily mean of surface relative humidity (REL thereafter); 186 diurnal change as well as daily standard deviation of 2-meter temperature or DT2M and T2MS, 187 respectively; daily mean of 10-meter zonal and meridional wind speed or U10 and V10, 188 189 respectively; daily mean of total column water (TCW); daily mean (TCV) and diurnal change (DTCV) of total column water vapor; daily mean of planetary boundary layer height (BLH); 190 daily mean soil water volume in soil layer 1 and 2 or SW1 and SW2, respectively; daily mean of 191 total cloud cover (TCC); daily mean geopotential heights at 500 (Z500) and 850 (Z850) hPa 192 pressure levels along with their diurnal changes D500 and D850, respectively. All input features 193 194 have been normalized into a range of [-1, +1] (Fig. S2 in Supplementary).

Before the training, the entire samples of labels alongside corresponding input features were randomly shuffled first then split as: 2/3 of the samples went to training set and 1/3 to validation set, each is used duly for its designated purpose throughout the entire training process without switch. The above procedure treats each of the events as an independent one. For the convenience in comparing performance or restarting training based on a saved machine, a pair of saved training <u>dataset and alongside a holdout</u> validation <u>dataset that has never been used in</u> training, <u>datasets</u> produced following the above procedure, was used.

202 The number of samples used in training HazeNet is rather limited in deep learning standard. However, to associate 16 joint two-dimensional maps with targeted labels even with the current 203 204 number of samples is still a demanding task, requiring a deep CCN to accomplish. Furthermore, targeted severe hazes are a low probability event. Its frequency of appearance is about 20.0% in 205 Beijing and Shanghai cases. Therefore, trained machine would easily bias toward the 206 overwhelming non-haze events. To resolve these issues, a combination of class-weight and batch 207 208 normalization has been implemented in HazeNet, both using corresponding Keras functions. The class weight is to change the weight of training loss of each class, normally by increasing the 209 210 weight of the low frequency class. Class weight coefficient was calculated based on the ratio of class 0 to class 1 frequency. Batch normalization (Ioffe and Szegedy, 2015) is an algorithm to 211 212 renormalize the input distribution at certain step (e.g., each mini batch) to eliminate the shift of 213 such distribution during optimization. This The above approach has effectively reduced the 214 overfitting while overcome the data imbalance issue, making the long training of a deep CNN 215 become possible (Wang, 2020). Entire trainings have been conducted using a NVIDIA Tesla 216 V100-SXM2 GPU cluster, costing 25s and 17s per epoch for the machine of Beijing and 217 Shanghai, respectively. 218 Kernel size and CNN performance optimization. As in the cases of other CNNs, there are 219 many hyperparameters need to be determined or optimized. These have been done through 220 numerous testing trainings. In practice, it occurs that, the The deep architecture of HazeNet and 221 the long training procedure have actually made the performance less sensitive to many 222 hyperparameters of the network. One hyperparameter, however, is specifically interesting to 223 explore for an application using large quantity of meteorological maps, that is the kernel size of

224 <u>the first convolutional layer, where the input data, *i.e.*, meteorological and hydrological maps are</u>

- 225 <u>convoluted then propagated into the subsequent layers. Meteorological maps or images often</u>
- 226 <u>contain characteristic patterns with different spatial scales. Intuitively, preserving these patterns</u>
- 227 <u>could be important in predicting the targeted extremes. Apparently, a larger kernel size produces</u>
- <u>smoother output images from the first convolutional layer, while a smaller kernel size can</u>
   preserve many spatial details of the meteorological maps as demonstrated from the layer output
- preserve many spatial details of the meteorological maps as demonstrated from the layer output
   shown in Fig. 53. In practice, however, the patterns produced by the latter configuration might be
- too complicated for the networks to recognize and to perform classification, whereas patterns
- resulted from a relatively larger kernel size for the first convolutional layer might be more
- 233 <u>characteristic for the task. The actual result suggests that HazeNet configured with a first-layer</u>
- kernel size of 20 to 26 or close to 5 6 <del>degree</del>degrees in spatial 'resolution', consistently
- produces a better performance (about a 10% improvement in *F1 score*; see next section and
- 236 <u>Method</u>) than that by a smaller kernel size of 3 or 6. As a result, a kernel size of 20 has been
- adopted as the default configuration for the first 2 convolutional layers in this study.



238

239 Figure 53. (Left column) Weight coefficients of the first filter set  $(W_{N,l})$ , (Middle column) partial output 240 for each feature  $(Z_{N,1})$ , and (Right column) the output (Z) of the first convolution layer (CONV2d 1) with 241 two selected kernel sizes or ks: (upper panels) 20x20 and (lower panels) 3x3. Here W represents the filters 242 and Z the output of convolution, the subsets of Z before the feature dimension is merged can be expressed <u>as:  $Z_{N,i} = W_{N,i}(ks, ks) \cdot f_N^T(ks, ks)$ , with the order of input features  $N = 1, \dots 16$  and *i* represents the</u> 243 244 convolutional layer index, *i.e.*, 1 is the first layer or CONV2d 1. For the first layer, input feature size is 245 (h,w) = (64, 64), the sets of filters is 92, thus the final output Z has a dimension of (h-ks+1, w-ks+1, 92). 246 Shown are results from the trainings for Shanghai haze cases.-

247

# 248 **3** Training and Validation Results of Haze Forecasting

Currently, it is still difficult to find any practical score in forecasting the occurrence of severehazes for comparison. Therefore, the performance of HazeNet has been mainly measured by

using certain commonly adopted metrics for classification largely derived from the concept of
the so-called confusion matrix (*e.g.*, Swets, 1988; Table A), including *accuracy*, *precision*, *recall*, *F1 score*, *equitable threat score* or *ETS*, and *Heidke skill score* or *HSS* (Appendix A).
Unless otherwise indicated, the discussions on the performance scores are hereafter referring to
the severe haze class, or class 1, and obtained from validation rather than training. In all the

- cases, the performance metrics referring to non-haze or class 0 has much better scores.
- 257 In order to train a stable machine, trainings with 2000 epochs or longer have been conducted 258 instead of using certain commonly used-adopted skills such as early stop. As a result, the validation performance metrics of the trained machines all appeared to be stabilized by 259 260 approaching the end of training (Fig.  $\frac{34}{2}$ ). These scores were consistent with the results of 261 ensemble training with the same configuration but different randomly selected training and 262 validation datasets, and also comparable among trainings with different configurations. Overfitting has been clearly overcome due to such a long training procedure alongside the 263 264 adoption of elass0weight class weight and batch normalization. In a 2-class classification (haze vs. non-haze), trained deep HazeNet can always reach an almost perfect training accuracy (e.g., 265 0.9956 for Beijing cases) and a validation accuracy of 80% (frequency of non-haze events or no-266 267 skill forecasting accuracy) in both Beijing and Shanghai cases, or the no-skill forecast accuracy 268 for no-haze (Fig. 34, left). At the same time, the performance scores in predicting specifically severe hazes are also very reasonable, e.g., for Beijing cases either precision or recall exceeds 0.5 269 270 (they normally evolve in opposite direction), leading to a nearly 0.5 F1 Score (Fig.34, right). The 271 corresponding scores in training are obviously much higher, e.g., with precision, recall, and F1 as 272 0.9804, 0.9980, and 0.9880, respectively for Beijing cases, owing to the deep and thus powerful 273 CNNs. HazeNet performed slightly better than several known deep CNNs such as Inception Net V3 (Szegedy et al., 2015), ResNet50 (He et al., 2015), and VGG-19 (Simonyan and Zisserman, 274 2015) in the same haze forecasting task (Wang, 2020). Nevertheless, as indicated previously that 275 276 a nearly perfect validation performance is not realistic since meteorological and hydrological 277 conditions are not the only factors behind the occurrence of haze events.
- 278



