Environmental Modelling and Software 154 (2022) 105424

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

Importance of 3D convolution and physics on a deep learning coastal fog model

Hamid Kamangir[a,d,*], Evan Krell[b,d,1], Waylon Collins[c], Scott A. King[b,d], Philippe Tissot[a,d]

[a] Conrad Blucher Institute for Surveying and Science, Texas A&M University - Corpus Christi, Corpus Christi, TX, USA
[b] Department of Computing Sciences, Texas A&M University - Corpus Christi, Corpus Christi, TX, USA
[c] National Weather Service, Corpus Christi, TX, USA
[d] NSF AI Institute for Research on Trustworthy AI in Weather, Climate and Coastal Oceanography, USA

ARTICLE INFO

Keywords:
3D convolutional neural network
SHAP
Explainable AI
Permutation feature importance
Atmospheric prediction

ABSTRACT

The forecasting of hazardous atmospheric phenomena is often challenging. Artificial intelligence (AI) models have been applied to atmospheric science problems. Model complexity provides a motivation to quantify the importance of model architecture components. We studied the relative importance of the components of the FogNet model that was designed for big atmospheric data: 1) 3D versus 2D convolution, 2) physics-based grouping and ordering of meteorological input features, 3) different auxiliary CNN-based feature learning modules and 4) parallel versus sequential spatial-variable-wise feature learning. We investigate the relative importance of these CNN architectural features by predicting coastal fog,a complex spatiotemporal dynamical process. We use four explainable AI techniques to better understand input feature contributions. The results of the experiments demonstrate that 3D-CNN based models better capture the complexity of the fog prediction process than the 2D-CNNs. We also show that physics-based feature grouping, and the order in which they are fed into the CNNs, significantly impacts performance.

Software and data availability

Name of software: FogNet v1.0.
Developer: Hamid Kamangir, Evan Krell.
Source: https://github.com/conrad-blucher-institute/FogNet.
Programming Language: Python 3.
Dependecies: Tensorflow, Keras, Numpy.
Licence: MIT License.
Data availability: North American Mesoscale (NAM) 12 km available in grib2 format archived at ftp://ftp.ncep.noaa.gov/pub/data/nccf/com/nam/prod/nam.YYYYMMDD.

1. Introduction

Convolutional neural networks (CNNs) have been applied extensively to atmospheric science applications in recent years. These models are often based on very large-scale spatio-temporal datasets, and the CNN may be required to learn complex, non-linear relationships to achieve acceptable performance. These datasets are often highly imbalanced — and predicting infrequent yet impactful events is complicated by the relatively few observations of the weather hazard as compared to the non-event occurrences. For example, the number of non-fog cases is much larger than the number of fog cases. A model can achieve high accuracy, but with no predictive skill, by always predicting the non-event Kumler-Bonfanti et al. (2020). If 95% of the test instances are the non-event, then 95% accuracy is achieved by simply always predicting no occurrence of, for example, fog. Complex CNN architectures have the potential to bring significant improvements for these problems. Given the large number of network parameters, such as the depth and width of the hidden layers, the choice of convolutional kernels, etc., it is challenging to develop high performance architectures.

HazeNet is an example of a CNN for an atmospheric application, forecasting severe haze, that achieves validation accuracy $> 95.2\%$, but that produces a large number of false negatives Wang et al. (2019). The architecture is a conventional CNN based on the popular VGG network Simonyan and Zisserman (2014). However, Wang et al. Wang et al. (2021) demonstrated that a more complicated CNN-based architecture that incorporated a spatiotemporal attention module performed better

[*] Corresponding author. Conrad Blucher Institute for Surveying and Science, Texas A&M University - Corpus Christi, Corpus Christi, TX, USA.
E-mail address: hamid.kamangir@tamucc.edu (H. Kamangir).
[1] Authors contributed equally.

https://doi.org/10.1016/j.envsoft.2022.105424
Received 19 August 2021; Received in revised form 18 March 2022; Accepted 18 May 2022
Available online 23 May 2022

than a conventional CNN for quantitative precipitation estimation. To better learn the underrepresented severe precipitation, the loss function was weighted by rain intensity.

FogNet (Kamangir et al., 2021) is an example of a CNN architecture that achieved high performance predicting coastal fog, outperforming, for example, the High Resolution Ensemble Forecast (HREF), an operational ensemble of numerical weather prediction (NWP) models. A complex CNN based architecture was developed to avoid overfitting and learn the process dynamic despite the relatively low number of fog cases. The architecture is based on 3D convolutions, physically-based feature groupings, dense blocks, and attention maps. This study uses the case of coastal fog predictions to investigate the benefits of some of these more complex features of CNNs.

It has been shown that when processing spatio-temporal images or images with a large number of bands such as hyperspectral imagery, 3D CNN-based models outperform the conventional 2D CNN-based variety by learning the complexity of the auto-correlated input dataset (Ma et al., 2019; He et al., 2017). The 3D CNN-based models are able to learn not only 2D spatial patterns and correlations between groups of pixels and a target but also learn spectral correlations between bands or temporal correlations between input variables. However, 3D CNN-based models are more computationally expensive compared to 1D- or 2D CNN-based models due to the larger number of parameters to train. Furthermore, if the order of the input variables or bands in a 3D image cube do not matter, there is no reason for the use of a 3D CNN-based architecture (Li et al., 2017). FogNet is a recent 3D CNN-based model trained on an atmospheric data cube with a large number of inputs (384 input variable maps) and with an architecture that is hypothesized to benefit from a physics-based ordering of the input variables.

Fog is a meteorological phenomenon consisting of very small water droplets near the earth's surface that reduces visibility to less than 1 km (Glickman, 2000; WMO, 2020). The low visibility associated with fog has an adverse effect on the transportation sector and contributes to vehicular and aviation accidents (Gultepe et al., 2019; Das et al., 2018). Fog droplets develop due to condensation within an environment characterized by high relative humidity which can range from unsaturated to supersaturated (Gultepe et al., 2007). High relative humidity can occur due to the addition of water vapor, cooling, or near surface mixing of air parcels with different temperatures (Glickman, 2000; Gultepe et al., 2007). Condensation into water droplets is aided by hygroscopic aerosol particles known as cloud condensation nuclei. The visibility reduction is due to what is termed the first indirect effect, whereby aerosols contribute to a cloud drop-size distribution characterized by a preponderance of smaller droplets, resulting in a larger surface-to-volume ratio and subsequent extinction and lower visibility (Twomey, 1974; Koračin et al., 2014).

Fog occurs within the planetary boundary layer (PBL), the lowest layer of the atmosphere that is directly influenced by the earth's surface. The PBL responds to surface forcings in a timescale of 1 hour or less (Stull, 1988). In particular, the warming and cooling of the earth's surface in response to radiation results in PBL changes via transport processes; the vertical transport of moisture, heat and momentum is dominated by turbulence, while horizontal transport is accomplished by the mean wind (Stull, 1988; Stensrud, 2009). Thus, these 3D transport processes directly contribute to PBL structure. The thickness of the PBL can range from 100 m to 3 km in time and space (Stull, 1988). Specific fog types tend to occur in association with unique atmospheric vertical structures. For example, an atmosphere characterized by a thin moist layer near the surface, and much drier air aloft, under clear skies and light wind, is conducive to radiation fog. Further, warm moist air (in the lower levels) approaching the Middle Texas Coast (United States), with or without stratus clouds aloft, contributes to the development of advection fog along the coast. Thus, the incorporation of 3D cubes of meteorological fog predictor variables within the lower atmosphere would capture the vertical and horizontal patterns corresponding to fog, and possibly account for the 3D non-linear processes contributing to fog formation, potentially resulting in skillful fog predictions. Gultepe et al. (2007) emphasized the importance of 3D prediction models to better predict various fog types.

1.1. Contributions

Our work makes the following contributions:

- Quantifying the advantage of using 3D vs. 2D kernels to capture interactions between variables and within vertical atmospheric profiles in addition to spatial features.
- Investigating the impact of physics based grouping and ordering of atmospheric input variables for a 3D CNN model.
- Investigating the importance of several feature learning modules for 3D CNN to better capture the complex interactions of meteorological input variables for fog prediction including dense block, attention mechanism and multiscale feature learning.
- Comparing the impact of learning spatial-wise and channel-wise features in parallel or in sequence.
- Investigating the importance and contribution of individual meteorological variables (features), and each input feature group, for fog forecasting by using four explainable artificial intelligence (XAI) techniques.

2. Methods

2.1. 2D convolutional feature learning

The core operation of CNNs is convolution over images to extract lower-dimensional features. A convolutional kernel is defined that repeatedly operates in a local window defined by the size of the kernel. The kernel acts as a moving window across the image's dimensions to calculate all the pixel values of the output feature map.

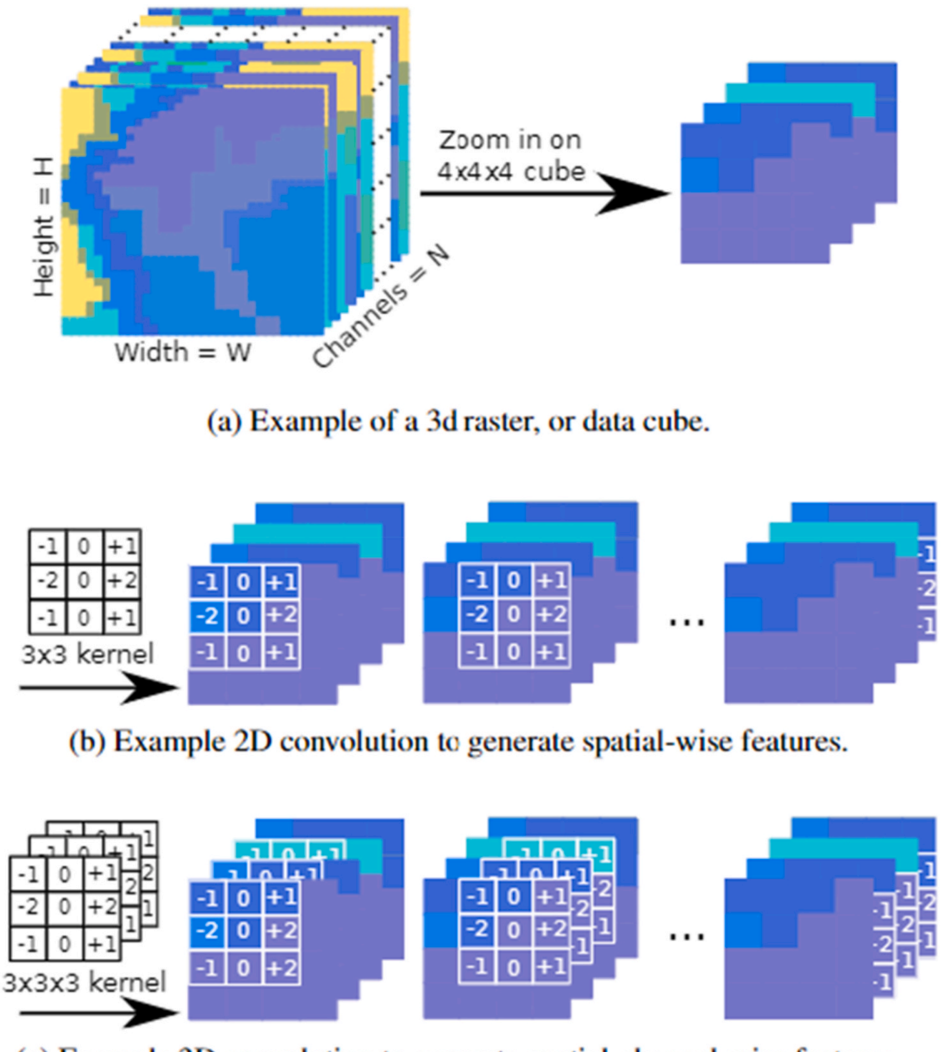
In traditional image processing, kernels are manually-designed to detect desired features (Simonyan and Zisserman, 2014). For example, the Sobel operator uses two 3 × 3 kernels to detect edges, one for horizontal edges and the other for vertical edges, based on an approximation of the gradient at the center pixel location (Kanopoulos et al., 1988). Many other kernels exist to perform operations such as blur, sharpen, detect other edge angles, etc.

These kernels are routinely used for image processing, including recognition tasks such as classification. Specific classes can be characterized by the combined outputs of a set of manually-selected kernels. Traditionally, image classification was based on hand-crafted kernels for feature extraction.

However, it is challenging to select the kernel values that best support image recognition tasks. This can be formulated as an optimization problem to select the values that minimize classification error. Thus, manual feature extraction can be replaced with data-driven learned feature extraction. CNNs have been shown to be an effective machine learning approach for automatic image feature extraction (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Huang et al., 2017). Convolutional layers perform convolution using kernels whose values are trainable parameters. A typical architecture contains convolution layers for feature extraction followed by fully-connected layers to make a prediction based on potentially highly nonlinear relationships between the features and the target class.

This approach is not limited to gray-scale (2D) or RGB (3D) visual images, but rather rasters of arbitrary dimensions. In the case of meteorological applications, the rasters' spatial dimensions typically represent a discretized spatial region while the channels (bands) represent separate environmental variables such as temperature, wind speed, or relative humidity. A 3D raster of meteorological variables is illustrated in Fig. 1a.

Even when working with multi-channel inputs such as RGB images, the majority of CNN applications focus on 2D convolution. That is, a 2D



(a) Example of a 3d raster, or data cube.

(b) Example 2D convolution to generate spatial-wise features.

(c) Example 3D convolution to generate spatial-channel-wise features.

Fig. 1. 2D and 3D convolution on 3D data cube to generate feature maps.

kernel is applied, as illustrated in Fig. 1b. Even though the kernel is calculated over multiple image channels, each operation involves raster values within a single channel. Thus, the output is 3D but each output channel contains only spatial-wise features. Alternatively, 3D convolution uses a 3D kernel to operate across channels, as well as spatially, to extract 3D features (see Fig. 1c). 3D convolution will be discussed in Section 2.2.

2.1.1. *Benchmark 2D CNNs*

In this section, three of the most common 2D CNN-based benchmarks, AlexNet (Krizhevsky et al., 2012), **ResNet** (He et al., 2015) and **DenseNet** (Huang et al., 2017, 2019), are discussed. Each of these architectures advanced the state of the art in visual recognition and are commonly used image classification benchmarks. While initially designed for RGB images, they can be adapted to support an arbitrary number of channels.

In 2012, AlexNet (Krizhevsky et al., 2012) won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012), and demonstrated that increasing the number of hidden layers dramatically enhances model performance. Relatively shallow networks were the norm, given the computational expense of learning the larger number of weights in deeper networks. However, GPUs were used to make it feasible to train the 8 hidden layers of AlexNet, and demonstrated that deep learning can dramatically outperform models based on human-selected features. The success of AlexNet led deep learning researchers to explore increasingly deeper and more complex architectures. ILSVRC-2014 was won by using even more hidden layers. Two variants of the VGG architecture, VGG-16 and VGG-19, were used where the *-16* and *-19* designations refer to the number of hidden layers. Eventually, however, additional layers were providing diminishing returns or even worse performance. The major problem was the vanishing or exploding gradient: applying backpropagation along the deep hidden layers results in multiplying so many weights that they either become 0 or arbitrarily large.

