A Novel Method for Improving Seasonal Atlantic Hurricane Activity Prediction Based on Convolutional Neural Networks and Visualization Maps

ANTONIA COMANICIU[a] AND HIROYUKI MURAKAMI[b]

[a] *Princeton University, Princeton, New Jersey*
[b] *NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey*

(Manuscript received 2 February 2025, in final form 26 October 2025, accepted 28 November 2025)

ABSTRACT: In this study, we propose a method to predict seasonal Atlantic hurricane activity by utilizing generated maps that depict the spatial distribution of two key predictors: sea surface temperature (SST) and outgoing longwave radiation (OLR). Our approach employs a convolutional neural network (CNN) system that processes SST and OLR maps as input images to predict the accumulated cyclone energy (ACE) index, which is classified into three activity levels: high, medium, and low. Using these two different sets of input data, we developed two separate models based on our CNN architecture. The first model provides an early season prediction, 4 months in advance, using SST data from January to identify high-activity seasons. The second model delivers high forecast skill for three-class prediction closer to the season start, using July OLR data. Both models offer higher accuracy compared to past seasonal predictions issued in May and August.

KEYWORDS: Seasonal forecasting; Artificial intelligence; Neural networks

1. Introduction

Atlantic hurricanes severely disrupt people's lives every year. The regions affected by hurricanes may suffer extensive damage from destructive winds and flooding, resulting in loss of life and billions of dollars of property damage. In fact, hurricanes have caused over 6890 deaths and 1.3 trillion U.S. dollars in damage since 1980 (NOAA 2025). Equipping government officials and townships with accurate seasonal hurricane activity predictions well in advance can help them better prepare for an effective response.

Operational seasonal Atlantic hurricane predictions are currently released by numerous forecasting centers (Caron et al. 2020). We focus on predictions issued by NOAA (NOAA 2023a) and the Colorado State University (CSU 2022), which are widely recognized for their established methodologies and long-standing contributions to hurricane prediction. They provide predictions multiple times per year and use the accumulated cyclone energy (ACE) index (Bell et al. 2000), which is defined as the sum of the square of the maximum surface wind velocity throughout the lifetime of a tropical cyclone (TC), to characterize the upcoming hurricane season's activity. These predictions are based on a combination of statistical and dynamical models (Takaya et al. 2023).

CSU provides statistical predictions for ACE based on the historical skills of different forecast models, with outputs fit to a Weibull distribution (Klotzbach et al. 2023). As input, these predictions use a variety of large-scale oceanic and atmospheric parameters spatially distributed across the globe, such as metrics that capture El Niño–Southern Oscillation (ENSO), Sahel precipitation and Saharan dust, and Atlantic Ocean thermodynamics. It is worth noting that, within Atlantic Ocean thermodynamics, an important emphasis is placed on relative sea surface temperatures (SSTs), which are determined as the difference between the tropical North Atlantic (10°–25°N, 80°–20°W) SST and SST across the tropical region (30°S–30°N latitude) (CSU 2024).

Similarly, for its seasonal hurricane activity prediction, NOAA assesses many atmospheric and oceanic parameters, drawing insights from both dynamic geophysical model outputs and expert judgment from human forecasters. These parameters typically encompass SST, relative humidity, low-level relative vorticity, and vertical wind shear in the tropical Atlantic region, as well as remote phenomena outside of the tropical Atlantic, such as ENSO and the West African monsoon (Karnauskas and Li 2016; NOAA CPC 2023). NOAA also considers the Atlantic multidecadal oscillation (AMO), which is an SST pattern that increases the likelihood of a high-activity season in its cold phase and of a low-activity season in its warm phase (NOAA CPC 2025).

Both CSU and NOAA's methodologies combine tropical Atlantic parameters with global-scale inputs, suggesting that the spatial distribution of SST across both the Atlantic and Pacific Ocean basins plays a critical role in enhancing hurricane forecast skill.