279 280 Figure 34. (Left) Validation accuracy (top panel) and loss (lower panel) of HazeNet with 16 features for 281 Beijing and Shanghai cases, kernel size for the first filter is 20x20. (Right Top) Prediction outcomes in 282 reference to haze events or class -1 of Beijing and Shanghai. Here TP = true positive, TN = true negative, 283 FP = false positive, and FN = false negative prediction outcomes. (Right Bottom) Scores of performance 284 metrics as last 100 epoch means for Beijing and Shanghai with 16 and 9 features, respectively.

Looking into the specific prediction outcomes in referring to severe haze, the trained machine 285 has produced considerably higher ratio of true positive or TP outcomes than in the Southeast 286 Asia cases (Wang, 2020) despite a number of outcomes of false positive or FP (*i.e.*, false alarm) 287 288 and false negative or FN (i.e., missing forecast). In forecasting the severe hazes in Beijing, the 289 trained machine performs reasonably well throughout all months except for April and May or the 290 major dusty season there, producing F1 score, ETS, and HSS all exceed or near 0.5 as well as the 291 number of TP outcomes is higher than that of FN (Fig. 45). The performance of HazeNet actually improves performs better in months with higher more observed haze events. For Beijing, 292 the lowest haze season is during the dusty April and May when all the major performance 293 metrics are lower than 0.4, and the machine produces more missing forecasts than true positive 294 295 outcomes. The relatively poor performance in spring suggests that the weather and hydrological features associated with dust-dominated haze events during this period might differ from the 296 297 situations in the other seasons when hazes are mainly caused by local particulate pollution. For 298 Shanghai cases, HazeNet performs better during late autumn and entire winter (from November 299 to February) when haze occurs most frequently (not shown). The worst performance comes from

the monsoon season (July to October), or the season with lowest haze cases. 300



301

302

Figure 45. (Top) Predicted Monthly counts of predicted TP, FP, and FN outcomes and (Bottom)
 performance scores for each month. All from validation of Beijing cases with 16 features.

305 Kernel size and CNN performance. The deep architecture of HazeNet and the long training 306 procedure have actually made the performance less sensitive to many hyperparameters of the 307 network. One hyperparameter, however, is specifically interesting to explore for an application 308 using large quantity of meteorological maps, that is the kernel size of the first convolutional layer, where the input data, i.e., meteorological and hydrological maps are convoluted then 309 310 propagated into the subsequent layers. Meteorological maps or images often contain 311 characteristic patterns with different spatial scales. Intuitively, preserving these patterns could be 312 important in predicting the targeted extremes. Apparently, a larger kernel size produces smoother 313 output images from the first convolutional layer, while a smaller kernel size can preserve many spatial details of the meteorological maps as demonstrated from the layer output shown in Fig. 5. 314 315 In practice, however, the patterns produced by the latter configuration might be too complicated 316 for the networks to recognize and to perform classification, whereas patterns resulted from a 317 relatively larger kernel size for the first convolutional layer might be more characteristic for the 318 task. The actual result suggests that HazeNet configured with a first layer kernel size of 20 to 26 319 or close to 5 - 6 degree in spatial 'resolution', consistently produces a better performance (about 320 a 10% improvement in F1 score) than that by a smaller kernel size of 3 or 6. As a result, a kernel 321 size of 20 has been adopted as the default configuration for the first 2 convolutional layers in this 322 study.

	CONV2d_1 W <sub>N.1</sub>	CONV2d_1 Z <sub>N.1</sub>	CONV2d_1 Z
	SS 🔛 💓 🚋 -	🚰 🌇 🚞	
K 1 00	N 💽 🔛 📰 📰	🦉 💻 🚍 📬	
Kernel = 20	🔫 👬 🎇 👬 🕺	5 🔄 🔍 ざ	
	🎇 🌨 🌉 🥌 e	<b>N N N</b>	
		i i i i i i i i i i i i i i i i i i i	
		27 🚝 🚝 🛒	
Kernel = 3		KA 🔛 🌌 💒	

323 324 Figure 5. (Left column) Weight coefficients of the first filter set (W<sub>N-1</sub>), (Middle column) partial output for each feature  $(Z_{N,1})$ , and (Right column) the output (Z) of the first convolution layer (CONV2d -1) with 325 two selected kernel sizes or ks: (upper panels) 20x20 and (lower panels) 3x3. Here W represents the filters 326 327 and Z the output of convolution, the subsets of Z before the feature dimension is merged can be expressed as:  $Z_{Ni} = W_{mi}(ks, ks) \cdot f_{m}^{\mp}(ks, ks)$ , with the order of input features  $N = 1, \dots, 16$  and *i* represents the 328 convolutional layer index, i.e., 1 is the first layer or CONV2d 1. For the first layer, input feature size is 329 330 (h,w) = (64, 64), the sets of filters is 92, thus the final output Z has a dimension of (h-ks+1, w-ks+1, 92). 331 Shown are results from the trainings for Shanghai haze cases.

**Reducing the number of input features**. One recognized advantage of deep CNN in practice is its capacity to directly link the targeted outcome with a large quantity of raw data. thus to avoid human misjudgment in selecting and abstracting input features due to a lack of knowledge about the application task. Nevertheless, for an application such as this one that uses a large number of meteorological and hydrological variables (or channels in machine learning term), reducing the number of input features with <u>a</u> minimized influence on the performance can still benefit the efforts of establishing physical or dynamical causal relations and beyond.

There are certain available methods to rank features then reduce some unimportant ones. These do not work straightforwardly for deep CNNs (*e.g.*, McGovern *et al.*, 2019). In the previous effort, this has been done by testing the sensitivity of the full network performance in real training with either a single feature <u>only</u> or all but one features (Wang, 2020), which apparently is also a demanding task. Here, another attempt has been made to use a trained then saved machine to examine the sensitivity of the network to various features (Appendix B).

B45 The sensitivity analyses for using trained machines for Beijing and Shanghai cases have obtained largely consistent results, indicating that the network is more sensitive to the same 9 features than the other 7 (Fig. S3). The highest-ranking features though differ, with diurnal change of column vapor (DTCV) and soil water content in the second soil layer (SW2) as the most sensitive features for Beijing, while relative humidity (REL) and planetary boundary layer height (BLH) for Shanghai. Most importantly, trainings using only the top 9 most sensitive features have produced a performance equivalent to or even better than the same training but with 16 features (Fig. <u>34</u>). With reduced number of features, many further analyses can be
conducted with less workload and produce results that are easily understood.