The major contribution of ResNet, winner of ILSVRC-2015, was architecture design techniques that would allow CNNs to efficiently scale to hundreds of hidden layers for improved model performance (He et al., 2015). Skip connections were introduced that mitigate the vanishing/exploding gradients. These are connections that flow from the input to each layer, skipping over the convolutions, to promote gradient flow. Also, bottleneck layers were included throughout the network for dimension reduction to limit the number of parameters to learn. By doing so, ResNet is able to have much deeper models than VGG while being significantly less complex. This was shown to enable feasible training of large models such as ResNet-152 (that is, a specific configuration of the ResNet architecture with 152 hidden layers) that actually have less parameters to learn than VGG-16 (He et al., 2015).

In 2017, Huang et al. developed DenseNet that was able to outperform ResNet by using dense blocks (Huang et al., 2017). With the dense block, every layer is connected to all subsequent layers. At each layer, input is the channel-wise concatenation of feature maps output from all previous layers. This is an extension of the skip connection concept, but promotes learning by allowing each layer to consider information from all previous layers. Thus, features learned at each layer are used more efficiently since all the subsequent layers have access. This allows the network to learn with fewer hidden layers. In addition, like skip connections, the feed-forward propagation of features avoids vanishing/exploding gradients.

These architectures have been shown to be effective in many domains besides RGB image recognition, including recent weather forecasting applications. ResNet was used by Rasp and Thuerey (Rasp and Thuerey (2021) for 5-day weather forecasting. First, climate simulation data was used to train an initial model. Then, transfer learning was used for additional training on real climate data. The model predicts geopotential, temperature, and precipitation. Given satellite imagery, Zanchetta and Zecchetto (2021) trained a ResNet model with Sentinel-1 satellite data to estimate wind direction over sea. Convolutions were performed over Synthetic Aperture Radar (SAR) images to learn to predict 2 km × 2 km wind direction fields. A modified ResNet was implemented by Bosma and Nazari (2021) to predict solar and wind energy production. Like FogNet, the input raster channels were weather data rather than visual imagery. The raster was a 155 × 108 spatial grid with 6 data channels: pressure, temperature, humidity, wind speed, wind direction, and cloud cover.

2.2. 3D convolutional feature learning

2D convolutional kernels extract the spatial correlation between pixels for each feature map. However, 2D convolutional kernels take a single map as input, so they fail to leverage context from adjacent feature maps. 3D convolutional kernels address this issue by moving the kernel in 3 dimensions (depth, height and width) as illustrated in Fig. 1a. The ability to leverage inter depth of the image and to learn context in correlation between different feature maps and channels can lead to improved performance for meteorological applications since there are meaningful relationships between different meteorological variables for event occurrence, especially variables of the same type (such as wind speed or temperature at various heights above the ground). But, using 3D ConvNets comes with a computational cost as a result of the increased number of parameters required by a 3D CNN-based architecture. Recently, 3D convolution kernels have been used in different deep learning architectures for weather and meteorological prediction (Niu et al., 2020; Wang et al., 2020; Castro et al., 2021).

Niu et al. (2020) proposed a new architecture for short time precipitation prediction based on a multi-channel ConvLSTM (Convolutional Long-Short-Term-Memory) and 3D-CNN. This architecture was trained on radar echo intensity data for 2017–2018 of south China with 1 km spatial resolution and 12 min intervals. They have shown for such time series data, having LSTM (Long-Short-Term-Memory) with 3D-CNN works better than only a 3D-CNN based model. Wang et al. Wang et al. (2020) applied a 3D CNN-based model for tropical cyclone intensity change prediction over a short temporal range of 24 h. They used 8 input variables including temperature, relative humidity, wind velocity (u and v wind components), geopotential height and sea surface temperature with 0.125° × 0.125° spatial resolution and a 6 h temporal interval. Castro et al. Castro et al. (2021) proposed a 3D-CNN based model, called STConvS2S, for weather forecasting and specifically air temperature and rainfall were tested. This architecture uses two different blocks to extract spatial and temporal representations of the input sequence data and they also used a temporal generator block on top of a spatial block to increase the sequence length of the time prediction.

To extend the applicability of 3D CNN-based models, the 3D CNN-based model used for fog prediction developed by Kamangir et al. (2021), called **FogNet3D**, is explained in the next subsection (section 2.2.1 FogNet3D). Also, different auxiliary modules for feature learning used by FogNet including dense block, attention mechanism and multiscale feature learning using 3D dilated convolutions has been explained.

2.2.1. FogNet3D

The FogNet3D (Kamangir et al., 2021) model (shown in Fig. 2) starts with separating the processing of input variables into five different groups based on their similar physical relationship to fog development. Each subgroup consists of a double parallel branch dense block feature extraction (spatial-wise and variable-wise) with an attention mechanism. The variable-wise and spatial-wise feature outputs for each subgroup from step 1 are concatenated into two main feature groups. In the next step, for each feature type, a 3D multiscale layer using dilated feature learning is used to extract new representation maps at different resolutions. At the end, the variable- and spatial-wise features are fused by using global average pooling and then a binary classifier used to generate a probability for fog or no fog. We investigate in detail the impact of the following five different characteristics on FogNet3D performance:

- **Physical grouping of meteorological variables:** Overfitting is a big challenge for all machine learning models especially when there is a high correlation between input variables which make the generalization of the model harder. Specifically for FogNet there are between 288–384 input variables that have physical correlation with other variables across and within input categories. The input data is categorized into 5 different groups based on their similar physical relationship to fog development as described below:
  - Group 1 emphasizes the influence of wind and contains the wind-related features $\mathrm{FRICV}_{\mathrm{surface}}$ (surface frictional velocity), $\mathrm{U}_{\text{10-meters}}$, $\mathrm{V}_{\text{10-meters}}$ (u and v wind components at 10-m height/elevation), $\mathrm{U}_{975\text{-}700}$, and $\mathrm{V}_{975\text{-}700}$ (u and v wind components at atmospheric pressure levels 975 mb–700 mb, at 25 mb increments).
  - Group 2 focuses on the influence of the combined effect of turbulence kinetic energy (TKE) and specific humidity (Q), and contains features $\mathrm{TKE}_{975\text{-}700}$ and $\mathrm{Q}_{975\text{-}700}$ (turbulence kinetic energy and specific humidity, respectively, at pressure levels 975 mb–700 mb, at 25 mb increments).
  - Group 3 incorporates the thermodynamic profile of the lower atmosphere and contains the features $\mathrm{TMP}_{\text{2-meters}}$, $\mathrm{DPT}_{\text{2-meters}}$, $\mathrm{RH}_{\text{2-meters}}$, (air temperature, dew point temperature, and relative humidity, respectively, at 2-m height/elevation), $\mathrm{TMP}_{975\text{-}700}$, and $\mathrm{RH}_{975\text{-}700}$ (temperature and relative humidity, respectively, at pressure levels 975 mb–700 mb, at 25 mb increments).
  - Group 4 accounts for the influence of surface atmospheric moisture and microphysics, and includes VIS (surface visibility), $\mathrm{Q}_{\mathrm{surface}}$ (2 m specific humidity), TLCL (temperature at the lifted condensation level) and $\mathrm{VV}_{975\text{-}700}$ (vertical velocity at pressure levels 975 mb–700 mb, at 25 mb increments).
  - Group 5 accounts for surface variables that control advection fog formation, including features SST (sea surface temperature), $\mathrm{DPT}_{\text{2-meters}}$ - SST (difference between 2 m dew point temperature and SST), and $\mathrm{TMP}_{\text{2-meters}}$ - SST (difference between air temperature and SST). Also, $\mathrm{TMP}_{\text{2-meters}}$ - $\mathrm{DPT}_{\text{2-meters}}$ (difference between 2 m temperature and 2 m dew point, otherwise known as the 2 m dew point depression), which is proportional to relative humidity.

  This helps to decrease the complexity of the input data and extract the correlated features individually from each group and then combine them for the next step to provide more distinguishable features for the classifier.

- **Parallel learning of spatial- and channel-wise features:** Previously, for spatiotemporal meteorological data (Castro et al., 2021) and remotely sensed hyperspectral data (Ma et al., 2019) it has been shown that separately learning representations of spatial-wise and

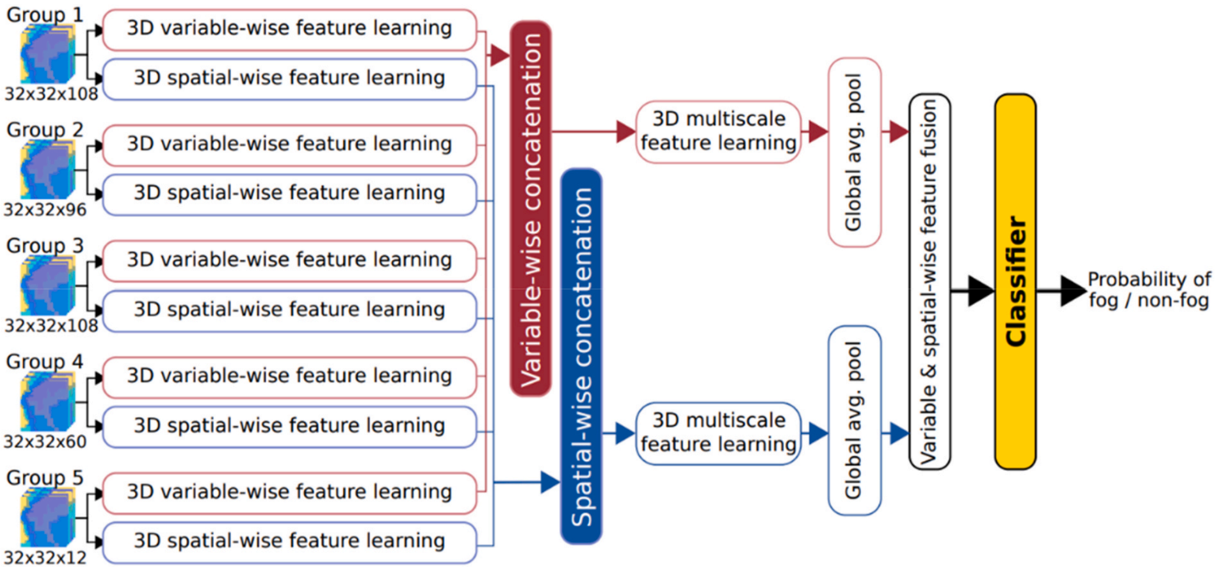

Group 1
32x32x108
Group 2
32x32x96
Group 3
32x32x108
Group 4
32x32x60
Group 5
32x32x12
3D variable-wise feature learning
3D spatial-wise feature learning
Variable-wise concatenation
Spatial-wise concatenation
3D multiscale feature learning
Global avg. pool
Variable & spatial-wise feature fusion
Classifier
Probability of fog / non-fog

Fig. 2. Overview of the FogNet3D parallel processing of features – spatial-wise (blue) and variable-wise (red). See Kamangir et al. (2021) for a detailed explanation.

channel-wise (or temporal-wise) features lead to better performance. Specifically, the impact of parallelizing the feature extraction of meteorological variables is investigated in the FogNet architecture.

- **Spatial- and variable-wise dense blocks:** When CNNs go deeper, the path for information from the input layer to the output layer becomes too long and gradient vanishing in the opposite direction from the output to input layer is a challenge. DenseNets (Huang et al., 2019) address this issue by simply connecting every layer directly with each other and all the previous layers and reusing instead of drawing representation power from extremely deep or wide architectures. FogNet takes advantages of two different dense blocks (step 1, Fig. 2), a spatial dense block with a kernel size of 3 × 3 × 1 to learn representation in the spatial domain of each feature map and a variable-wise dense block with a kernel size of 1 × 1 × 9 to learn the correlation between different input variables.
- **Spatial- and variable-wise attention blocks:** Attention mechanism (Xu et al., 2015) has been proposed to pay more attention to certain features when processing the data by CNNs. Attention mechanism manages and quantifies the interdependence between the input variables and the output elements by focusing on the most informative parts and suppressing the weights of other regions. FogNet consists of two different attention modules including a variable-wise attention module (step 2, Fig. 2) to focus on informative input variables and a spatial-wise attention module to extract informative areas from each input variable map.
- **Multiscale feature learning:** Multiscale feature extraction using convolutional kernels has been effective for classification problems (He et al., 2017; Srivastava et al., 2014). This is in part because multiscale convolutions have the power to extract more complex combined spatial-spectral features. The meteorological data includes 3D patterns with different spatial resolutions which have the potential to be quantified by using different kernels and receptive fields. Dilated convolution using expansion of receptive fields aggregates multiscale contextual information without loss of resolution or coverage (Yu and Koltun, 2015). In fact, dilated convolution modifies the convolution filter in different ways at different ranges using different dilation factors. In FogNet (step 4, Fig. 2), a multiscale 3D dilated convolution block is used to learn more complicated meteorological features.

3. Results/discussion

3.1. Study area and features

The FogNet study domain includes a portion of the Texas coast and the adjacent western Gulf of Mexico (see Fig. 3), and is organized as a 32 × 32 horizontal grid with 12 km grid spacing. The domain, (384 km × 384 km), is sufficiently large to account for atmospheric processes driving the formation of fog at the target location over a 24 h period (Orlanski, 1975) (the maximum forecast length of the FogNet predictions.)

The FogNet features (predictor variables) originate from a numerical weather prediction (NWP) model, the North American Mesoscale (NAM) modeling system, used operationally by meteorologists in the National Weather Service (United States), and from satellite imagery. The specific features used were chosen to predict fog by capturing the fog development process, or the lower atmospheric structure consistent with fog development, for the specific fog types that typically occur in the study domain. These fog types (corresponding mechanisms) include radiation fog (nighttime radiational cooling of moist air to saturation within a stagnant environment under clear skies in association with a high pressure system), advection fog (typically, the cooling to saturation of moist onshore flow by cool shelf waters along the Texas coast), advection-radiation fog (advection of near surface moisture onshore during the day followed by the radiation fog development at night), frontal fog (3 types, 2 of which involve rainfall which evaporates and moistens the sub-cloud layer to saturation, either in a post-cold frontal or pre-warm frontal environment, and one involving the mixing of distinct airmasses during frontal passage), and stratus-lowering fog (radiational cooling of the air at cloud top, which is transported downward by turbulent mixing and cools the sub-cloud layer to saturation and/or settling of cloud/drizzle drops that fall below cloud base, evaporate and cool the sub-cloud layer to saturation, resulting in the lowering of the cloud base to the surface.) See Table 1 in Kamangir et al. (2021) for detailed information regarding the NAM and the selection of the features.