Seasonal activity forecasting tends to focus on predicting exact ACE values, which is a challenging task. However, NOAA predicts ranges for ACE and releases forecasts that categorize classes of activity for the upcoming hurricane season. According to NOAA, a season is classified as high activity if ACE $> 111 \times 10^4$ kt^2 (1 kt ≈ 0.51 m s^{-1}) and low activity if ACE $< 66 \times 10^4$ kt^2. For ACE values in between, the season is considered medium (near-normal) activity (CPC NOAA 2022). The determined ACE values represent the cumulative ACE index for the entire season, which spans from June to November. NOAA releases operational seasonal forecasts for the Atlantic hurricane season's activity in May (early prediction) and August (updated prediction). For the period 2004–23, NOAA's three-class prediction accuracy was about 50% (NOAA 2023a). Similarly, computing an equivalent three-class

Corresponding author: Antonia Comaniciu, acomaniciu25@gmail.com

DOI: 10.1175/AIES-D-25-0010.1 e250010

characterization of CSU ACE predictions (using NOAA's classification) for the period 2007–22, for which ACE predictions are published, yields an accuracy of 56.25%. While these predictions demonstrate skill, room for improvement remains.

In this paper, we propose a seasonal prediction focused on activity classes for ACE aligned with NOAA's seasonal activity classification. Similar to NOAA, we introduce both an early prediction (issued a couple of months before the hurricane season starts) and an updated prediction released midseason in August. Our early prediction aims to provide an extended lead time alert for seasons with heightened hurricane activity by using a binary categorization system (high vs medium–low activity). For midseason predictions, we propose a refined approach, classifying hurricane season activity into three distinct categories: high, medium, or low. NOAA estimates the mean number of named storms and hurricanes that is expected in correlation with the predicted activity class (above normal, near normal, and below normal) of each hurricane season, which are defined based on the ACE index (CPC NOAA 2022). Thus, classifying into three classes allows us to leverage NOAA's definition and provide a more detailed outlook of the upcoming hurricane season.

Our proposed method classifies hurricane season activity using convolutional neural networks (CNNs), which associate nonlinear, complex patterns in input images with ACE classes. For our early prediction, we use SST maps as input, demonstrating that accurate prediction can be obtained as early as February using SST images for the month of January. Preliminary results on this predictor were reported in Comaniciu and Murakami (2022). However, in this paper, we reconducted all experiments using raw data, which were processed and saved as grayscale images using Python libraries. This adjustment enhances the reproducibility of our results. In addition, to improve performance, in this paper, we propose a majority voting classifier that aggregates outputs from multiple independently trained models. This ensemble approach mitigates errors from outlier models and increases overall predictive accuracy, and it has been shown to be effective in improving model robustness (Opitz and Maclin 1999). Our choice of January SST data for our early prediction method is inspired by the work of Villarini and Vecchi (2013), who implemented a hybrid statistical–dynamical seasonal forecasting system that leveraged tropical Atlantic and tropical-mean SSTs to predict the ACE values for the upcoming hurricane season. Their results indicated that prediction accuracy was highest when using January SST data, achieving a correlation of 0.65 with hurricane activity.

For our updated midseason predictor, we use July outgoing longwave radiation (OLR) maps as input to our CNN architecture, classifying seasonal hurricane activity into three ACE classes: high, medium, and low. As with our early predictor, we conducted experiments using raw OLR data, which were processed and saved as grayscale images using Python libraries. Our choice of July OLR data is motivated by the work of Karnauskas and Li (2016), who demonstrated the utility of July OLR data for Atlantic hurricane prediction. Their study employed the meridional gradient of July OLR data across Africa in a logistic-regression-based predictor to estimate the number of hurricanes and storms in the upcoming season, achieving a high accuracy of 87%. However, their approach did not outperform NOAA's predictions when estimating hurricane season activity based on the ACE index. The meridional OLR gradient is used as a proxy for deep convective activity, which is closely tied to the West African monsoon rainfall pattern (NCAR 2022). The West African monsoon system includes features such as the African easterly jet and African easterly waves, which are seed disturbances for the development of Atlantic TCs (Alaka and Maloney 2017). Thus, variations in the OLR gradient over Africa reflect environmental conditions favorable for TCs.