# 4 Identifying and Categorizing the Typical Regional Meteorological and Hydrological Regimes Associated with Haze and Non-Haze Events

A major purpose of this study is to identify the meteorological and hydrological conditions 356 357 favoring the occurrence of severe hazes in the targeted cities. When using a dataset with a large number of samples, this type of analyses could be better accomplished by applying, e.g., cluster 358 analysis (e.g., Steinhaus, 1957), a standard unsupervised ML algorithm that groups data samples 359 360 into various clusters in such a way that samples in the same cluster are more similar to each other than to those in other clusters. Specifically for this study, the derived clusters would likely 361 represent various regimes in terms of combined meteorological and hydrological conditions for 362 associated events. However, applying cluster analysis directly to a large number of samples, each 363 with a feature volume of ~50000 is an uneasy task. A dimensionality reduction is apparently 364 needed to reduce the feature volume of data. 365

In practice, a trained CNN is actually an excellent tool for this purpose. It encodes 366 (downscales) the input with large feature volume into data with a much smaller size in the so-367 368 called latent space (*i.e.*, the output of the layer before the output layer) while equal predictability for the targeted events. This feature functionality of CNN has been used in developing various 369 370 generative DL algorithms from variational autoencoder or VAE to different generative 371 adversarial networks or GANs (e.g., Forest, 2019). Therefore, the trained HazeNet for Beijing 372 and Shanghai have been used in this study to produce data with reduced size suitable for 373 clustering (Fig. 6; see also Appendix C). The new sample-feature set with a size of 14,975×512

- 374 produced from this procedure was then used in cluster analysis.
- 375



376

Figure 6. A diagram of the cluster analysis procedure. Here 96, 64, and 9 represent the number of

longitudinal, latitudinal grids, and number of features (variables), or the size of the input feature volume
of a trained HazeNet for Beijing cases, while 512 is the size of the output from the second dense layer
before output layer of HazeNet or the new feature volume.

In order to provide useful information for understanding the performance of the trained 381 382 networks, the clustering has been performed for each of the prediction outcomes rather than just haze versus non-haze events (Appendix C). In this configuration, haze associated regimes are 383 384 represented by derived clusters of TP plus FN outcomes, while non-haze regimes by those of TN 385 plus FP. Since the clusters were actually derived using the indices of samples as the record for members, the actual feature maps of the members in any cluster thus can be conveniently 386 387 retrieved then used to identify the representative regimes in terms of combined 9 meteorological 388 and hydrological features of various prediction outcomes or haze versus non-haze events. Here the clustering results have been analyzed using the feature maps in both normalized (machine 389 native) and unnormalized (original reanalysis data) format. The characteristics of various 390 391 regimes can be easily identified from the former as they represent anomalies to climatological means. An added benefit is to advance the understanding of the performance of the trained 392 networks. The analysis using the latter maps aims to better appreciate the conventional regional 393 and local meteorological and hydrological patterns associated with various regimes. The feature 394 395 maps used in both analyses have been averaged across each cluster for clarity.

## **4.1 Results based on normalized feature maps**

397 As shown in Figure 7, the 4 clusters of true positive or TP in Beijing cases exhibit a clear 398 similarity in general feature patterns closely surrounding Beijing (marked by a navy dot in the 399 figure) among themselves., differing only in rather minor details. The differences between 400 elusters are more evident in the daily change of column water vapor or DTCV and in two soil 401 water contents (SW1 and SW2). On the other hand, FN elusters (also associated with haze events but missed in prediction) also display a clear similarity to the patterns of TP clusters across most 402 features except DTCV, SW1, and SW2. These common patterns include an isolated small 403 404 positive relative humidity (REL) center covering Beijing, associated with mild diurnal variation 405 change (DT2M) and standard deviation (T2MS) of surface temperature as well as zonal wind (U10), and a lower boundary layer height (BLH). Weatherwise, Beijing and its immediate 406 407 surrounding area appear to be located between two sharply different airmasses occupying respectively the northwestern and southeastern part of the domain (weather systems usually 408 409 progress from northwest to southeast in this region). When relating this to the other feature 410 characteristics, it is likely that Beijing and nearby area is not experiencing a drastic weather system change such as fronts when haze occurs, hence the high REL- a critical condition for 411 aerosol to effectively scatter sunlight - can be easily formed, aided by a stable boundary layer 412 413 with mild surface wind to allow aerosols well mix vertically near the ground while without being 414 significantly reduced through advection diffusion. In addition, relatively high soil water content could fuel the humidity in the air, and thin while stable low clouds, if exists (judged based on 415 temperature change) could signal a lack of persistent precipitation. Altogether, these conditions 416 417 can apparently allow the haze to easily form, to last, and to effectively scatter sunlight thus reduce visibility. These conditions are also in a noticeably contrast to those associated with non-418 haze events represented by TN outcomes (Fig. S4). 419 420 Note that each cluster consists of a collection of 3D data volumes or images, any two clusters could be sufficiently differentiated should only one of their images differs based on the 421 422 clustering derivation algorithm, even though statistically speaking, they very likely belong to the 423 same population (*i.e.*, should be tested statistically). As shown in Fig. 7, the distinctions between 424 TP clusters are largely reflected from the two different airmasses distant from Beijing, in both 425 strength and spatial extent particularly from DTCV patterns, likely representing different types