The target used for FogNet originates from visibility measurements from the Automated Weather Observing System (AWOS) site at the Mustang Beach Airport (KRAS) (latitude 27.8118333°N, longitude 97.0887500°W) in the coastal city of Port Aransas, Texas. This AWOS was provided by Vaisala Inc, which provides AWOS model AW20, which is certified by the U.S. Department of Transportation Federal Aviation

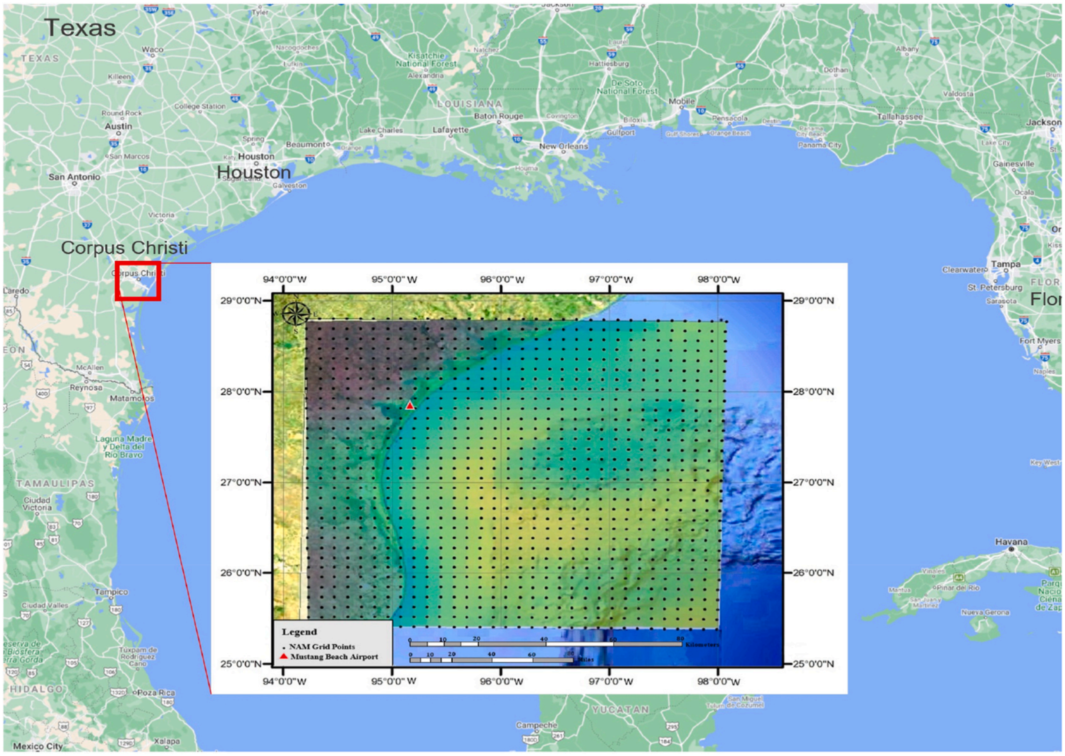


Fig. 3. The inset is an example of a 2D map of surface (2 m) air temperature within the domain used in this study: a 32 × 32 grid covering a 384 km × 384 km area of the Texas Coastal Bend including nearshore waters.

Table 1
Results for all of the models compared in our experiments described in Sections 3.2-3.4. The results are for binary fog predictions (24 h prediction of mist or fog with visibility ≤ 1600 meters or not). The values are performance, computational costs, the number of parameters for the model and the section the model is introduced. Performance metrics include false alarm rate (F), false alarm ratio (FAR), probability of detection (POD), critical success index (CSI), Peirce's skill score (PSS), Heidke skill score (HSS), odds ratio skill score (ORSS), and Clayton skill score (CSS). The performance values are the mean values based on 5 iterations of the 2D models and 10 iterations of the 3D models using the same multi-year data set used to develop FogNet3D.

Model	F	FAR	POD	CSI	PSS	HSS	ORSS	CSS	$\frac{TIME}{EPOCH}$(s)	Parameters($\times 10^6$)	Section
ResNet152	0.01	0.17	0.07	0.02	0.06	0.04	0.16	0.03	23	60.	3.2
DenseNet121	0.02	0.47	0.36	0.22	0.35	0.33	0.76	0.32	23	8.	3.2
DenseNet201	0.02	0.35	0.21	0.10	0.19	0.16	0.45	0.14	39	19.	3.2
FogNet2D	0.02	0.55	0.44	0.28	0.42	0.43	0.95	0.43	256	16.	3.2
FogNet3D-WithoutAttention	0.01	0.56	0.49	0.31	0.47	0.45	0.96	0.44	512	23.	3.3
FogNet3D-WithoutMScale	0.02	0.56	0.47	0.29	0.46	0.43	0.96	0.43	493	22.	3.3
FogNet3D-WithoutSpatial	0.02	0.53	0.40	0.27	0.38	0.41	0.95	0.45	190	8.	3.3
FogNet3D-WithoutSpectral	0.02	0.57	0.53	0.31	0.51	0.45	0.96	0.41	190	8.	3.3
FogNet3D-Sequential	0.02	0.58	0.53	0.29	0.51	0.44	0.96	0.40	196	8.	3.3
FogNet3D-WithoutGrouping	0.02	0.52	0.36	0.25	0.34	0.39	0.95	0.45	522	23.	3.4
FogNet3D-Shuffled	0.01	0.62	0.16	0.10	0.14	0.16	0.90	0.35	540	23.	3.4
FogNet3D	**0.02**	**0.50**	**0.54**	**0.35**	**0.52**	**0.50**	**0.97**	**0.48**	**544**	**23.**	**2.2.1**

Administration (FAA) and meets the FAA AWOS Advisory Circular 150/5220–16 for facilities that are not Federally owned (Vaisala, 2015; FAA, 2017). The AW20 Vaisala Present Weather Detector sensor (PWD22) generates 15 s visibility values that are averaged to generate 1-min and 10-min output values (Vaisala, 2004). The visibility value generated by AWOS for the user is the 10-min harmonic average. The accuracy of the PWD22 sensor for visibility is ±10 percent from 10 to 10,000 m and ±15 percent from 10 to 20 km (Vaisala, 2018). Vaisla Inc. provided the AWOS instrumentation for KRAS from 2011 to the present; the AW20 model was installed in 2018 (11 March 2022 personal communication from Randy Hansen, Airport Manager, Mustang Beach Airport).

The target vector was developed as follows: Each KRAS visibility measurement in the dataset is converted to one of 4 visibility categories ($\leq 1600\ m$, $\leq 3200\ m$, $\leq 6400\ m$, $> 6400\ m$). All visibility measurements $\leq 6400\ m$ caused by a weather phenomenon other than fog or mist were removed from the dataset. Thus, FogNet was trained to predict visibility restrictions due only to fog or mist. The training, validation, and testing data were extracted from the 2009 −2020 time series of NAM NWPs. The 2012–2017 part of the data was used for the training of the model, 5,460 cases (50%), 2009–2012 data was used for validation, 3,328 cases (30%), and the remaining of the data, 2018–2020, 2,228 cases (20%), was used for an independent assessment of the model after the completion of the calibration.

3.2. 2D vs 3D convolutional feature learning

For this comparison, three 2D CNN architectures were selected for comparison with FogNet. Each of the three architectures were trained on multiple hidden layer depths. The three models trained were ResNet-152, DenseNet-121, and DenseNet-201. The numbers refer to the model's number of hidden layers. Each model was trained using the adam optimizer for 100 epochs with a batch size of 64. A dynamic learning rate was used, beginning with an initial value of 0.1.

The deep learning framework PyTorch (Paszke et al., 2019) was used to train each of these models. We use the TorchSat (sshuair, 2020) package which includes PyTorch implementations of AlexNet, ResNet, DenseNet, and other popular CNN architectures. TorchSat is similar to the popular TorchVision, but supports an arbitrary number of channels where TorchVision supports only grayscale (1 channel) and RGB (3 channels). TorchSat allows us to train a fog detection model using the 384-channel input raster.

It is extremely common to use transfer learning when training these CNNs. That is, the initial weights are based on training on very large datasets such as ImageNet. The new model is able to take advantage of features already learned on the large-scale dataset and is adjusted with additional training to suit the new problem domain. Transfer learning has been shown to be effective even when the target dataset differs considerably from the original such as satellite images (Gadiraju and Vatsavai, 2020). However, based on the substantially greater number of channels, their non-visual nature, and for a fairer comparison with FogNet, which did not use prior training, transfer learning was not performed to construct these benchmarks.

A drawback of using off-the-shelf CNNs is that they expect a single raster input. For visual image inputs, it is reasonable to assume that all the channels will have the same dimensions. But when the grids are temperature, wind, etc. they may be of various sizes. Important features may be lost if scaling is used to construct a single raster. FogNet performs the scaling with a dimension reduction component of the model. Specifically, the SST is transformed from 384 × 384 to 32 × 32. Since the scaling is performed through convolution, the scaling that best helps the model to extract discriminating features is learned. For the benchmarks, however, the SST is simply downsampled with Gaussian smoothing for anti-aliasing.

For all the experiments run with FogNet3D the same hyperparameters have been used. To find the best FogNet hyperparameters, a grid search (section 11.4.3 of Goodfellow et al. (2016)) was applied. In all the experiments, the model is trained for 50 epochs with 32 batches per epoch on 5,460 training samples and 3,328 validation samples. For all the experiments, the learning rate (lr) is held constant at 0.0009, and dropout and L2 regularization are 0.4 and 0.001 respectively. The same architecture as FogNet3D was implemented to create a **FogNet2D** by using 2D convolutional kernels instead of 3D convolutional kernels. The purpose of FogNet2D is to investigate the impact of 3D convolutions used in FogNet3D as compared to the 2D kernels in FogNet2D. All experiments in this work, besides the 2D benchmarks, were trained using the Keras Python package (Chollet et al., 2018).

In this section, the results (for 10 iterations) for FogNet3D has been compared with 2D kernel-based models including FogNet2D, **DenseNet121**, **DenseNet201** and **ResNet152** based on the Peirce skill score (PSS), Heideke skill score (HSS), and the Clayton skill score (CSS) verification performance metrics for deterministic forecasts of binary events, and area under the receiver (or relative) operating characteristic curve (AUC) for probabilistic predictions of binary events.

The HSS and PSS measure the accuracy relative to the accuracy achieved by random forecasts. The accuracy measure used by the HSS (PSS) is the proportion correct (hit rate). The proportion correct is the fraction of all forecasts that were correct. The hit rate measures the fraction of observed events that were correctly forecast. The values of both metrics are within the [−1,1] range, and skill is demonstrated with values greater than zero. The CSS measures the difference between the conditional probability of an event given a forecast that the event will occur, and the conditional probability of an event given a forecast that the event will not occur. Skill is achieved when CSS > 0, which indicates that the event occurs more frequently when forecast than when not forecast. The value of 1 for any of these 3 metrics demonstrates a perfect forecast system. The PSS and CSS metrics are related to economic value. PSS represents the maximum potential economic value realized by users of the forecast system with cost/loss ratios equal to the base rate (climatology). The CSS represents the range of cost/loss ratios for which users gain economic value from the forecasts. See (Jolliffe and Stephenson, 2003; Wilks, 2011) for more information regarding these metrics.

Based on the results, shown as a box-plot in Fig. 4, FogNet3D has the best performance with an AUC of 0.94 (with CI 0.95%) and highest score for PSS ($Avg = 0.52$), HSS ($Avg = 0.50$) and CSS ($Avg = 0.48$) without overlapping of the interquartile range with all other 2D CNN-based models. Also, the interquartile range boxes, especially for HSS and CSS, of FogNet3D show a better stability and low variability of training process.

In contrast, the best 2D-CNN model was FogNet2D, which has an average AUC of 0.93, while 2D-CNN benchmarks, DenseNet121, Dense201 and ResNet152 are the next best performing models, with average AUC of 0.90, 0.88 and 0.74, respectively. Results show that for such a meteorological prediction application with having large number of variables, 3D convolutional feature learning is better able to learn the complex 3*D* structure of the atmospheric profile, in order to generate more accurate and skillful predictions. This performance enhancement is not surprising since atmospheric processes in nature occur in 3D. The better performance of FogNet2D in comparison to the 2D-CNN based benchmarks, including DenseNet121-201 and ResNet152, shows that the auxiliary feature learning modules used in the FogNet architecture (discussed in subsection 3.3) improve performance.

We also summarized the computational cost of the 2D/3D-CNN models based on our desktop (4 GPUs: NVIDIA RTX 1080S). The time of training per epoch was much higher for the 3D models than for the 2D models with the same batch-size. It took more than 3 h to train a 3D model, whereas the training time of 2D models was only half that time. Once the trained models were used for prediction, the difference of the computation costs between the 2D and 3D models was narrow, namely 0.02s vs 0.03s for processing a single image.

3.3. Ablation study of FogNet components

In the last few years, several methods were developed to improve the performance and efficiency of CNN-based models. These new methods include DenseNet and ResNet which extract features at different resolutions using dilated convolutions, applying the attention mechanism, etc. FogNet applies several of these modules to better approximate the complex relationship between input meteorological variables and fog prediction, to control overfitting and to better generalize when applied to novel data. In this section, we discuss an ablation study to evaluate the modules used by FogNet. The value of using an attention module, multiscale feature extraction, and spatial and spectral dense blocks was investigated by removing those modules from the FogNet3D architecture and then comparing the results with the base performance of FogNet3D. Also, FogNet3D uses two parallel branches, one for variable-wise and one for spatial-wise feature learning to extract these two types of features separately and then fuses them before classifying. To better understand the contribution of this strategy, the parallelism is removed and instead the spectral features are extracted first using a variable-wise dense block and then those feature maps are fed into the spatial-wise dense block to extract spatial features. The alternative of extracting the spatial features was not attempted since if we first extract spatial features then the outputs are feature maps generated by kernels and there are no more raw input variables to extract and build correlations from. Hence, in this experiment, we only test the first strategy and compare its performance with parallel feature learning. As shown in Fig. 5, FogNet3D shows improvement over most but not all of the derivative models in this ablation study, indicating that the FogNet3D modules are generally beneficial to the performance of FogNet3D.

As we mentioned in Section 2.2.1, attention mechanism is a new methodology to magnify the most important areas for each map and the

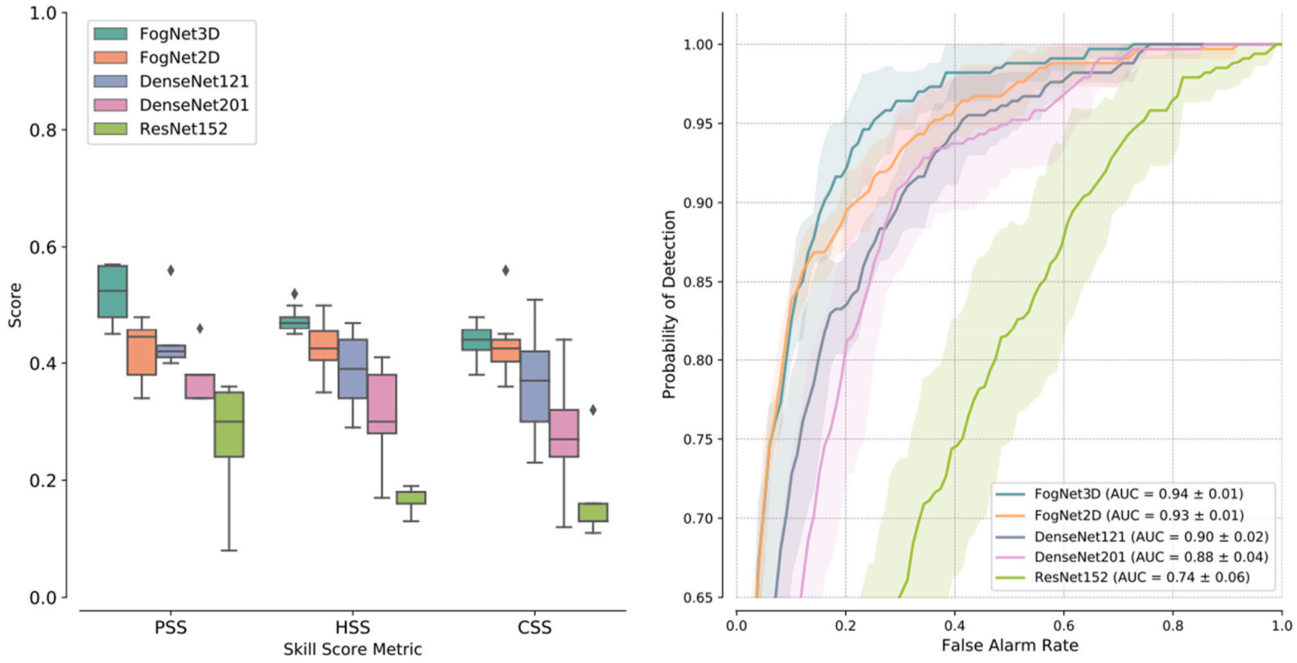


Fig. 4. Importance of 3D convolution. Left: Boxplot of the FogNet3D CNN performance vs 2D CNN-based models, when calibrated ten times, using three skill scores: Peirce's skill score (PSS), Heidke skill score (HSS), and Clayton skill score (CSS). The diamonds are outliers, outside the respective 1.5 * interquartile ranges' whiskers. Right: ROC curves for the same models with shading indicating standard deviation over the ten runs.