426 of systems or background regimes. Specifically, a strong DTCV anomalous center seen in cluster 427 1 and 4 patterns occupies most of the domain west of Beijing and directly influence Beijing and its nearby area. In contrast, DTCV distributions in cluster 2 and 3 are much weaker, where 428 429 Beijing and its immediate neighboring area even appear to be influence more by the southeaster system. In addition, surface wind distributions of the first two clusters clearly differ from those 430 of cluster 3 and 4, and the patterns of BLH alongside SW1 and SW2 over Beijing and its 431 432 immediate neighboring area of cluster3 also suggests a land-atmosphere exchange condition 433 differing from that of others. The combinations of these differences across various TP clusters 434 apparently well defines the various regimes of surrounding weather systems as well as their 435 influence on Beijing. For TP clusters of Shanghai, the above similarities alongside differences 436 among various clusters also exist, except where the cluster 1, 2, and 4 maintain more similarities in feature patterns of distant airmasses from Shanghai, while cluster 3 offers certain evident 437 diversity in many feature patterns comparing to other clusters (Fig. S5). Even more interestingly, 438 the distribution of the number of members within various TP clusters does not differ evidently in 439 different months (Table S1) (note that the number of haze events itself differs seasonally – Fig. 440 5). Therefore, it is very likely that the characteristic weather conditions favoring haze occurrence 441 442 and being captured by HazeNet cannot be simply differentiated by locations (Beijing vs. 443 Shanghai) and seasons. 444 On the other hand, among three FN clusters (also associated with haze events but missed in 445 prediction), only the first cluster display a clear similarity to TP clusters across most features, 446 though the characters of the airmasses distantly surrounding Beijing differ substantially from those of TP clusters. Such differences appear to be even more evidently in the two other clusters 447 448 alongside some of the common features in TP clusters, e.g., the size and strength of high relative 449 humidity center covering Beijing are even different. This result suggests a possible reason for why HazeNet missed these targets, that is haze might occurr under unfavorable weather and 450 451 hydrological conditions owing to, e.g., certain energy consumption scenarios. Again, the 452 distribution of members of these latter two clusters does not exhibit clear seasonality (Table S1). 453 Generally speaking, the common patterns in normalized feature maps shared by most clusters associated with observed haze events (i.e., TP plus FN outcomes) include an isolated positive 454 455 relative humidity (REL) center in the southeast region covering Beijing associated with mild temperature variations (DT2M and T2MS) as well as zonal wind (U10) and lower boundary 456 457 layer height (BLH). Note that the mild daily temperature variation alongside lower BLH indicates that the haze region is not experiencing drastic weather system change such as fronts 458 and likely covered by low cloud, hence the high REL can be easily formed. All these characters 459 reflect a stable regional weather conditions over the southeastern half of the domain where 460 targeted hazes occurred. They are also in a sharp contrast to the conditions in the northwestern 461 half of the domain as well as the conditions associated with non-haze events represented by TN 462 outcomes (Fig. S4). On the other hand, FN clusters (also associated with haze events but missed 463 in prediction) also display a clear similarity to the patterns of TP clusters across most features 464 465 except DTCV, SW1, and SW2. Interestingly, first two of the 4four FP (false alarm) clusters actually display display a more 466 467 clear similarity in normalized feature patterns to those of TP as those of than FN in Beijing and 468 its immidiate surrounding area -(Fig. 7). As in FN cases, however, two other clusters differ more evidently. In addition, despite an anticipated diversity in feature patterns across TN clusters (Fig. 469 470 S4), four of its clusters (i.e., 2, 5, 12, and 13) exhibit a certain level of similarity to those of TP 471 clusters. All these could offer an explanation for explain the forecast false alarming errors made

472 by the machine, *i.e.*, the machine could have simply been confused by such similarities between

473 certain <del>FN and TN members, or between certain</del> TP and FP members. Nevertheless, these could

- also suggest an alternative reason behind the incorrect forecasts that is certain pollution
- 475 <u>mitigation measures were in place</u>. <u>The results of Shanghai FP clusters and the last FN cluster</u>
- 476 <u>besides TP are largely of Shanghai cases also share some similar characters as analyzed here the</u> 477 same as in Baijing case (Fig S5 & S6)
- 477 <u>same as in Beijing case (Fig S5 & S6).</u>
- <u>Therefore, it</u>It is worth indicating again that meteorological or hydrological conditions are
   not the only factors determining the occurrence of hazes. Other factors such as abnormal energy
   consumption events or long-range transport of aerosols could all cause haze to occur even under
   unfavorable weather and hydrological conditions. This could well be the reason for some of the
- 482 missing forecasts (FN outcomes) when haze occurred under unfavorable conditions, <u>as suggested</u>
- <u>above</u>, or for false alarms (FP outcomes) when low aerosol events occurred even under a weather
   condition favorable to haze. Future improvement of the skill could benefit from this knowledge.
- 485 The results of Shanghai are largely the same as in Beijing case (Fig S5 & S6).
- 486
- 487





489

Figure 7. Maps of 9 features in normalized format for 4 clusters of true positive or TP outcome, 5-3
clusters of false negative or FN outcome, and 4 clusters of false positive or FP outcome. Here TP plus FN
haze events. Results shown are cluster averages for Beijing (location marked by navy dot) cases.

# 493 **4.2 Results based on original unnormalized feature maps**

494 Utilizing feature maps in their original unnormalized format represented by actual physical
495 quantities could provide a convenience to appreciate the conventional regional and local
496 meteorological and hydrological patterns associated with various events. Note that the visual
497 differences between unnormalized feature maps particularly in cluster-mean format might be
498 subtle for bare eyes to recognize.

For haze events in Beijing (*i.e.*, TP and FN outcomes; Fig. 8), the associated cluster-mean regional meteorological and hydrological patterns of most features except DTCV contain two regions with sharply contrasting quantities, roughly separated by a line linking the southwest and northeast corner of the domain, likely due to the <u>nature typical progression direction</u> of weather systems in this region besides meridional variation of general climate. Beijing (at ~1/3 domain width from the east boundary and nearly the north south center) locates in the southeastern half of the domain. In comparison, as same as shown in the previous analysis using normalized

17

- feature maps, the patterns of <u>the first</u> FN clusters share many <u>common patterns characters</u> with
- 507 those of TP clusters. Their differences among TP and FN clusters are more evident in DTCV
- 508 (specifically cluster 1 and 4 versus cluster 2 and 3), SW1, and SW2, and surface winds
- 509 particularly for the 2<sup>nd</sup> and 3<sup>rd</sup> FN clusters. In addition, cluster 5 of FN shows more diverse
- 510 patterns than the rest. FP clusters also display a similarity to those of TP clusters (Fig. <u>\$5\$7</u>),

whereas TN clusters show more visible differences particularly in patterns of meridional wind
(V10) and daily change of column water vapor or DTCV (Fig. <u>\$658</u>).

512 REL DT2M T2MS U10 V10 DTCV BLH SW1 SW2 (=1 <=2 TΡ <=4 C REL DT2M T2MS U10 V10 DTCV BLH SW1 SW2 =1 FN E I 513



514

Figure 8. Feature maps associated with severe haze events in Beijing represented by 4 clusters of TP (4 top rows) and 5-3 clusters of FN (5-3 lower rows) predicted outcomes. Shown are cluster means of unnormalized data of relative humidity or REL (ratio), diurnal change (DT2M) and daily standard deviation (T2MS) of 2-meter temperature in degree, 10-meter winds U10 and V10 in m/s, diurnal change of column water vapor or DTCV (kg/m<sup>2</sup>), planetary boundary hheight of BLH in meter, and soil water content in soil level 1 (SW1) and level 2 (SW2) in kg/m<sup>2</sup>.

521 The general regional meteorological and hydrological conditions during haze events in the southeastern in comparison to the northwestern portion of the domain include a higher relative 522 523 humidity, lower variation of surface temperature, largely northward or northwestward wind, 524 lower planetary boundary layer height, and higher soil water content, and quantity wise these are 525 all in a sharp contrast to the situations in the other half of the domain. Based on the surface wind 526 direction. Beijing and its immediate surrounding area is clearly located between two airmasses 527 both with anticyclonic surface winds. The strengths of these two centers differ particularly in the 528 last two FN clusters, implying regimes with systems having different strengths or in different 529 development phases. The Such a difference is also clearly related to the visually recognized 530 cross-cluster differences of haze events mainly exist in DTCV patterns, represented by a strong 531 negative center in the middle of the domain with varying extent and strength across different clusters. To a less extent, patterns of surface wind V10 and U10 also offer some different 532 533 characteristics among various clusters particularly of FN clusters. Consistent to the analysis 534 result using normalized feature maps, all these indicate a stable weather condition over the 535 southeastern half of the domainBeijing and its neighboring area for during haze events while surrounded by two (or more) different weather systems in Beijing. It is known that dust can 536 cause low visibility events in Beijing. During dust seasons, the condition of the northwestern half 537 of the domain, represented by a dominant eastward wind and lower soil water content likely 538 favors dust transport from desert to Beijing. However, the details would need an in-depth 539

analysis to examine since most clusters having members rather well distributed through differentmonths (Table S1).

The cluster-means of 9 features for haze events (TP plus FN) versus non-haze (TN plus FP) 542 543 at the grid point of Beijing are also derived and listed in Table 1 for reference. Specifically, the common local conditions associated with hazes in Beijing in comparison to those with non-haze 544 545 events include a higher humidity, less drastic variations in surface temperature, a northwestward rather than southeastward wind, a lower planetary boundary layer height, and higher soil water 546 547 contents. Again, the most recognizable cross-cluster differences appear in DTCV (*i.e.*, cluster 1 548 versus others), followed by surface wind (cluster 1 and 2 versus 3 and 4). In most of the local 549 features, variabilities of FN clusters tend to be larger than those of TP clusters. Notably, such 550 differences in local feature quantities for FN clusters are not necessarily more evident than in the regional maps over distant airmasses. One interesting result of the local weather conditions 551 shown in Table 1 is that the cluster means of TN are sharply different than those of TP and FN, 552 553 while the cluster means of FP and those of TP+FN are likely to be statistically indifferent except 554 for DTCV, providing an evidence to support the assumption that FP outcomes might simply 555 represent the non-haze events caused by reasons other than weather and hydrological conditions. 556





558

559 560

**Figure 9.** The same as Figure 9 except for Shanghai with 3-4 clusters for TP and 4-3 for FN outcomes.

561 For the case of Shanghai, the general weather conditions associated with haze events are 562 likely stable, with characters similar to the cases of Beijing except for that Shanghai appears to be located between a northwest airmass with anticyclonic surface wind and a southeast one with 563 cyclonic wind (Fig. 9). Quantities of most feature patterns display a sharply southeast versus 564 northwest contrast. DTCV maps display a negative center over a large area, its distribution and 565 extent vary significantly among different clusters in particular for the first two FN clusters. The 566 567 patterns of soil water content in both soil layers exhibit a sharp meridional contrast, much higher 568 in the south part of the domain than in the north part, largely separated by the Yellow River. Local quantities of all the features associated with haze events (TP plus FN) in Shanghai display 569 clear differences with those of non-haze prediction outcomes (TN) (Table 1). The most 570 571 recognizable cross-cluster differences for TP appear in U10 of cluster 4 and V10 of cluster 3, differing from the cases of Beijing, and DTCV particularly of cluster 3 for FN. Similar toLike 572 573 the cases of Beijing, the cluster mean of the FP outcomes is statistically indifferent to those of 574 haze (TP and FN) than predicted non-haze (TN) events. Again, this result implies that even a 575 weather pattern favoring haze appeared and was correctly recognized by HazeNet, due to other

576 factors such as energy consumption variations, haze could still not to occur.

577 It is worth indicating that the current analysis discussed here is only applied to the included

578 <u>features in clustering, and the presented figures in cluster-wise averaging format might have</u>

579 <u>effectively smoothed out certain variability among members. A full-scale analysis would</u>

580 <u>necessarily go beyond this to provide further synoptical or large-scale hydrological insights and</u>
 581 <u>better define different regimes.</u>

**Table 1**. Cluster means of features associated with haze events (TP and FN) in Beijing and Shanghai

versus means of all clusters of non-haze events of TN and FP, respectively. Number of cluster membersof each cluster are listed in bracket.

Cluster	REL (0-1)	DT2 (°C)	T2MS (°C)	U10 (m/s)	V10 (m/s)	DTCV (kg/m <sup>2</sup> )	BLH (m)	SW1 (kg/m <sup>2</sup> )	SW2 (kg/m <sup>2</sup> )
Beijing									
TP1 (848)	0.64	-5.99	3.24	-0.29	0.20	0.04	379.71	0.23	0.22
TP2 (181)	0.65	-5.80	3.14	-0.28	0.19	0.57	378.33	0.23	0.23
TP3 (354)	0.65	-5.39	2.98	-0.45	0.29	0.31	400.20	0.23	0.22
TP4 (1208)	0.64	-5.82	3.18	-0.34	0.28	0.27	381.28	0.23	0.22
FN1	<u>0.63</u> 0	<u>-6.24</u> -	<u>3.32</u> <del>3.1</del>	<u>-0.25</u> -	<u>0.20</u> 0.3	<u>0.07</u> <del>0.1</del>	<u>422.60</u> 379.9	<u>0.23</u> <del>0.2</del>	<u>0.22</u> 0.2
( <del>157<u>392</u>)</del> EN2	. <del>66</del> 0.650	<del>5.83</del> -5.71-	<del>6</del> 3.052.0	<del>0.43</del> -0.20-	4	<del>5</del> 0.10-	+ 106 65422 2	<del>3</del> 0 23 <u>0 2</u>	$\frac{1}{0.220-2}$
$(\frac{1390}{1})$	<u>0.05</u> 0 . <del>65</del>	<u>-5.71</u> - <u>5.05</u>	<u>3.03</u> 2.9	$\frac{-0.20}{0.52}$	<u>0.17</u> 0.4 8	$\frac{0.19}{1.88}$	<u>+00.05</u> <del>5</del>	<u>0.23</u> 0.2	<u>0.22</u> 0.2 2
FN3						Ξ			
( <del>29</del> <u>26</u> )	<u>0.69</u> 0	<u>-5.37</u> - 5.90	<u>2.94</u> 3.0	$\frac{-0.61}{0.41}$	<u>0.39</u> 0.3	$\frac{0.170.9}{9}$	<u>410.95</u> 393.5	$\frac{0.250.2}{4}$	$\frac{0.230.2}{3}$
<del>FN4 (86)</del>	0.64	-5.64	3.02	- <u>0.19</u>	0.11	0.10	420.49	0.23	0.22
<del>FN5 (223)</del>	<del>0.60</del>	<del>-6.56</del>	<del>3.45</del>	<del>-0.14</del>	<del>0.11</del>	<del>0.01</del>	44 <del>9.48</del>	0.23	<del>0.22</del>
TN mean	0.51	-7.13	3.65	0.15	-0.15	0.36	552.90	0.22	0.21
FP mean	0.65	-5.84	3.15	-0.35	0.25	-0.26	386.27	0.24	0.23
Shanghai									
TP1 (1228)	0.81	-3.44	1.79	-0.16	-0.55	-2.25	415.59	0.35	0.35
TP2 (135)	0.81	-3.10	1.71	-0.12	-0.66	-2.08	422.04	0.36	0.36
TP3 (689)	0.81	-2.95	1.59	-0.17	-1.28	-2.29	472.74	0.36	0.35
TP4 (355)	0.81	-3.52	1.82	0.03	-0.57	-2.74	411.96	0.35	0.35
FN1	<u>0.80</u> 0	<u>-3.48</u> -	<u>1.80</u> 1.8	<u>-0.41</u> -	<u>-0.42</u> -	<u>-0.84</u> -	<u>421.13</u> 4 <del>09.5</del>	<u>0.35</u> 0.3	<u>0.35</u> 0.3
$(\frac{102372}{112})$	<del>.82</del>	3.33	θ	<del>0.67</del>	<del>0.36</del>	0.14	5	<del>5</del>	5
FN2 (113)	<u>0.80</u> 0	<u>-3.64</u> - <del>3.64</del>	$\frac{1.841.8}{4}$	$\frac{-0.34}{0.34}$	$\frac{-0.51}{0.51}$	$\frac{-1.21}{1.21}$	<u>423.09</u> 423.0 9	<u>0.35</u> 0.3	$\frac{0.340.3}{4}$
FN3	<u>0.82</u> 0	<u>-3.28</u> -	<u>1.77</u> 1.8	<u>-0.68</u> -	<u>-0.49</u> -	<u>0.10</u> -	<u>422.36</u> 421.3	<u>0.35</u> 0.3	<u>0.35</u> 0.3
( <u>370<u>107</u>)</u>	<del>.80</del>	<del>3.47</del>	0	<del>0.41</del>	<del>0.42</del>	<del>0.84</del>	6	5	5
<del>FN4 (7)</del>	<del>0.80</del>	-2.82	<del>1.39</del>	<del>-1.19</del>	-2.18	3.63	<del>596.53</del>	<del>0.36</del>	<del>0.36</del>
TN mean	0.77	-3.29	1.57	-2.86	1.40	0.62	739.75	0.31	0.32
FP mean	0.82	-3.26	1.71	-0.48	-0.85	-2.26	438.55	0.35	0.35