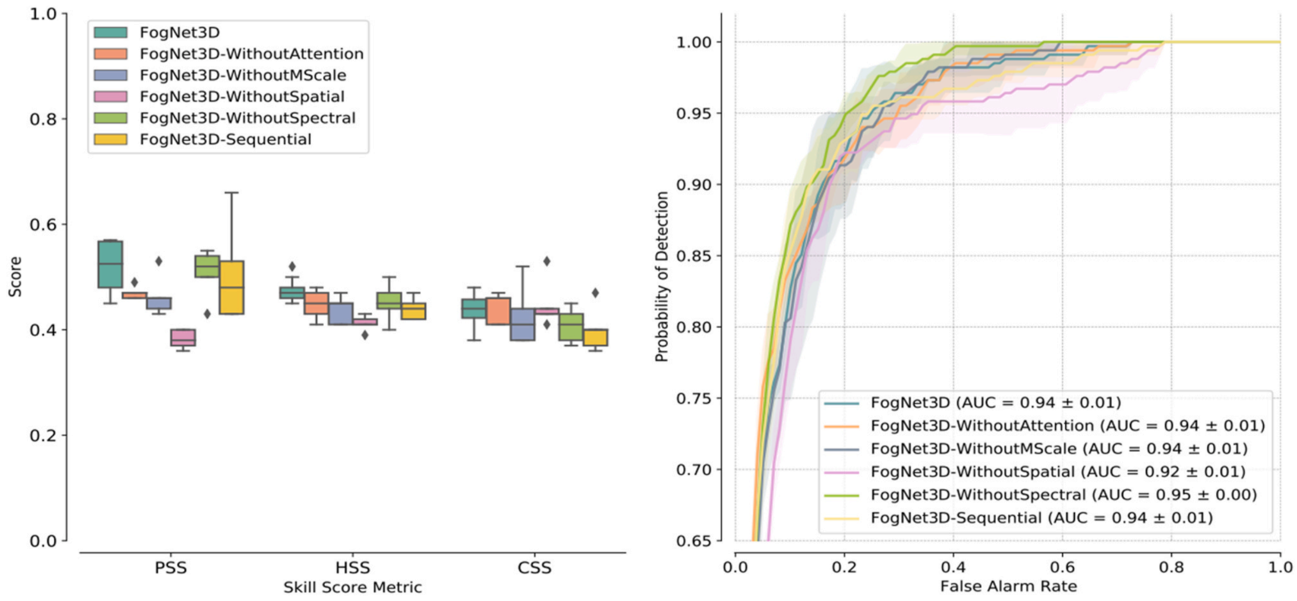


Fig. 5. Importance of FogNet3D architecture components. Left: Boxplot of the results of the ablation study for the FogNet3D, when calibrated ten times, using three skill scores: Peirce's skill score (PSS), Heidke skill score (HSS), and Clayton skill score (CSS). The diamonds are outliers, outside the respective 1.5 * interquartile ranges' whiskers. Right: ROC curves for the same models with shading indicating standard deviation over the ten runs.

most important feature maps for the classifier or decision maker. To investigate the importance of applying an attention mechanism for meteorological prediction applications, we compare the results of FogNet3D with and without attention mechanism. The results for skill metrics in Fig. 5 show that the performance of FogNet3D has decreased mainly for PSS since there is no overlapping between range values of their box-plots. Based on the results in Table 1, on average the results for FogNet3D without attention mechanism has a score between 4–5 points lower for PSS, HSS, and CSS.

To consider the importance of spatial-wise correlation and variable-wise feature map correlation learning, we ignore the impact of dense block feature learning by itself, more intuitively we are considering the importance of the spatial correlation between pixels and autocorrelation between different variables for meteorological applications. To do so, we remove the spatial-dense block in FogNet3D and the same thought for variable-wise feature learning by removing the variable-wise dense block. As we can see in Fig. 5 by removing the spatial-wise dense block the performance for FogNet3D decreased between 10–14 points for PSS, between 5–10 points for HSS and between 2–3 points for CSS. In comparison with spatial-wise, variable-wise feature learning decreases the performance less, only between 2–5 points for each of the skill metrics. The results for these two experiments may not be very informative since we have both spatial and variable-wise feature learning in multiscale block before the classifier for a combination of all different groups but as we can see these results show the impact of spatial correlation between pixels and variable-wise correlation between maps for meteorological problems, in this case fog prediction.

It has been shown that for 3D CNN-based hyper-spectral remote sensing image processing, separately learning spatial and spectral correlation can result in better performance (Ma et al., 2019). To do so, one strategy is learning these types of features in parallel and then fuse them before decision making. FogNet3D uses a parallel strategy, where each of the feature types, spatial- and variable-wise, is learned separately in parallel and then fused before classification. Another strategy is sequential learning, which is a common approach in CNN-based models, where variable-wise features are learned first and then fed into spatial-wise feature learning (or vice versa). The results in Fig. 5 shows that applying the sequential strategy decreased the performance of FogNet3D (mainly for HSS and CSS) with no overlap in the respective ranges of the second and third quartiles. In fact, this experiment introduces a new idea that the separate learning of spatial correlation between pixels for each map and auto-correlation between different input maps might help improve the performance of CNN-based models for meteorological applications.

Due to the complex interaction between meteorological variables and event occurrence for meteorological applications along with imbalanced conditions and high complex correlation between the variables, using only current CNN-based computer vision techniques is insufficient. In this section, several modules to improve the CNN-based model with their specific contribution to FogNet3D's performance are introduced. Based on our results, shown in Fig. 5, each of the modules (attention mechanism, spatially and spectral dense block feature learning, multiscale feature extraction by using dilated convolution, parallel extraction of spatial and variable-wise features) has made varying levels of contribution to the performance of FogNet3D suggesting that they are useful modules for CNN-based meteorological applications.

For FogNet3D the more than 200 input meteorological variables were categorized into 5 different input groups each based on their similar physical relationship to fog development. We investigated the impact of grouping input variables in such a 3D CNN-based model by training the FogNet3D model using all input variables in only one cube. The results in Fig. 6 show that AUC decreases from 0.95 to 0.91 for FogNet3D without grouping the input variables. The box-plot results for PSS and HSS also clearly indicate that performance deteriorates with similar results for CSS. Also, based on the results for the average of 10 training runs (Table 1), PSS is 35% lower for FogNet3D without grouping (0.52 for FogNet3D compared to 0.34 for FogNet3D without grouping), 22% lower for HSS (0.50 for FogNet compared to 0.39 for FogNet3D without grouping), and 6% lower for PSS (0.48 for FogNet3D compared to 0.45 for FogNet3D without grouping).

The results for this experiment show that tying the parameters of several parallel networks for each group leads to an improvement of FogNet3D performance (6–35 for the skill metrics). This means that reducing the number of free parameters by sharing them between group feature learning leads to better generalization by reducing overfitting.

3.4. *Importance of meteorological variable order in atmospheric cube*

Given a 3D weight tensor, channel order is important which means changing the order of input channels would change the performance of the CNN model. Also, for meteorological applications, due to the 3D nature of atmospheric processes and the associated correlation between different variables in the atmospheric cube, the order of variables in a 3D cube for a 3D-CNN is important. In this experiment, to investigate the importance of physically ordering of the input variables in a cube, we shuffled the order of the input variables for all variables in FogNet3D model and check the performance with FogNet3D without shuffling.

The features in FogNet were chosen to capture the vertical structure of the lower atmosphere, including the PBL, and to utilize atmospheric variables that modulate fog development. Thus, within groups 1 through 4, many features were ordered sequentially in the vertical direction by atmospheric pressure level (975 mb–700 mb, at 25 mb increments) such that the feature representations to the machine learning model are vertical profiles of variables that influence fog development. The ordering of features in group 5 was arbitrary since the focus was only to capture the advection fog process based only on surface and near surface

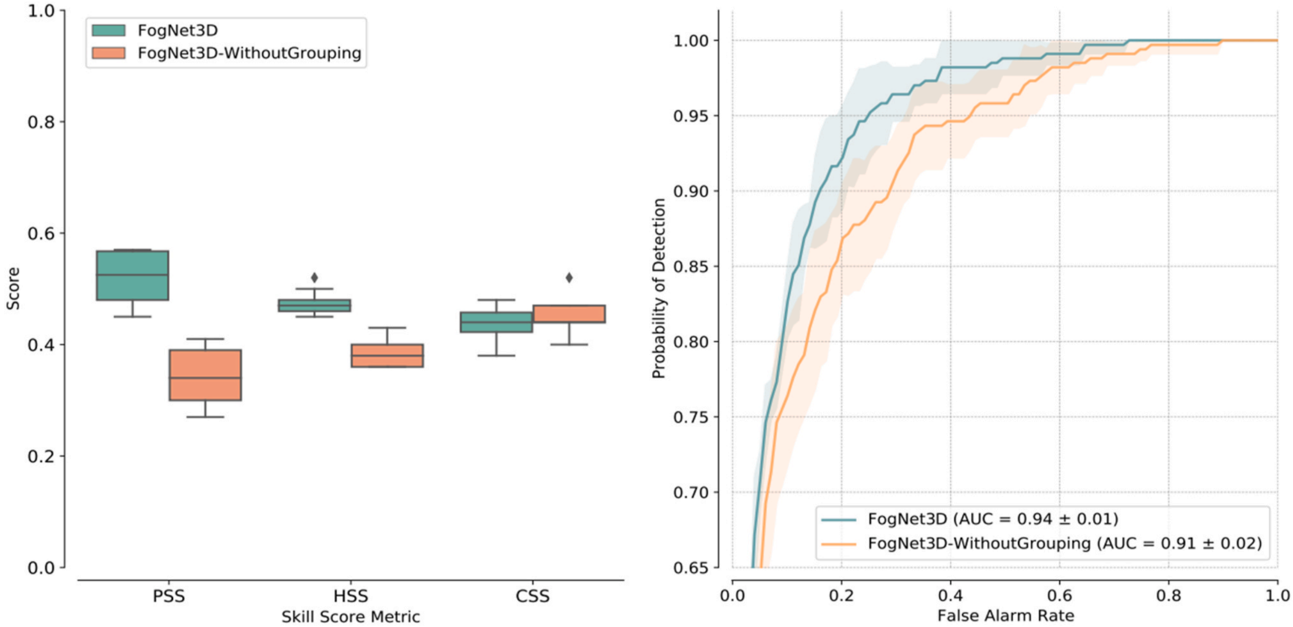


Fig. 6. Importance of physically grouping atmospheric variables. Left: Boxplot comparison of FogNet3D performance with and without Physics-based grouping, when calibrated ten times, using three skill scores: Peirce's skill score (PSS), Heidke skill score (HSS), and Clayton skill score (CSS). The diamonds are outliers, outside the respective 1.5 * interquartile ranges' whiskers. Right: ROC curves for the same models with shading indicating standard deviation over the ten runs.

variables.

Fig. 7 shows the results for FogNet3D with and without shuffling. The average of 10 training AUC scores for FogNet3D with shuffling is 0.06 lower than FogNet3D, showing degradation in performance for FogNet3D with shuffling. Also, the box-plots of the results show a large gap in performance between FogNet3D with and without shuffling. PSS for FogNet3D with shuffling is around 0.13–0.16 vs. 0.48–0.58 for FogNet3D and HSS for FogNet3D with shuffling is around 0.15–0.20 vs. 0.47–0.51 for FogNet3D. For CSS there is overlap between the respective second and third quartile of the result distributions. However, this is due to the large range of the third quartile for the shuffled cases which could be due to two or three outliers given that the results are based on ten cases. It is likely that more repetitions would result in a narrower confidence interval. And the median CSS for the shuffled case is substantially lower than for FogNet3D.

Also, based on the average of 10 training runs given in Table 1, the PSS score is only 0.14 for FogNet3D with shuffling compared to 0.52 for FogNet3D, HSS is only 0.16 with shuffling compared to 0.50 for FogNet3D, and for CSS the score with shuffling is 0.35 compared to 0.48 for FogNet3D. The results after shuffling the feature maps show that for 3D kernel CNN-based models the order of input variables for meteorological application is important because in this situation, the potential connections between input variables that have a high correlation regarding the event prediction have been removed so the feature maps generated by 3D convolutional kernels are not meaningful for the model to make a skilled decision.

To contextualize from a meteorological and representative learning perspective, temperature, moisture, and wind-related features in groups 1 through 4 within FogNet3D were ordered sequentially in the vertical direction (at atmospheric pressure levels from 975 mb to 700 mb at 25 mb increments) to form vertical profiles of the lower atmosphere both physically consistent with those that occur in nature and strongly related to fog development. These profiles became feature representations to the FogNet3D architecture, which allowed FogNet to relate these profiles to fog prediction during model training. For example, consider the wind profile generated by the u and v wind components (G1 features). A clockwise turning of wind direction with increasing height above KRAS corresponds to warm air advection, a component of the advection fog process along the Texas coast (United States) during the fog season. In addition, strong vertical wind shear near the surface (the u,v wind at the adjacent 10-meter and 975 mb levels) can preclude radiation fog. Further, an increase in q (specific humidity) with height in the lower levels (G2) is essential for radiation fog. Finally, radiation fog generally requires specific temperature and relative humidity profiles (G3) which depict a thin nearly saturated layer near the surface, followed by much lower relative humidity values aloft, within a temperature inversion. When the features were shuffled, training of the FogNet3D-Shuffled model was likely unsuccessful in capturing the relationship that maps the profiles of temperature, moisture, and wind to fog. In other words, the G1, G2, and G3 profiles mentioned above that correlate to fog are destroyed when the features are shuffled. Hence, the significant drop in performance (e.g HSS drops from 0.50 to 0.16).

Table 1 presents an extended set of metrics for all of the models tested. This includes FogNet3D, the 3D CNN-based model, all the variations of the FogNet3t3D of the Ablation Study (Section 3.3), removal of the physical grouping (Section 3.4) along with several popular 2D CNN-based models (ResNet152, Dense121, DenseNet201 - Section 2.1.1) and FogNet2D (Section 3.2).

3.5. *Investigating the influence of meteorological variables*

Due to their black box nature, it is difficult to determine how a trained model uses the data to make predictions. However, it is useful to have some understanding of the strategies learned by the model. This has motivated the rapidly developing field of XAI, where various methods have been proposed that probe the model in some way to learn about the model's input-output relationships (see Murdoch et al. (2019)). For example, Lapuschkin et al. (2019) demonstrates that a model performing well even on the testing dataset may rely on spurious associations in the dataset that would lead to poor real-world performance. McGovern et al. (2019) provide a detailed discussion of the application of XAI for meteorological models. Unfortunately there is no

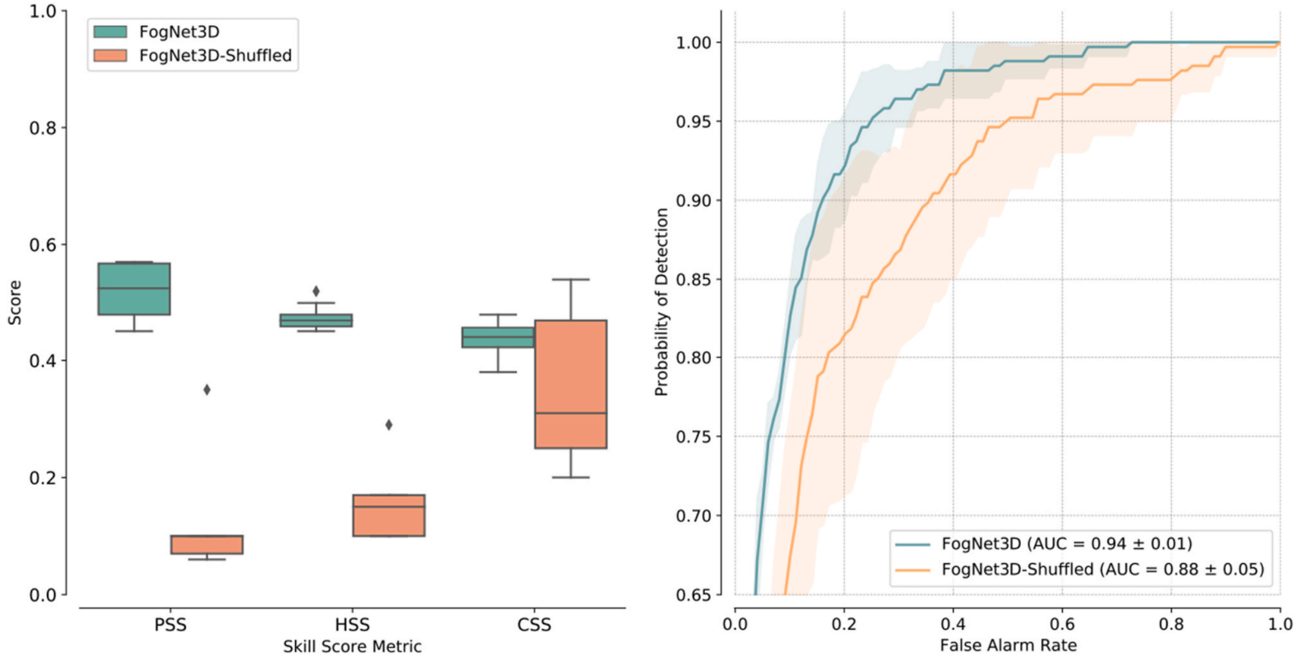


Fig. 7. Importance of Physically ordering atmospheric variables in 3D CNN. Left: Boxplot comparison of FogNet3D performance with and without shuffling of the order of the feature maps within the groups, when calibrated ten times, using three skill scores: Peirce's skill score (PSS), Heidke skill score (HSS), and Clayton skill score (CSS). The diamonds are outliers, outside the respective 1.5 * interquartile ranges' whiskers. Right: ROC curves for the same models with shading indicating standard deviation over the ten runs.

single technique guaranteed to provide a complete and accurate explanation of the model Molnar et al. (2020). Properties of the data such as dependencies and interactions can yield misleading explanations Molnar et al. (2020). Similarly, model architecture can influence XAI usefulness, such as the gradient shattering that may occur in very deep models Mamalakis et al. (2022). Since the true explanation is unknown, McGovern et al. (2019) suggests applying multiple XAI methods. When multiple methods consistently highlight certain features, it suggests that those features are truly influential for the model.

We have applied four XAI methods to investigate the influence of input raster features on FogNet3D. Three of these, Group-hold-out, Permutation Feature Importance, and LossSHAP, are used to analyze the importance of the five metocean groups. Feature importance is based on how much each feature, here a group of adjacent raster channels, affects the overall model loss. The three methods and their results are described in Section 3.5.1. We have also use Channel-wise Partition-SHAP to analyze the feature effect of individual raster channels. Feature effect is based on how much the feature (channel) contributes to an output prediction. This technique is discussed in Section 3.5.2. In a discussion of XAI pitfalls, Molnar et al. (2020) warns against confusing feature importance and effect when interpreting XAI outputs. Here, we use the term *feature influence* to collectively refer to both importance and effect.

3.5.1. *Group-wise feature importance*

We are interested in the relative importance of each of the five metocean variable groups. The features where chosen based on their predictive relationship with fog, but that does not guarantee that the model learned to take advantage of them. This could be because it did not need to, that is, they did not provide significant additional information compared to other relationships learned. Or, they would be useful for the model but the model was unable to learn them, perhaps because of their complexity or a lack of variation in the training data.

To analyze the group importance, we use three XAI techniques that calculate global feature importance scores based on how the absence of each group changes the HSS. Intuitively, if removing a group G_i causes a 10% reduction in HSS and another group G_j causes only 4% reduction, then G_i is assigned higher importance toward model performance.

Group-hold-out: The most straightforward strategy is to simply remove entire groups and retrain. For each group, that group is removed from the FogNet3D architecture and the model is retrained. Because of variations between trained models, 10 trials are performed for each group. For each trained model, we calculate the change in HSS to that of FogNet3D (HSS = 0.5). Each group's importance score is the average of the 10 trials.

The advantage of Group-hold-out is that it directly tests the impact of the input features on model performance. This is in contrast to the other methods that, as will be discussed, use techniques to imperfectly simulate the removal of features. The Group-hold-out approach is often not used in XAI studies, because every hold-out requires retraining the model. Or, as here, multiple times per feature to obtain an average. To perform Group-hold-out to evaluate every element of the 32 × 32 × 384 FogNet input raster would be computationally impractical. But to do so for only five grouped features is tractable.

Given that we are able to use Group-hold-out directly, a reasonable question is why other feature importance methods are needed for group-wise analysis. The major disadvantage of Group-hold-out is that it does not reveal the importance for the specific trained model under investigation. It is reasonable to expect that the group-hold-out results are similar to that of an individual model, but the variation is such that a given model might have learned strategies not represented by the repeated retrainings used to generate the Group-hold-out score. Molnar (2020) discusses disadvantages of testing importance with retraining, with an example of how it can potentially produce misleading interpretations of feature importance. Instead, Molnar (2020) recommends using Permutation Feature Importance.

Permutation Feature Importance (PFI): Similar to Group-hold-out, PFI (McGovern et al. (2019)) calculates feature importance based on model performance (here, HSS) with and without features being present. But since models almost always take in a fixed-size raster input, the feature cannot be simply removed. Instead of retraining, PFI simulates feature removal by permuting the feature's values. Here, the values within the permuted group are randomly shuffled to break the relationship between input variables and the target.

One issue with PFI is that the input features generated from shuffling are not actually the same as completely removing the feature, and the model output may simply reflect the response to unrealistic (out-of-distribution) data rather than no data as desired. Also, PFI may struggle with correlated features, as the importance scores may be divided among the group of correlated features. With FogNet, we expect input data to have strong correlation, by design, across the 32 × 32 spatial maps and across channels that represent vertical atmospheric profiles. With PFI, the importance scores may be diluted across the features such that none appear important despite being used by model. A way to mitigate this is by grouping features (Molnar et al. (2020)). Here, we expect that the 5 groups are distinct enough to allow meaningful PFI scores. Again, see Molnar (2020) for a discussion of the advantages and disadvantages of PFI.

LossSHAP: Lundberg et al. (2019) proposed a game-theoretic alternative to PFI inspired by Shapley values. Shapley values have been proposed to add rigor to XAI (Messalas et al. (2019); Fryer et al. (2021)). Based on cooperative game theory, Shapley values are a fair assignment of payout to each player in a game based on their contribution to the outcome. Applied to XAI, the game is an individual model prediction and the players are the features. The Shapley values describe each feature's contribution to the model output.

SHAP (Lundberg and Lee (2017)) is an implementation for approximating Shapley values to calculate feature effect. This is a local XAI technique, meaning it explains the contribution of the features for a single input instance. LossSHAP is a version of SHAP used to calculate global feature importance (Lundberg et al. (2019)). With SHAP, effect scores are based on the contribution of a feature to a specific output. Instead, LossSHAP applies SHAP's strategy for permuting the features based on the marginal contribution, but the entire dataset is used to calculate the difference in loss (here, HSS) to measure performance.

For group-wise LossSHAP, an importance score is calculated for each of the groups by averaging over the marginal contribution of that group based on all combinations of including or not including the other four. A group is said to be removed by replacing all cells with random values. This is similar to PFI, but with a critical distinction that allows it to take into account dependencies between groups. Suppose the feature under evaluation is group 3 (G3). PFI simply compares the model output with and without replacing G3 with permuted values. LossSHAP does this multiple times, but each time with some of the other groups also permuted. To calculate the LossSHAP value for G3, it is necessary to test with groups 1, 2, 4, and 5 present, then with 1, 2, 4, present and 5 permuted, and so forth for all possible combinations. Because minor variations between LossSHAP implementations exist, the following equation shows exactly how we calculate the HSS-based importance score for G3 using input raster X:

$$\begin{aligned} LossSHAP_{G3}(X) = \quad & w_1 MC_{G3,\{G3,G1\}}(X) + w_2 MC_{G3,\{G3,G2\}}(X) \\ & + w_3 MC_{G3,\{G3,G4\}}(X) + \dots \\ & + w_{16} MC_{G1,\{G1,G2,G3,G4,G5\}}(X) \end{aligned}$$

Where MC is the subtraction of prediction for G1 and G3 and prediction for only G3 based on the HSS value:

$$\begin{aligned} MC_{G3,\{G3,G1\}}(X_i) = \quad & HSS(Predict_{G3,G1}(X)) \\ & -HSS(Predict_{G3}(X)) \end{aligned}$$

Also, W is the weight of the marginal contribution where 1) the sum of all the weights are equal to 1: $W_1 + \cdots + W_{16} = 1$, 2) the sum of the weights for weights in the same level of combination are equal:

$$W_1 = W_2 + \ldots + W_5 === \\ W_6 + \ldots + W_{11} = \\ W_{12} + \ldots + W_{15} = \\ W_{16}$$

and 3) all the weights in the same level are equal, for example for level 2 with having two groups combination: $W_2 = W_3 = W_4 = W_5$.

SHAP is becoming increasingly popular because of fairness guarantees and convergence to a single global optimum (Lundberg and Lee (2017)). Even so, it may still be susceptible to the out-of-sample input and correlated features problems of PFI. It is also substantially more complex since each feature requires computing with combinations of removing the other features. Again, the complexity is less of an issue when dealing with only 5 grouped features. Molnar (2020) further discusses the advantages and disadvantages of SHAP.

Fig. 8 presents the results of these techniques. For each technique the scores of the groups are normalized so that the relative importance of each group for each technique can be compared. Despite using grouping to mitigate issues that stem from correlated features, it is well-documented that the accuracy of explanation produced by XAI techniques is affected by a variety of subtle issues (McGovern et al. (2019)). Also, grouping does not completely partition the input raster into uncorrelated features. Within each feature group, the features chosen have a similar physical relationship to fog development. However, correlations exist across the groups. G1 and G2 are correlated by definition. Instantaneous wind velocity (U) (group 1) can be written as U = $\overline{U} + U'$, where $\overline{U}$ represents the mean wind over a period of time, and U' represents the turbulent part which is related to TKE in G2. (Stull, 1988). Also, G1 and G4 are related since divergence (convergence) of the 2D wind field at the surface (G1) result in downward (upward) vertical velocities (G4) immediately aloft, owing to the conservation of mass. G1, G2 and G3 are also correlated. For example, a temperature inversion (G3), representing atmospheric static stability, can suppress turbulence (TKE in G2) (Stull, 1988), and also prevent the vertical mixing of greater momentum from aloft to the surface (thus lower surface wind speeds represented in group 1). Also, during advection fog, groups 3, 4, and 5 are correlated given that advection fog (accounted for in the group 5 features), implies high surface relative humidity and low visibility, group 3 and 4 features, respectively. This makes it challenging to determine which method is most trustworthy. Observing disagreement between the methods in Fig. 8, we suggest not overemphasising the exact values between methods but rather the overall impression they suggest concerning the model.

G1, G3 and G4 appear to be most important based on these methods. G2 and G5 are also important, but seemingly less so. G2 is the most ambiguous; LossSHAP and Group-hold-out suggest very low importance, but the PFI ranking is comparable to G4. It also has the largest difference between methods (LossSHAP and PFI). Based on these results, all groups appear to have a substantial impact on model performance, validating their inclusion as FogNet inputs. Only the LossSHAP ranking for G2 would suggest that a group might not be necessary. In aggregate, there is evidence that G2 does help FogNet but less so than the others. This could be used to spin off new experiments with G2 such as varying the horizontal spacing between channels or increasing total vertical distance with additional channels.