## 585 **5** Summary and Conclusions

Following an earlier preliminary attempt for hazes in Singapore, a deep convolutional neural
network containing more than 20 million parameters, namely HazeNet, has been further

developed to test forecasting the occurrence of severe haze events during 1979-2019 in two
metropolitans of Asia, Beijing and Shanghai. By training the machine to recognize regional
patterns of meteorological and hydrological features associated with haze events, the study
would advance our knowledge about this still poorly known environmental extreme. The deep
CNN has been trained in a supervised learning procedure using the time sequential maps of up to
16 meteorological and hydrological variables or features as inputs and surface visibility
observations as the labels.

Even with a rather limited samples (14,975), the trained machine has displayed a reasonable performance measured by commonly adopted validation metrics. Its performance is clearly better during months with high haze frequency, *i.e.*, all months except dusty April and May in Beijing and from late autumn through entire winter in Shanghai. Relatively larger spatial patterns appear to be more effective than the smaller ones to influence the performance of forecasting. On the other hand, in-depth analysis on performance results has also indicated certain limitations of current approach of solely using meteorological and hydrological data in performing forecast.

The trained machine has also been used to examine the sensitivity of the CNN to various input features and thus to identify then remove features ineffective to the performance of the machine. In addition, in order toto further categorize typical regional weather and hydrological patterns associated with severe haze versus non-haze events, an unsupervised cluster analysis has been subsequently conducted, benefited from using features with greatly reduced dimensionality produced by the trained machine.

The cluster analysis has, arguably for the first time, successfully categorized major regional 608 meteorological and hydrological patterns associated with severe haze and non-haze events in 609 610 Beijing and Shanghai into a limited number of representative groups, with the typical feature 611 patterns of these clustered groups derived. It has been found that the typical weather and hydrological regimes of haze events in Beijing and Shanghai are rather stable conditions, 612 613 represented by anomalously increasing high relative humidity, low planetary boundary layer height, mild daily temperature change that likely associated with a thin low cloud cover over the 614 615 haze occurring regions. The result has further revealed a rather strong similarity between the meteorological and hydrological patterns associated with haze events and those with either false 616 617 alarm or missing forecast prediction outcomes, implying that factors other than meteorological and hydrological ones such as energy consumption variations, long range transport of aerosols, 618 619 or beyond, could cause haze events to occur even under unfavorite weather conditions.

620 Due to the exploratory nature of this specific effort, several aspects could be further 621 optimized including the rather arbitrary though statistically meaningful labeling. Also, an indepth analysis on weather regimes would necessarily involve the use of certain features that are 622 623 not included in the current clustering, which, however, exceeds the extent of this paper and can 624 only be discussed properly in a future work. Nevertheless, this study has demonstrated the potential of applying deep CNNs with extensive multi-dimensional and time sequential 625 626 environmental images to advance our understandings about poorly known environmental and 627 weather extremes. The methodology, results alongside experience obtained from this study could benefit future improvement of the skills. Besides, the trained machines can be used in many 628 629 other types of machine learning and deep learning applications as partially demonstrated here.

## 630 Appendix A. Performance metrics

631 Several commonly used performance metrics have been used in this study. They are largely derived based on 632 the so-called confusion matrix (e.g., Swets, 1988) as defined in the following Table A.

633

634 **Table A.** Confusion matrix for measuring the prediction outcomes of a given class.

	Observed			
		Positive	Negative	
Predicted	Positive	True Positive or TP	False Positive or FP	
	Negative	False Negative or FN	True Negative or TN	

635 Here, positive or negative is referring to the outcome of a given event or class in the classification, e.g., severe haze 636 or non-haze events. Hence, the prediction outcome TP is a correct forecast of a severe haze while TN a correct 637 forecast of a non-haze event, FP represents a false alarm, and FN a missing forecast. The context of outcomes 638 changes when the designated class is switched. The major performance metrics used in this paper include:

639	$accuracy = \frac{TP+TN}{N}$	(A1)
640	$precision = \frac{TP}{TP + FP}$	(A2)
641	$recall = \frac{TP}{TP+FN}$	(A3)
642	$F1 \ score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$	(A4)
643	$ETS = \frac{TP - Hit_{random}}{TP + FP + FN - Hit_{random}};$	(A5a)

644 where: 
$$Hit_{random} = \frac{(TP+FN)\cdot(TP+FP)}{N}$$
 (A5b)  
645  $HSS = \frac{2\cdot(TP\cdot TN-FP\cdot FN)}{(TP+TP)\cdot(TP+FP)}$  (A6)

$$HSS = \frac{2 \cdot (TP \cdot TN - FP \cdot FN)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (TP + TN)}$$

646 Note that *accuracy* has the same value for all the classes and thus is a good metrics for the overall classification. 647 Values of all the other metrics differ depending on the referred specific class. Here, F1 score is the F-score with  $\beta =$ 648 1 (van Rijsbergen, 1974), ETS represents equitable threat score (or Gilbert skill score; Gilbert, 1884; range = [-1/3, -1/3]

649 1), HSS represents Heidke skill score (Heidke, 1926; range =  $[-\infty, 1]$ ), and N is the number of total outcomes.