Fog is strongly controlled by wind (G1), and certain fog types that occur over KRAS (the target location used to develop FogNet3D) are statistically correlated to specific wind velocities. Radiation and advection-radiation fog generally cannot develop with surface (10-m) wind speeds greater than around 2.5 ms-1 (Tardif and Rasmussen, 2007; Koračin et al., 2014). Most of the advection fog events at KRAS occur when wind is onshore. Further, the cold front post-frontal and warm front pre-frontal fog types at KRAS generally occur when the wind has a significant north component. The wind profile represented in G1 has an influence on radiation fog since vertical wind shear (change in wind velocity with height) and horizontal wind velocities influence the vertical structure of radiative cooling associated with radiation fog (Dupont et al., 2016). Veering (clockwise-turning) wind with height corresponds to warm air advection (Wallace and Hobbs, 1977); the advection of warm moist air over the cooler sea surface contributes to advection fog (Koračin et al., 2014; Huang et al., 2015). If the turbulence-generation effect of wind shear (change in wind velocity with height) exceeds the turbulence-suppression effect of atmospheric buoyancy in the PBL,

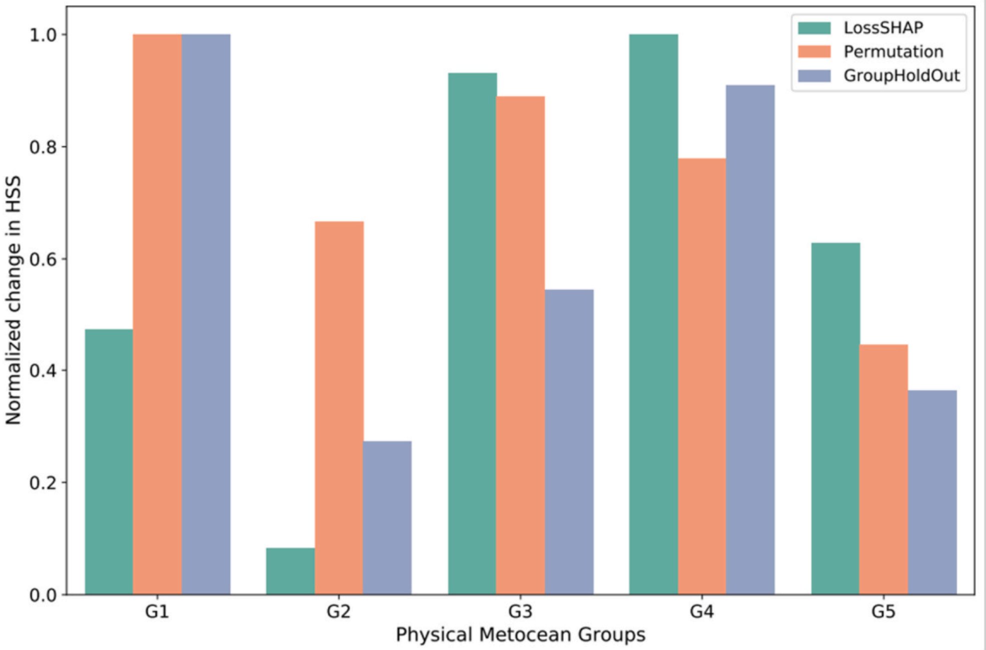


Fig. 8. Feature importance of metocean groups using XAI methods LossSHAP, group permutation and group holdout using the normalized changes in Heidke Skill Score (HSS). G1: influence of wind, G2: influence of turbulence kinetic energy and specific humidity, G3: influence of thermodynamic profile, G4: influence of surface moisture and microphysics, G5: influence of sea surface temperature (see Section 2.2.1).

radiation fog is not likely to occur (Baker et al., 2002). Lastly, frictional velocity magnitudes become miniscule during radiation fog events (Liu et al., 2011)

PFI also identifies G2, G3, and G4 as important, with normalized HSS changes ranging from approximately 0.65 to 0.85. Although the purpose of G2 was to account for cases whereby drier air aloft mixes vertically downward and dissipates or precludes fog (Toth et al., 2010), the G2 features influence fog in other ways. For example, very low TKE magnitudes and high TKE dissipation rates are correlated to stratus-lowering fog (Dupont et al., 2016). Further, surface TKE magnitudes become very small during radiation fog development (Liu et al., 2011). However, along coastal regions when conditions are favorable for advection fog (advection of warm moist air over the cooler sea surface), mechanical turbulence (due to vertical shear) in the statically-stable layer within a few hundred meters of the surface, can force a pre-existing stratus or stratocumulus cloud to the surface to produce fog (Huang et al., 2015). Furthermore, it is theoretically possible for the turbulent mixing of nearly saturated wind eddies near the surface to produce fog under certain conditions (Price, 2019). G3 is important since this group was developed to capture the vertical profile of atmospheric temperature and relative humidity, which has a strong influence on fog development, and the fog type. Based on an assessment of fog cases at KRAS for the 2009–2020 period used to train, validate, and test FogNet3D (not shown), radiation fog requires a vertical profile characterized by a thin moist/saturated layer near the surface, followed by significantly drier conditions aloft, while advection and stratus-lowering fog cases tend to occur with a slightly deeper moist layer than associated with radiation fog. The lower radiation versus stratus-lowering fog saturated layer depths are consistent with results in Dupont et al. (2016). Lower moist layer depths associated with radiation relative to advection fog are consistent with results from Croft et al. (1997). Cold front post-frontal and warm front pre-frontal fogs are correlated with an even deeper moist layer. Results from (Oliver et al., 1978; Dupont et al., 2012, 2016) suggests that for stratus-lowering fog events, the initial altitude of the pre-existing stratus cloud layer (which can be identified by the relative humidity profile below 750 mb in G3) should be 1 km or less. With respect to the temperature profile, radiation, advection-radiation, advection, warm front pre-frontal fog, and cold front post-frontal fog occur under strong temperature inversions (temperature increase with height) (Stull, 1988; Glickman, 2000; Koračin et al., 2014; Dupont et al., 2016). FogNet3D involves the post-procesing, via deep learning, of a select group of variables from the NAM (and from satellite-derived SST data) to make predictions of visibility categories associated with fog and mist. Knowledge of the magnitudes and spatial distribution of microphysical variables liquid water content, fog droplet number concentration, and particle size, is essential for skillful prediction of low visibility due to fog in NWP models (Gultepe et al., 2017). The NAM lacks the resolution necessary to resolve these microphysical-based variables and thus are parameterized (formulate implicit effects in terms of resolved fields). The NAM bulk microphysics parameterization scheme predicts cloud water mixing ratio yet retains a constant cloud droplet number concentration (is single moment with respect to cloud water). The G4 feature VIS is the NAM prediction/diagnosis of visibility based on an empirical relationship between the mass of cloud liquid water to the extinction coefficient. Although a direct relationship between fog and the actual microphysical processes responsible for fog is not accounted for in the NAM, the VIS feature represents an attempt to relate such based on microphysics parameterization. As mentioned earlier, stratus-lowering fog involves the lowering of pre-existing stratus or stratocumulus clouds, with cloud bases ≤ 1-km elevation/height, to the surface. Clouds develop in response to the activation of cloud condensation nuclei; this activation process is modulated by vertical velocity and cloud base temperature (Gultepe et al., 2017). G5 scored the worst with normalized HSS changes of around 0.45. This group accounts for the development of advection fog when moist onshore flow moves over the cooler shelf waters near the Middle Texas Coast. Although the majority of fog cases analyzed in this study were of the advection type, the features in this group neither capture processes responsible for the other fog types, nor accounts for the relationship between the vertical profile of various features and advection fog. Since G5 features are only relevant during advection fog cases, yet the other 4 groups are relevant to all fog types, we speculate that a global XAI technique would rank G5 near the bottom.

LossSHAP suggests that the vertical structure of temperature and relative humidity, and cloud microphysics (G3 and G4, with normalized HSS changes between 0.90 and 1.0, respectively) were much more important than the wind profile, surface features that capture the advection fog process (G1 and G5, with normalized HSS changes between 0.45 and 0.6), and TKE and Q profiles (G2 with normalized HSS changes less than 0.10) when predicting fog via FogNet (Fig. 8). It is not surprising that the contribution of microphysics (G4) was greatest using LossSHAP since the low visibility associated with fog is the direct result of a microphysical process known as the first indirect effect mentioned in the Introduction, and captured by the NAM VIS output. Group 4 feature VIS represents the NAM prediction/diagnosis of visibility due to various hydrometeors, including fog. We speculate that the post-processing of the NAM VIS by FogNet removed systematic errors in VIS, resulting in more accurate/skillful visibility predictions. The normalized HSS changes clearly indicate that the vertical temperature and relative humidity profiles (G3) are essential to skillful fog prediction. For all fog types in this study, a strong lower-level temperature inversion (temperature increase with height) is essential for suppressing the fog dissipative effects of turbulence/vertical mixing (Baker et al., 2002; Toth et al., 2010). Although wind (G1) has the greatest importance than the other 4 groups from a PFI perspective, when wind is forced to “compete” with the other 4 groups (LossSHAP), the vertical temperature and relative humidity structure and microphysics exerted a greater contribution to fog prediction skill. The strength of G5 in this competition is likely due to the fact that the majority of fog cases in this study were of the advection type and thus resulting in a significant contribution with the third greatest normalized HSS change. LossSHAP suggests that G2 has limited influence to the overall skill of FogNet. However, the vertical profile of TKE and Q are strongly related to fog development. An atmospheric layer characterized by a decrease in Q with height, combined with TKE, can result in fog dissipation (Baker et al., 2002; Toth et al., 2010). The collinearity of TKE and Q with other features may explain the low importance of G2 per LossSHAP.

With respect to the Group-Hold-Out XAI method applied to FogNet predictions, the group feature importance scores were similar to that of PFI wherein G1, G3, and G4 have greater importance than G2 and G5. This is not surprising given the similarities of the group-hold-out and PFI. However, only G4 importance improved (greater normalized HSS) when the PFI was replaced with Group-Hold-Out.

3.5.2. *Channel-wise feature effect*

It is also of interest to learn which individual feature maps (raster channels) are influential for FogNet. Because of potentially high correlation within the groups, we are interested to find out if the entire group is used or if a small subset of channels dominate. Here, we want to understand FogNet at a more granular level: to know how FogNet works, even when that hurts the performance. To do so, we use a feature effect measurement (SHAP) instead of feature importance. However, SHAP values are calculated locally, for a single data sample. To see what channels are used overall, we present a strategy to aggregate SHAP values to rank channels globally.

One option is to use SHAP directly on the channels, similar to our approach for LossSHAP on grouped channels. However, we are most interested in channels that have spatial locations with strong effect on prediction. Thus, we apply SHAP to superpixels within each channel, then rank channels based on the summed absolute SHAP values of those superpixels. Otherwise, the positive and negative contributions could cancel out if calculating SHAP directly on the channels.

3.5.2.1Channel-wise PartitionSHAP (CwPS). SHAP values were calculated using Channel-wise PartitionSHAP (CwPS), our modification of PartitionSHAP by Hamilton et al. (2021). PartitionShap can be used to explain image-based models. An image is recursively divided along the rows and columns to generate a partition tree. Given a user-supplied maximum number of evaluations, SHAP values are calculated for the superpixels defined by the partitions. The SHAP values are based on evaluating the prediction with and without removing pixels. Like SHAP, a superpixel's contribution is not based just on masking that superpixel but on a number of evaluations that include other superpixels to take into account dependencies. Since the superpixels are based on the hierarchical partitions, the number of evaluations controls the size of the final superpixels (explanation granularity) and the amount of computation time required.

The output of PartitionShap is a heatmap overlaid on the input image. This makes sense for RGB images, for example by showing that superpixel with a bird's eye was important to the classification of an *egret*. But here, a spatial explanation does not reveal the influence of the 384 channels variables. To explain the important of the FogNet raster channels, CwPS partitions along the channels before partitioning along the rows and columns. More details about CwPS are available on our GitHub repository Krell (2021).

Where SHAP replaces features with random values, PartitionShap has the choice of blurring kernels or replacement with a constant value. We experimented with 6: 3 blurring kernels and 3 constant values. The kernel sizes were 10 × 10, 20 × 20, and 32 × 32, and the constant values were 0.0, 0.5, and 1.0 since FogNet data is normalized. Plotting the SHAP values showed that the blurring kernels produced inconsistent explanations, while the constant values were very consistent. Our hypothesis is that blurring, while useful to break up the edges that are the basis of typical image classification, are less effective for variables such as sea surface temperature. A blurred SST may be very similar to the original, and not sufficiently removing the original information. Thus, the constant value of 0.5 was chosen due to its consistency.

CwPS was performed on a set of data samples, using both test and validation because of the low number of fog instances. The set included all 67 hits, 64 misses, 78 false alarms, and 84 randomly selected correct rejections. The maximum number of SHAP evaluations was set to 250,000. Fig. 9 shows an example of CwPS heatmap on the top three ranked channels for a randomly selected hit sample.

Fig. 10 shows the number of times that each channel occurs among the top and bottom 50 channels within each classification category. First, each instance's channels are ranked based on the summed absolute SHAP values as shown in Fig. 9. Then, the top and bottom counts are obtained by searching for occurrences of each channel within the top and bottom of the ordered channel lists. The number 50 was selected since it shows most of the channels from each of the groups for a comparison. It is also interesting to see the effect of varying the number of top channels from 1 to 384; this shows how some features consistently remain in the top and other are sluggish to leave the bottom bands plot. We present this as an animation in Fig. 11. To analyze the contribution of specific channels, Table 2 shows the ordered top 10 frequently occurring channels across the four classification categories.

An immediate observation is the very large effect of G5 channels compared to those of the other groups. According to CwPS, G5 channels consistently are among those with greatest effect on the prediction. However, G5 was not given the highest importance according to the group-based XAI methods. There appears to be a discrepancy, but we have three comments as to why this is not an unexpected XAI outcome.

First, it is important to keep in mind the difference between feature importance and effect as previously discussed. Fig. 8 shows the results of XAI based on the change in overall HSS. Fig. 10 shows the results of XAI based on the change in output for individual predictions, aggregated for a global model summary. Features may be used by the model without increasing performing, or even hurting performance. While CwPS suggests that G5 channels are heavily relied upon for the hits and correct rejects (which would increase performance), they are also used for misses and false alarms (lowering performance). Thus, it is reasonable that the high effect reported by CwPS would not be reflected in feature importance study shown in Fig. 8.

Second, even if CwPS were based on performance, it is not true that XAI at smaller levels of granularity sum to the equal to the output of XAI performed at a higher level. XAI techniques are highly susceptible to the feature grouping used Au et al. (2021). This can be illustrated with a simple 2D example: classification of a bird photograph. Given a robust model, permuting a single pixel in the bird's beak might have practically no effect on the model's ability to recognize the bird. But removing (permuting) all the beak pixels together might trigger a significant response, perhaps causing the bird to be mislabeled. Thus, simply summing the pixel-level XAI output does not provide the same explanation as the superpixel-level XAI. This example actually occurred when we applied PartitionSHAP to a CNN trained on ImageNet data.

In the case of FogNet, there is a physical interpretation for some of these groups that suggests that a discrepancy between XAI at the group and channel-level is not unexpected. For example, G3 represents the atmospheric profile where each channels are the variable at consecutive heights. Taking a single channel and evaluating superpixels within it might not be sufficient to break up the overall across-channel gradient pattern learned by the CNN. However, removing the entire group completely removes that pattern and triggers an appreciable influence on model performance.