#### 650 Appendix B. Examining the network's sensitivity to features using trained machine

651 A method has been adopted in this study to use a trained machine from basic training to examine the sensitivity 652 of the network to a random perturbation applied to the values of different features. The saved machine contains all 653 the coefficients in different network layers and can be used to predict output from any of these layers using same 654 input features for training or validation. The sensitivity of the network to a given feature is determined by comparing 655 the prediction using input feature maps containing randomly perturbation applied to the map of this feature with the prediction using original input feature maps, and measured by the content loss between these two predictions, with 656 657 *img1* with MxN pixels as the unperturbed and *img2* as perturbed network output:

658 
$$Content \ Loss = \frac{1}{M \times N} \sum_{i,j}^{M,N} (img \mathbf{1}_{i,j} - img \mathbf{2}_{i,j})^2 \tag{B1}$$

659 The perturbation is applied as random patch with addition of -0.2 or 0.2 to 10% of the pixels of the input map of 660 the targeted feature in each sample while maps of all the other features remain unperturbed. To reduce the workload, 661 only validation input set corresponding to the class-1 events (about 1020 samples) are used. Therefore, the 662 sensitivity tested here is actually the sensitivity of the network to a given feature in predicting class-1 events. To preserve the spatial information of the perturbation field, the output of the 9<sup>th</sup> layer, or the MaxPooling layer 663 664 following the second convolutional layer (Fig. 1) is used as the prediction. It has a size of (15, 31, 92) for Beijing cases and (15, 15, 92) for Shanghai cases when a kernel size of 20x20 is adopted. A higher content loss represents 665 666 that the performance of the network is more sensitive to the variations in value of this feature.

#### 667 **Appendix C. Cluster analysis**

668 The cluster analysis of this study was conducted in the following three steps (see also Fig. 6).

(i) Firstly, the trained and saved HazeNet for both Beijing and Shanghai cases with 9 input features have been
used to perform prediction using the entire 14,975 input samples in original raw data format, *i.e.*, with a feature
volume size of 96x64x9 for Beijing and 64x64x9 for Shanghai for each sample. The prediction results were then
summarized into various outcomes, *e.g.*, as true positive (TP), true negative (TN), false positive (FP), or false
negative (FN) in referring to the haze class. In the meantime, the output of the second dense layer just before the
output layer or latent space (see Fig. 1 & Fig. 6) were further used to form the new data of each sample with reduced
feature volume of 512. This new dataset with 14075 samples and 512 feature volume were ready for clustering.

676 (ii) The second step is to actually perform clustering using the new data with reduced size resulted from the 677 previous step. For this purpose, it should be conducted separately for different types of samples or events, e.g., 678 categorizing all the samples for haze into characteristic groups with similarity and same for non-haze events. In 679 order to provide additional information to further the understanding of the network's performance, the clustering 680 was actually conducted for different prediction outcomes, by taking corresponding samples from the new dataset. In 681 this case, TP plus FN would lead to haze events, and TN plus FP to non-haze events. The clustering calculations 682 were done by directly using the k-mean (Steinhaus, 1957) function of scikit-learn library (https://scikit-683 learn.org/stable/modules/clustering.html#clustering). For Beijing cases, the trained machine with 9 features 684 produced 2591 TP, 11368 TN, 508 FP, and 508 FN outcomes, and 2407 TP, 11484 TN, 492 FP, and 592 FN for 685 Shanghai. The cluster analysis was performed separately for each of these outcomes in an unsupervised learning 686 procedure to let the machine to categorize corresponding samples into groups based on similarities among them. In 687 practice, similarity is judged by the so-called inertia for a cluster with members of  $x_i$  and mean of  $\mu$ :

 $inertia = \sum_{i}^{N} (||x_{i} - \mu||)^{2}$ (C1)

689 The clustering is to seek a grouping with minimized inertia within each cluster. The overall measure is the 690 summation inertia that decreases almost exponentially with the increase of number of clusters. In practice, the 691 cluster analysis was first tested with various given number of clusters ranging from 1 to 100, to examine the values 692 alongside decay of the inertia. This provided a base to identify the smallest possible number of cluster centers with 693 reasonably low inertia in actual cluster analysis. This has actually been decided by using square root of the inertia 694 weighted by the number of samples to put the varying number of samples across various outcomes in consideration. 695 An optimized number of clusters was chosen with a weighted inertia lower than 1/e of that of the single cluster case. 696 For TN, due to the large sample number, this criterion was set to be half of 1/e. As a result, the optimized numbers 697 of clusters for TP, FN, FP, and TN outcomes are 4, 53, 4, and 15 for Beijing and 4, 4,3 3, and 10 for Shanghai, 698 respectively,

(iii) The members of each cluster derived from (ii) were recorded by the actual sample indices with date
 attribute. Therefore, actual samples of input data grouped into various clusters can be thus conveniently identified
 with corresponding feature maps retrieved, either in the format of normalized or unnormalized (*i.e.*, in original
 quantity as in reanalysis dataset), and used for further analyses. In practice, cluster-averaged maps for various
 features were performed beforehand.

#### 704 Code and data availability

The Python script for network architecture, training and validation is rather straightforward and simple,
 basically consisting of directly adopted function calls from Keras interface library (<u>https://github.com/keras-</u>
 team/keras) with TensorFlow-GPU (https://www.tensorflow.org) as backend, or from scikit learn library

708 (https://scikit-learn.org/). All the data used for analyses are publicly available as indicated in the

709 Acknowledgements.

## 710 Competing interests

711 The author declares that he has no conflict of interest.

712

688

#### 713 Acknowledgements

- 714 This study is supported by L'Agence National de la Recherche (ANR) of France under "Programme
- 715 d'Investissements d'Avenir" (ANR-18-MPGA-003 EUROACE). The author thanks the European Centre for
- 716 Medium-range Weather Forecasts for making the ERA5 data publicly available under a license generated and a
- 717 service offered by the Copernicus Climate Change Service, and the National Center for Environmental Information
- of the US NOAA for making GSOD data available. All the related computations have been accomplished using the
- 719 GPU clusters of French Grand equipment national de calcul intensif (GENCI) (Project 101056) and the CNRS
- 720 Mesocenter of Computing of CALMIP (Project p18025). <u>Constructive comments and suggestions from two</u>
- 721 <u>anonymous reviewers have led to the improvement of the manuscript.</u>

## 722 References

- 723 Chan, C. K. and Yao, X.: Air pollution in mega cities in China, Atmos. Environ., 42, 1-42, 2008.
- Chattopadhyay, A., Nabizadeh, E. and Hassanzadeh, P.: Analog forecasting of extreme-causing
  weather patterns using deep learning. *J. Adv. Modeling Earth Sys.*, 12, e2019MS001958.
  Doi:/10.1029/2019MS001958, 2020.
- Forest, D.: *Generative Deep Learning*, O'Reilly Media, Inc., 2019.
- Gagne, D., Haupt, S. and Nychka, D.: Interpretable deep learning for spatial analysis of severe
  hailstorms. *Mon. Wea. Rev.*, 147, 2827–2845, Doi:/10.1175/MWR-D-18-0316.1, 2019.
- 730 Gilbert, G. K.: Finley's tornado predictions, *Amer. Meteor. J.*, 1, 166–172, 1884.
- 731 Goodfellow, I., Bengio, Y. and Courville, A.: *Deep Learning*, MIT Press, 800pp., 2017.
- Grover, A. Kapoor, A. and Horvitz, E.: A deep hybrid model for weather forecasting, *Proc. 21st ACM SIGKDD Intern'l Conf. KDD*, p.379-386, Sydney, Australia, August 10, 2015. ACM.
   ISBN 978-1-4503-3664-2/15/08. Doi:10.1145/2783258.2783275, 2016.
- He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, arXiv:1512.03385, 2015.
- Heidke, P.: Calculation of the success and goodness of strong wind forecasts in the storm
  warning service, *Geogr. Ann. Stockholm*, 8, 301–349, 1926.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J.,
  Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X.,
- Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P.,
  Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A.,
- Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P.,
- Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and
- 745 Thépaut, J.-N.: The ERA5 global reanalysis, O. J. R. Meteorol. Soc., 146, 1999-2049, 2020.
- Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing
   internal covariate shift, arXiv:1502.03167, 2015.
- Jiang, G.-Q., Xu, J. and Wei, J.: A deep learning algorithm of neural network for the
  parameterization o typhoon-ocean feedback in typhoon forecast models, *Geophys. Res. Lett.*,
  45, https://doi.org/10.1002/2018GL077004, 2018.
- Kiehl, J. T. and Briegleb, B. P.: The relative roles of sulfate aerosols and greenhouse gases in
  climate forcing, *Science*, 260, 311-314, 1993.
- Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., Mahesh, A.,
  Matheson, M., Deslippe, J., Fatica, M., Prabhat and Houston, M.: Exascale deep learning for
  climate analytics, arXiv:1810.01993, 2018.
- 756 Lagerquist, R., McGovern, A. and Gagne II, D.: Deep learning for spatially explicit prediction of
- synoptic-scale fronts. *Wea. Forecasting*, 34, 1137–1160. Doi:10.1175/WAF-D-18-0183.1,
  2019.

- LeCun, Y., Bengio, Y. and Hinton, G.: Depp learning, *Nature*, 521, 436-444,
- 760 doi:10.1038/nature14539, 2015.
- Lee, H.-H., Iraqui, O. and Wang, C.: The impacts of future fuel consumption on regional air
   quality in Southeast Asia, *Sci. Rep.*, 9:2648, doi:10.1038/s41598-019-39131-3, 2019.
- Lee, H.-H., Iraqui, O., Gu, Y., Yim, H.-L. S., Chulakadabba. A., Tonks, A. Y. M., Yang, Z. and
  Wang, C.: Impacts of air pollutants from fire and non-fire emissions on the regional air
  quality in Southeast Asia, *Atmos. Chem. Phys.*, 18, 6141–6156, doi:10.5194/acp-18-61412018, 2018.
- Lee, H.-H., Bar-Or, R. and Wang, C.: Biomass Burning Aerosols and the Low Visibility Events
  in Southeast Asia, *Atmos. Chem. Phys.*, 17, 965-980, doi:10.5194/acp-17-965-2017, 2017.
- Lin, Y., Wijedasa, L. S. and Chisholm, R. A.: Singapore's willingness to pay for mitigation of
  transboundary forest-fire haze from Indonesia, *Environ. Res. Lett.*, 12, 024017,
  doi:10.1088/1748-9326/aa5cf6, 2016.
- Liu, M., Huang, Y., Ma, Z., Jin, Z., Liu, X., Wang, H., Liu, Y., Wang, J., Jantunen, M., Bi, J. and
  Kinney, P. L.: Spatial and temporal trends in the mortality burden of air pollution in China:
  2004-2012, Environ. Int., 98, 75-81, 2017.
- Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M.
  and Collins, W.: Application of deep convolutional neural networks for detecting extreme
  weather in climate datasets. arXiv:1605.01156, 2016.
- McGovern, A., Lagerquist, R., Gagne II, D. J., Jergensen, G. E., ElmLMore, K. L., Homeyer, C.
  R. and Smith, T.: Making the black box more transparent: Understanding the physical
  implications of machine learning, *Bull. Amer. Meteor. Soc.*, 100, 2175-2199, 2019.
- Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional networks for biomedical image
   segmentation, arXiv:1505.04597, 2015.
- Shi, X., Chen, Z., Wang, H. and Yeung, D.-Y.: Convolutional LSTM network: A machine
   learning approach for precipitation nowcasting, arXiv:1506.04214, 2015.
- Silva, R. A., West, J. J., Zhang, Y., Anenberg, S. C., Lamarque, J.-F., Shindell, D. T., Collins,
  W. J., Dalsoren, S., Faluvegi, G., Folberth, G., Horowitz, L. W., Nagashima, T., Naik, V.,
  Rumbold, S., Skeie, R., Sudo, K., Takemura, T., Bergmann, D., Cameron-Smith, P., Cionni,
- 788I., Doherty, R. M., Eyring, V., Josse, B., MacKenzie, I. A., Plummer, D., Righi, M.,
- Stevenson, D. S., Strode, S. Szopa, S. and Zeng, G.: Global premature mortality due to
  anthropogenic outdoor air pollution and the contribution of past climate change, *Environ*.
- 791 *Res. Lett.*, 8, 034005, doi:10.1088/1748-9326/8/3/034005, 2013.
- Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image
   recognition, arXiv:1409.1556, 2015.
- Smith, A., Lott, N. and Vose, R.: The integrated surface database: Recent developments and
  partnerships, *Bull. Ameri. Meteorol. Soc.*, 92, 704-708, doi:10.1175/2011BAMS3015.1,
  2011.
- 797 Steinhaus, H.: Sur la division des corps matériels en parties, *Bull. Acad. Polon. Sci.*, 4, 801–804,
  798 1957.
- 799 Swets, J.: Measuring the accuracy of diagnostic systems, *Science*, 240, 1285–1293, 1988.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the inception
   architecture for computer vision, arXiv:1512.00567, 2015.
- van Rijsbergen, C., 1974: Foundation of evaluation, *J. Documentation*, 30, 365–373.
- Wang, C.: Exploiting deep learning in forecasting the occurrence of severe haze in Southeast
   Asia, arXiv:2003.05763, 2020.

Weyn, J. A., Durran, D. R. and Caruana, R.: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere, *J. Adv. Modeling Earth Sys.*, e2020MS002109, <u>https://doi.org/10.1029/2020MS002109</u>, 2020.

808 809