In the case of G3 from a meteorological perspective, the 2-m dew point temperature (non-conservative surface moisture proxy) and 2-m relative humidity (percent of saturation) are the only features that directly contribute to fog formation; high relative humidity and moist environments are essential for fog development. However, there is no

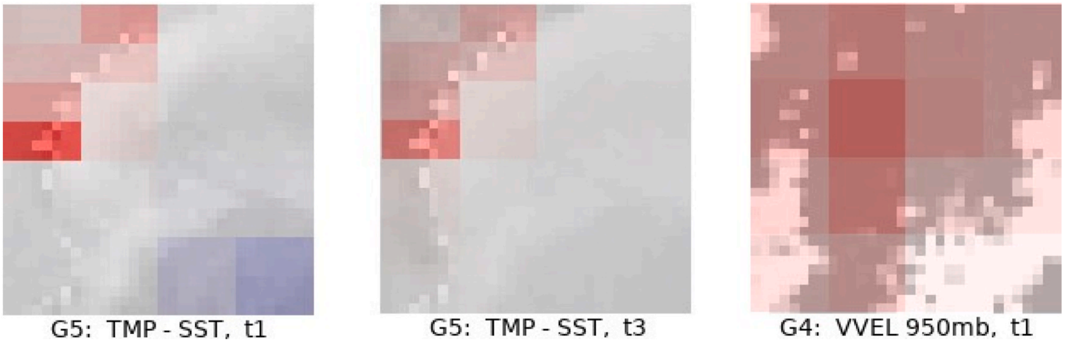
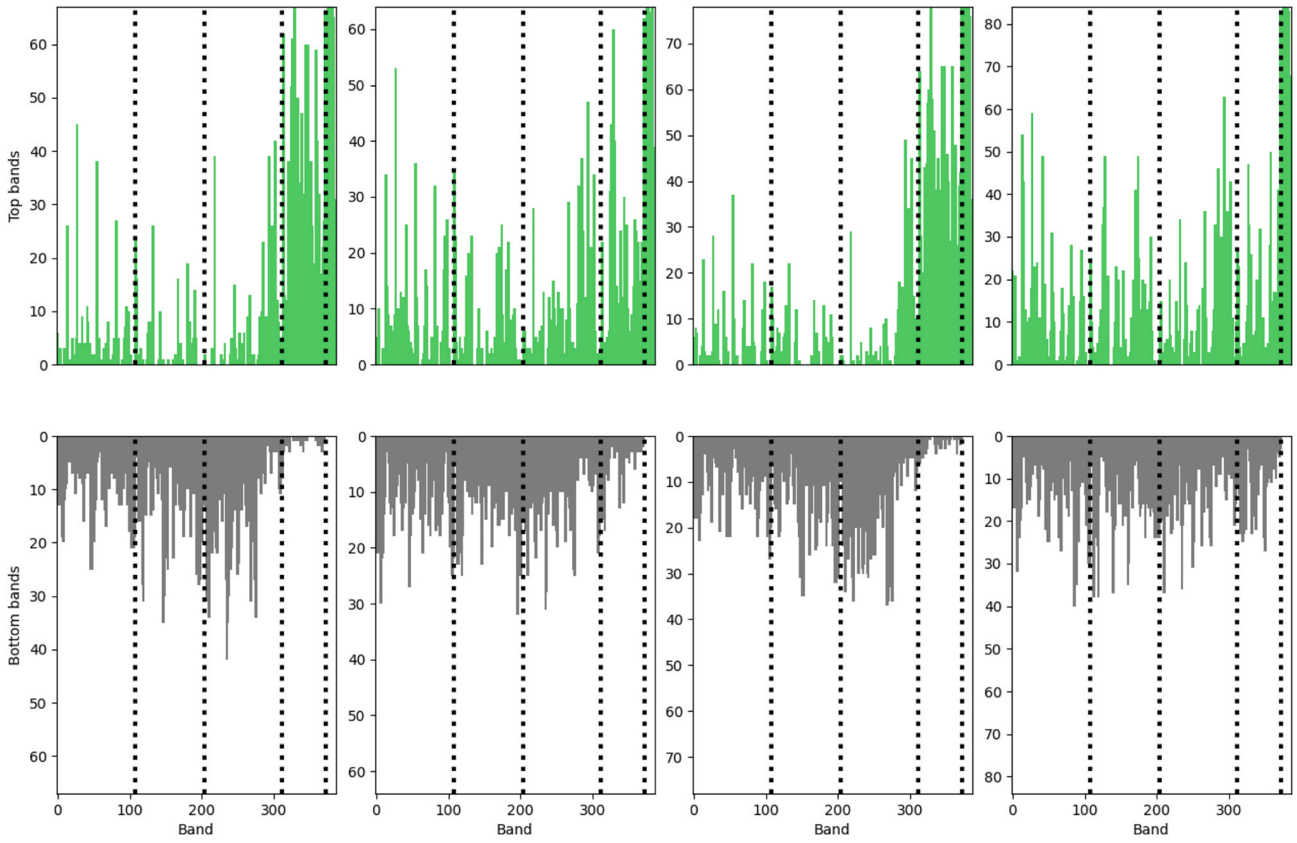


Fig. 9. Channel-wise PartitionShap results for a randomly selected *hit* case (advection fog). Three images represent heatmaps of the SHAP values for superpixels within the top three channels. The channels are ranked based on the sum of the absolute SHAP values. The red values indicate superpixels that contributed toward the prediction (fog) and blue away from the prediction.



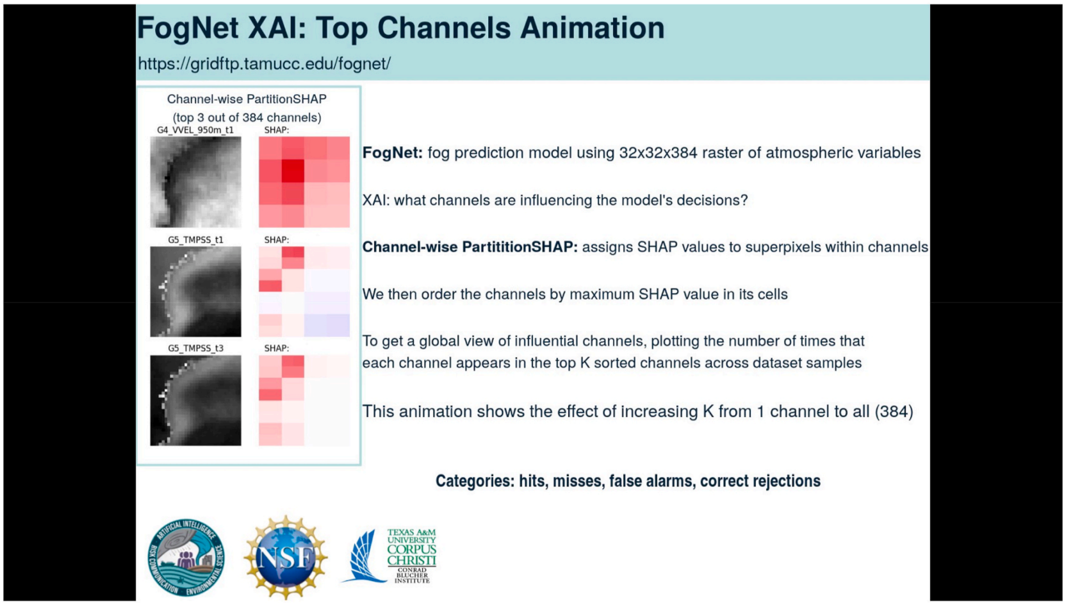
(a) Hits

(b) Misses

(c) False alarms

(d) Correct rejections

Fig. 10. Count of occurrences in top & bottom 50 channels when using Channel-wise PartitionSHAP. Dotted lines separate the 5 physical groups, from G1 on the left to G5 on the right.



FogNet XAI: Top Channels Animation

https://gridftp.tamucc.edu/fognet/

Channel-wise PartitionSHAP
(top 3 out of 384 channels)

FogNet: fog prediction model using 32x32x384 raster of atmospheric variables

XAI: what channels are influencing the model's decisions?

Channel-wise PartititionSHAP: assigns SHAP values to superpixels within channels

We then order the channels by maximum SHAP value in its cells

To get a global view of influential channels, plotting the number of times that each channel appears in the top K sorted channels across dataset samples

This animation shows the effect of increasing K from 1 channel to all (384)

Categories: hits, misses, false alarms, correct rejections

Fig. 11. Animation showing the occurrences of each channel in the top-K channels, as K is varied from 1, 2, …, 384 since there are 384 channels total.

direct relationship between the other features *individually* and fog development. However, *collectively*, all features in G3 are important; the profiles of temperature and relative humidity are strongly related to fog development. For example, radiation fog over south Texas typically requires a thin layer of moist/saturated conditions near the surface followed by significantly drier aloft, and is associated with an inversion (temperature increase with height) in the lower levels. The air temperature at 2 m and at each of the isobaric levels from 975 mb to 700 mb, are *individually* not related to fog. However, these temperature features are *collectively* related to fog since their combination determines

Table 2

Top channels based on Channel-wise PartitionSHAP. The most frequent feature (raster channel) selected at each rank for classification categories. That is, if channel *x* is at rank 1, then *x* was most commonly selected as the top channel in the ordered channels for each input case. The ordering is based on the highest absolute SHAP values that occur within each band. The Name is of the form *feature_level_time*, where *feature* is the symbol describing the feature as defined in Kamangir et al. (2021), *level* is the atmospheric pressure level (in mb) or height level (in m) corresponding to the feature, and *time* is the NAM prediction hour. The features are either output by or derived from a NWP model. The exception is TMP-SST that relies on both the NWP output and Multiscale Ultra High-resolution (MUR) satellite product. The features that appear here are vertical velocity (VV), specific humidity (Q), the difference between the 2-m air temperature and the MUR sea surface temperature (TMP-SST), and the difference between the 2-m air temperature and the 2-m dew point temperature (TMP-DPT).

(a) Hits

Rank	Most Common Feature		
	Band	Name	Group
1	329	VV 950 mb, t_1	G4
2	377	TMP - SST, t_1	G5
3	379	TMP - SST, t_3	G5
4	375	TMP - DPT, t_3	G5
5	375	TMP - DPT, t_3	G5
6	376	TMP - SST, t_0	G5
7	383	DPT - SST, t_3	G5
8	372	TMP - DPT, t_0	G5
9	372	TMP - DPT, t_0	G5
10	314	Q 2 mb, t_2	G4

(b) Misses

Rank	Most Common Feature		
	Band	Name	Group
1	377	TMP - SST, t_1	G5
2	377	TMP - SST, t_1	G5
3	375	TMP - DPT, t_3	G5
4	375	TMP - DPT, t_3	G5
5	378	TMP - SST, t_2	G5
6	383	DPT - SST, t_3	G5
7	380	DPT - SST, t_0	G5
8	373	TMP - DPT, t_1	G5
9	372	TMP - DPT, t_0	G5
10	372	TMP - DPT, t_0	G5

(c) False alarms

Rank	Most Common Feature		
	Band	Name	Group
1	329	VV 950 mb, t_1	G4
2	377	TMP - SST, t_1	G5
3	379	TMP - SST, t_3	G5
4	375	TMP - DPT, t_3	G5
5	375	TMP - DPT, t_3	G5
6	378	TMP - SST, t_2	G5
7	376	TMP - SST, t_0	G5
8	378	TMP - SST, t_2	G5
9	372	TMP - DPT, t_0	G5
10	383	DPT - SST, t_3	G5

(d) Correct rejections

Rank	Most Common Feature		
	Band	Name	Group
1	377	TMP - SST, t_1	G5
2	375	TMP - DPT, t_3	G5
3	375	TMP - DPT, t_3	G5
4	379	TMP - SST, t_3	G5
5	382	DPT - SST, t_2	G5
6	383	DPT - SST, t_3	G5
7	381	DPT - SST, t_1	G5
8	373	TMP - DPT, t_1	G5
9	380	DPT - SST, t_0	G5
10	372	TMP - DPT, t_0	G5

whether a temperature inversion exists, which is related to fog. Further, advection fog events (moist air moves over a cooler surface resulting in condensation) affecting the target in this study (KRAS) are also associated with a lower level temperature inversion, yet generally associated with a deeper moist layer than associated with radiation fog. The features in group 2 in Table 2 (a) are Q (specific humidity), which individually is strongly related to fog. Q is defined as the mass of water vapor to the total mass of air and is thus a good measure of moisture content, and sufficient moisture is required for fog development. More often that not, after advection fog develops, a persistent supply of moisture is necessary for fog maintenance (Yang et al., 2018). The features TMP-SST and DPT-SST in group 5 are individually important to advection fog development at the target (KRAS) during the October–April period; advection fog at KRAS typically occurs when moist onshore flow moves across the cooler shelf waters near the coast and maintains a temperature inversion and moist marine layer near the surface; the layer eventually condenses, resulting in fog formation. . This scenario requires the temperature (TMP) or dew point temperature (DPT) of the air near the surface (10 m) to exceed the sea surface temperature (SST).

Finally, when interpreting the relative influence of the groups in Fig. 8, it is important to keep in mind the sizes of the groups. While G3 and G4 are suggested to be more important than G5, the latter group has only 12 channels while G3 has 108 and G4 has 60. This could actually highlight G5 importance since it has an appreciable impact on loss despite the group having so few members.

Note from Fig. 10 that the number of G5 channel counts exceeded that of all other channels for each of the 4 locations on the confusion matrix. This is likely due to the following: the vast majority of fog events in the training and validation data sets were of the advection type. Further, the G5 features were included specifically to capture advection fog cases. Therefore, we conjecture that during training, FogNet learned the strong relationship between advection fog and G5 features (especially TMP-SST and DPT-SST). Thus, during every testing set instance, FogNet would nearly always use G5 channels when making a prediction. A similar argument can be made with respect to the G4 channels for hits and false alarms; note that the G4 channel counts for positive fog predictions (prediction of a fog event) far exceeded that of the G1, G2, and G3 counts, which illustrates FogNet's propensity to use the G4 channels, a behavior learned by this CNN during the training process. Table 2 complements Fig. 10 by specifying the top 10 channels used by FogNet (based on CwPS) for each confusion matrix scenario. The top 10 channels are only the following 5 different features at different NAM prediction hours: vertical velocity at 950 mb (VV 950), specific humidity (Q), the difference between the 2-m air temperatures and the MUR SST (TMP-SST), the difference between the 2-m dew point temperature and the MUR SST (DPT-SST), and the difference between the 2-m air and dew point temperatures, otherwise known as the dew point depression (TMP-DPT). All of these channels are in G4 or G5. The VV 950, Q, and TMP-DPT are related to all fog types. VV 950 is the most frequent channel used by FogNet in connection with positive fog predictions (hits and false alarms). This channel is included in the microphysics group given its relationship to the activation of cloud condensation nuclei. However, it is likely that the frequent use of VV 950 by FogNet is related to the strong dynamical, rather than microphysical, relationship to fog. In particular, fog is generally associated with small vertical velocity magnitudes. Furthermore, advection fog is associated with negative vertical velocities (Koračin et al., 2014; Huang et al., 2015). The DPT-SST and TMP-SST features are critically important to advection fog and only to advection fog. The condition DPT-SST ≥ 0 or TMP-SST ≥ 0 must be met for advection fog to develop. Again, the preponderance of advection fog cases in the training and validation data set, and the requirement that the foregoing condition be met for advection fog formation, suggests that FogNet learned, wrongly, that fog in general required that the 2-m dew point or air temperature approach or exceed the SST, rather than learning that TMP-SST and DPT-SST are important only to advection fog.

Based on FogNet's apparent overreliance on TMP-SST and DPT-SST to predict fog, the features used in all 4 positions on the confusion matrix in Table 2 can be explained as follows: With respect to hits, all 10 bands were the primary ones used to make the positive fog predictions, the net effect of the magnitudes of these bands prompted FogNet to render a prediction that fog will occur, and the vast majority of these cases were advection fogs. This is supported by the fact that FogNet demonstrated skill in predicting the advection fog events, yet the ability to predict radiation fog was poor (Kamangir et al., 2021). For the false alarms, the net effect of the magnitudes of the top 10 bands prompted FogNet to predict that fog would occur, yet was primarily motivated by DPT-SST ≥ 0 and/or TMP-SST ≥ 0, the condition for advection fog. However, features within G1, G2, or G3 were probably not conducive to advection fog. For example, the G2 features may not have been conducive to advection fog (decrease in Q with height combined with significant TKE would likely preclude fog per Toth et al. (2010)). In another example, the wind speeds (G1 feature) could be $< 2.5\ ms^{-1}$ and thus no advection fog due to insufficient advection (Tardif and Rasmussen, 2007). The missed events can be explained by the likelihood that the DPT-SST ≥ 0 and/or TMP-SST ≥ 0 conditions were not met, and thus given FogNet's strong reliance on this condition for fog, a negative fog prediction was made. However, G1, G2, and/or G3 features likely favored fog types other than advection fog. Examples include wind, TKE, and thermodynamic profiles favoring radiation or cloud base lowering fogs. (Croft et al., 1997; Dupont et al., 2016). For the correct rejections, there were likely many testing set cases where neither DPT-SST ≥ 0 nor TMP-SST ≥ 0 were met, and either G1, G2, and/or G3 conditions were not met for any other fog type, and thus FogNet rightly predicted no fog. The difficulties of the model to predict radiation cases is due to the small number of such cases in the data set, 23 radiation fog cases as compared to 183 advection fog cases. Although it is doubtful that changes in FogNet3D architecture would resolve this challenge, the introduction of new features, such as mean sea level pressure (MSLP), including TMP, u, v, TKE, RH, and VVEL at 1000-mb, and surface sensible and latent heat fluxes, may improve radiation fog prediction. Increasing the number of radiation fog cases, either by including additional sites beyond KRAS (with a greater percentage of radiation fog cases), and/or performing data augmentation, should also improve radiation fog prediction.

4. Conclusions

In this work, we present evidence showing the importance of using 3D convolutional neural networks for improving the accuracy of predictions for a complex atmospheric process, the occurrence of coastal fog. We show that the 3D convolutions are important to capture the 3D structure of the lower atmosphere (including the PBL), and to learn the nonlinear complexity of the relationship between the state of the atmosphere and the formation of fog. Given the 3D nature of extreme atmospheric events, the knowledge obtained from this study may apply more generally.

We also present several strategies to increase performance when using 3D CNNs. Parallel spatial and variable-wise feature learning showed a good improvement in fog forecasting versus sequential spatio-variable feature learning using 3D kernels. Given a large number of input variables, the grouping of features, based on their similar relationship to fog development, helps to generate more distinguishable features for the classifier and thereby improving the performance of forecasting. Due to the 3D structure within the data, a unique ordering of features helps the 3D convolutions discover important feature representations (such as vertical profiles of temperature, moisture, and wind). Also, given the complexity of meteorological forecasting, a simple CNN-based model is not sufficient for efficiently learning the relationships between inputs and targets. We introduced and evaluated several techniques that improve the CNN performance including dense block, attention mechanism, and multiscale feature learning.

Applying XAI to study the influence of the features to FogNet revealed challenges in interpreting seemingly conflicting results. The lack of a guarantee as to which of the many methods is giving the most accurate explanation makes analysis challenging. Still, we argue that overall consistencies among methods give confidence to some interpretations. It is encouraging to observe that all the groups are contributing to an increase in model performance. A salient observation is the strong effect of all G5 channels on the output. We see that G5 is relied upon strongly for correct and incorrect predictions; there are patterns learned that hold for the fog but are not rich enough to completely disambiguate fog and non-fog. Other groups contain channels that have a higher effect for misses or false alarms, making them candidates for removal for potential performance improvements. Alternatively, one could argue that given FogNet's apparent overreliance on features that were designed solely to capture advection fog, and that have no significant relationship to the other fog types (TMP-SST and DPT-SST), FogNet is thus more trustworthy with respect to the prediction of advection fog than with respect to radiation and advection-radiation fog. Thus, a solution is to increase the number of fog types other than advection fog in the training and validation data sets to allow the CNN to learn the relationships between the features and these other fog types. However, it will be challenging for FogNet to sufficiently learn the relationships for all fog types. According to Gultepe et al. (2007), it is very difficult to provide accurate numerical fog predictions when the predominate fog generation process is other than dynamical; this suggests that it would be easier to predict advection fog rather than radiation fog. It is clear that XAI techniques are sensitive to how features are grouped, and certain groups may provide explanations that are more useful to the modeler. A challenge is how to best partition the large raster input into meaningful feature groups. Here, domain knowledge was applied to perform XAI on the 5 metocean groups. But this may or may not be the optimal grouping. To achieve a grouping closer to optimal, it may be necessary to increase the number of groups to at least 6 by including horizontal wind, vertical velocity, and turbulence profiles as G1, retaining Q as the sole feature in G2 (Q should increase with height to support radiation fog development), retain the features in G3 (identify thermodynamic profiles conducive to radiation fog), retain VIS as the sole feature in G4, use G5 to account for surface processes, which retains the current G5 features, yet adds new ones such as sensible and latent heat fluxes (that modulates fog development), and add G6 to account for the surface synoptic condition by including MSLP (to acccount for the relationship between radiation fog and the proximity of the surface anticyclone (Northern Hemisphere) per (Meyer and Lala, 1990) and 2-m equivalent potential temperature (to better account for frontal fogs). We are currently investigating the role of spatial statistics to select the groups in a data-driven fashion, based on feature correlations and interactions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This material is based upon work supported by the National Science Foundation under awards 2019758 and 1828380.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2022.105424.

References

Au, Q., Herbinger, J., Stachl, C., Bischl, B., Casalicchio, G., 2021. Grouped Feature Importance and Combined Features Effect Plot arXiv preprint arXiv:2104.11688.

Baker, R., Cramer, J., Peters, J., 2002. Radiation fog: ups airlines conceptual models and forecast methods. In: Proc. 10th Conf. On Aviation, Range, and Aerospace Meteorology. American Meteorological Society, pp. 154–159.

Bosma, S., Nazari, N., 2021. Estimating california's Solar and Wind Energy Production Using Computer Vision Deep Learning Techniques on Weather Images, 08727 arXiv: 2103.

Castro, R., Souto, Y.M., Ogasawara, E., Porto, F., Bezerra, E., 2021. Stconvs2s: spatiotemporal convolutional sequence to sequence network for weather forecasting. Neurocomputing 426, 285–298.

Chollet, F., et al., 2018. Deep Learning with Python, vol. 361. Manning, New York.

Croft, P.J., Pfost, R.L., Medlin, J.M., Johnson, G.A., 1997. Fog forecasting for the southern region: a conceptual model approach. Weather Forecast. 12, 545–556.

Das, S., Brimley, B.K., Lindheimer, T.E., Zupancich, M., 2018. Association of reduced visibility with crash outcomes. IATSS Res. 42, 143–151.

Dupont, J., Haeffelin, M., Protat, A., Bouniol, D., 2012. Stratus–fog formation and dissipation: a 6-day case study. Boundary-Layer Meteorol. 143, 207–225.

Dupont, J.C., Haeffelin, M., Stolaki, S., Elias, T., 2016. Analysis of dynamical and thermal processes driving fog and quasi-fog life cycles using the 2010-2013 parisfog dataset. Pure Appl. Geophys. 173, 1337–1358.

FAA, 2017. Advisory Circular. Subject: Automated Weather Observing Systems (Awos) for Non-federal Applications. https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_150_5220-16E.pdf.

Fryer, D., Strümke, I., Nguyen, H., 2021. Shapley Values for Feature Selection: the Good, the Bad, and the Axioms arXiv preprint arXiv:2102.10936.

Gadiraju, K.K., Vatsavai, R.R., 2020. Comparative analysis of deep transfer learning performance on crop classification. In: Proceedings of the 9th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, pp. 1–8.

Glickman, T.S., 2000. Glossary of Meteorology. American Meteorological Society.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT press.

Gultepe, I., Milbrandt, J.A., Zhou, B., 2017. Marine fog: a review on microphysics and visibility prediction. In: Marine Fog: Challenges and Advancements in Observations, Modeling, and Forecasting. Springer, pp. 345–394.

Gultepe, I., Sharman, R., Williams, P., et al., 2019. A review of high impact weather for aviation meteorology. Pure Appl. Geophys. 176, 1869–1921.

Gultepe, I., Tardif, R., Michaelides, S., Cermak, J., Bott, A., Bendix, J., Müller, M.D., Pagowski, M., Hansen, B., Ellrod, G., et al., 2007. Fog research: a review of past achievements and future perspectives. Pure Appl. Geophys. 164, 1121–1159.

Hamilton, M., Lundberg, S., Zhang, L., Fu, S., Freeman, W.T., 2021. Model-agnostic Explainability for Visual Search arXiv preprint arXiv:2103.00370.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition arXiv preprint arXiv:1512.03385.

He, M., Li, B., Chen, H., 2017. Multi-scale 3d deep convolutional neural network for hyperspectral image classification. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 3904–3908.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., Weinberger, K., 2019. Convolutional networks with dense connectivity. In: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Huang, H., Liu, H., Huang, J., Mao, W., Bi, X., 2015. Atmospheric boundary layer structure and turbulence during sea fog on the southern China coast. Mon. Weather Rev. 143, 1907–1923.

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley & Sons Ltd.

Kamangir, H., Collins, W., Tissot, P., King, S.A., Dinh, H.T.H., Durham, N., Rizzo, J., 2021. Fognet: A Multiscale 3d Cnn with Double-Branch Dense Block and Attention Mechanism for Fog Prediction. Machine Learning with Applications, 100038.

Kanopoulos, N., Vasanthavada, N., Baker, R.L., 1988. Design of an image edge detection filter using the sobel operator. IEEE J. Solid State Circ. 23, 358–367.

Koračin, D., Dorman, C.E., Lewis, J.M., Hudson, J.G., Wilcox, E.M., Torregrosa, A., 2014. Marine fog: a review. Atmos. Res. 143, 142–175.

Krell, E., 2021. Partitionshap Multiband Demo. https://github.com/conrad-blucher-institute/partitionshap-multiband-demo.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Kumler-Bonfanti, C., Stewart, J., Hall, D., Govett, M., 2020. Tropical and extratropical cyclone detection using deep learning. J. Appl. Meteorol. Climatol. 59, 1971–1985.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R., 2019. Unmasking clever hans predictors and assessing what machines really learn. Nat. Commun. 10, 1–8.

Li, Y., Zhang, H., Shen, Q., 2017. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. Rem. Sens. 9, 67.

Liu, D., Yang, J., Niu, S., Li, Z., 2011. On the evolution and structure of a radiation fog event in nanjing. Adv. Atmos. Sci. 28, 223–237.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2019. Explainable Ai for Trees: from Local Explanations to Global Understanding arXiv preprint arXiv:1905.04610.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777.

Ma, W., Yang, Q., Wu, Y., Zhao, W., Zhang, X., 2019. Double-branch multi-attention mechanism network for hyperspectral image classification. Rem. Sens. 11, 1307.

Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I., 2022. Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience arXiv preprint arXiv:2202.03407.

McGovern, A., Lagerquist, R., Gagne, D.J., Jergensen, G.E., Elmore, K.L., Homeyer, C.R., Smith, T., 2019. Making the black box more transparent: understanding the physical implications of machine learning. Bull. Am. Meteorol. Soc. 100, 2175–2199.

Messalas, A., Kanellopoulos, Y., Makris, C., 2019. Model-agnostic interpretability with shapley values. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE, pp. 1–7.

Meyer, M.B., Lala, G.G., 1990. Climatological aspects of radiation fog occurrence at albany, New York. J. Clim. 3, 577–586.

Molnar, C., 2020. Interpretable Machine Learning. Lulu. com.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B., 2020. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models arXiv preprint arXiv: 2007.04131.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. Unit. States Am. 116, 22071–22080.

Niu, D., Diao, L., Xu, L., Zang, Z., Chen, X., Liang, S., 2020. Precipitation forecast based on multi-channel convlstm and 3d-cnn. In: 2020 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE, pp. 367–371.

Oliver, D.A., Lewellen, W.S., Williamson, G.G., 1978. The interaction between turbulent and radiative transport in the development of fog and low-level stratus. J. Atmos. Sci. 35, 301–316.

Orlanski, I., 1975. A rational subdivision of scales for atmospheric processes. Bull. Am. Meteorol. Soc. 56, 527–530.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Adv. Neural Inf. Process. Syst. 32, 8024–8035. Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Price, JD, 2019. On the formation and development of radiation fog: An observational study. Boundary-Layer Meteorol. 172 (2), 167–197. In this issue.

Rasp, S., Thuerey, N., 2021. Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: a new model for weatherbench. J. Adv. Model. Earth Syst. 13, e2020MS002405.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv:1409.1556.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

sshuair, 2020. Torchsat. https://github.com/sshuair/torchsat.

Stensrud, D.J., 2009. Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models. Cambridge University Press.

Stull, R., 1988. An Introduction to Boundary Layer Meteorology. kluwer academic publishers, p. 666.

Tardif, R., Rasmussen, R.M., 2007. Event-based climatology and typology of fog in the New York city region. J. Appl. Meteorol. Climatol. 46, 1141–1168.

Toth, G., Gultepe, I., Milbrandt, J., Hansen, B., Pearson, G., Fogarty, C., Burrows, W., 2010. The Environment canada Handbook on Fog and Fog Forecasting. Environment Canada.

Twomey, S., 1974. Pollution and the planetary albedo. Atmos. Environ. 8, 1251–1256, 1967.

Vaisala, 2004. Present weather detector pwd22. https://www.manualslib.com/manual/1226525/Vaisala-Pwd22.html.

Vaisala, 2015. Vaisala automated weather observing system aw20. https://manualzz.com/doc/11369536/vaisala-automated-weather-observing-system-aw20-faa.

Vaisala, 2018. Present weather and visibility sensors pwd10, pwd12, pwd20, and pwd22. https://www.vaisala.com/sites/default/files/documents/PWD-Series-Datasheet-B210385EN.pdf.

Wallace, J.M., Hobbs, P.V., 1977. Atmospheric Science: an Introductory Survey. Academic.

Wang, C., Richard, E., Pedeboy, S., 2019. Exploiting Deep Learning Algorithms in Forecasting the Occurrence of Intense Lightning Activities. AGU Fall Meeting Abstracts, pp. A33M–A2967.

Wang, C., Wang, P., Wang, P., Xue, B., Wang, D., 2021. A spatiotemporal attention model for severe precipitation estimation. IEEE Geosci. Rem. Sens. Lett. 19, 1–5.

Wang, X., Wang, W., Yan, B., 2020. Tropical cyclone intensity change prediction based on surrounding environmental conditions with deep learning. Water 12, 2685.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Academic Press.

WMO, 2020. Fog Compared with Mist. https://cloudatlas.wmo.int/en/fog-compared-with-mist.html.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057.

Yang, L., Liu, J.-W., Ren, Z.-P., Xie, S.-P., Zhang, S.-P., Gao, S.-H., 2018. Atmospheric conditions for advection-radiation fog over the western Yellow Sea. J. Geophys. Res. 123, 5455–5468. https://doi.org/10.1029/2017jd028088.

Yu, F., Koltun, V., 2015. Multi-scale Context Aggregation by Dilated Convolutions arXiv preprint arXiv:1511.07122.

Zanchetta, A., Zecchetto, S., 2021. Wind direction retrieval from sentinel-1 sar images using resnet. Remote Sens. Environ. 253, 112178.