CNNs have been successfully applied to hurricane prediction, such as activity based on satellite imagery and other data sources (Chen et al. 2018; Tan et al. 2022; Pradhan et al. 2018; Xu et al. 2023), hurricane track (Lian et al. 2020), and rapid intensification (Griffin et al. 2022). Wang and Li (2023) proposed deep learning models for TC intensity and wind radius estimation. McNeely et al. (2023) use a combination of deep autoregressive generative models and CNNs to predict short-term TC intensity changes for 6- and 12-h intervals. Prior work to explore the use of CNNs for seasonal ACE prediction includes that of Asthana et al. (2021) and Fu et al. (2023). Asthana et al. (2021) proposed a fused CNN model that processes different data modalities, such as SST, sea level pressure, and wind, to predict ACE values for an entire hurricane season. Despite the complexity of this model, it only achieved comparable accuracy with the current state of the art, albeit with longer lead times. In addition, the complexity of the model was not conducive to conclusions regarding the influences of each modality on the prediction accuracy. The work of Fu et al. (2023) proposed ensemble CNN models trained on a combination of environmental factors: SST anomalies, saturation deficit anomalies, 850-hPa relative vorticity anomalies, and vertical wind shear anomalies between 850 and 200 hPa. Their model successfully predicted cyclone frequency and showed high skill in forecasting the number of major TCs. Additionally, the model was able to capture the spatial distribution of cyclone activity in seven different TC basins. Fu et al. (2023) predicted ACE in the North Atlantic basin with accuracy ranging from 0.65 correlation (Pearson correlation coefficient) for 0-month lead time to 0.47 correlation for a 4-month lead time.

Our prediction method is based on a simple CNN architecture that exploits spatial information from SST and OLR maps and focuses exclusively on predicting ACE activity classes across North Atlantic basin. The simplicity of the model highlights that the spatial distributions of SST and OLR contain useful information correlated with upcoming seasonal hurricane activity. This simplicity also facilitates the interpretation of the results through saliency maps (McNeely et al. 2023; Wang and Li 2023), which identify regions of high importance for hurricane activity. The majority voting-based CNN early prediction is available 3 months ahead of NOAA's early prediction in May and demonstrates approximately 20% higher accuracy than NOAA's recomputed equivalent two-class classification for the period 2009–23. Similarly, our updated August prediction, which incorporates majority voting across 10 trained models, is approximately 30% more accurate than NOAA's three-class August prediction when compared with predictions over the

FIG. 1. January 2022 anomaly SST map example.

recent 20 years (2004–22). The remainder of this paper covers the data and methodology used in this study, our findings, and the conclusions we drew from those findings.

2. Data and methodology

a. Visualization input data

To enhance prediction accuracy, our approach leverages the spatial distribution information in both SST and OLR maps.

1) SST DATA

SST data were used for our early prediction. SST visualization images are generated based on the Extended Reconstructed SST (ERSST) dataset v5, which provides data from January 1854 to the present and is updated monthly (NOAA 2021a). This dataset spans latitudes from 88.75°S to 88.75°N in increments of 2° and longitudes from 0° to 358°E in increments of 2°. SST anomalies are calculated relative to a 1971–2000 climatology. For each year, January SST data were downloaded as a netCDF file (e.g., ersst.v5.201901.nc for January 2019 data). Data processing was performed using the netCDF Python library. SST anomaly values for each year were encoded as grayscale images with a single channel, using 255 gray levels. The encoding formula was

$$\text{gray level} = 255 \times \frac{\text{current SST values} - \text{min SST value}}{\text{max SST value} - \text{min SST value}}, \quad (1)$$

where max SST value and min SST value are the maximum and minimum SST anomaly values across all January months in the dataset, respectively. In the resulting grayscale images, white represents the highest SST anomalies, while dark gray or black represents the lowest values. The white continents do not represent SST data (Fig. 1).

SST images for the years 1856–2023 were encoded based on the dataset for the month of January, resulting in a dataset of 168 samples (images), each representing 1 year. The images had dimensions of 180 × 89 pixels. It is important to note that the ERSST v5 dataset is known to have more reliable data after the 1940s (NOAA 2021a).

As mentioned in the introduction, both the tropical region and the tropical Atlantic region are known to influence Atlantic TC activity. To better understand the contribution of prediction skill, we also generated an image dataset only corresponding to the tropical region. The tropical region images spanned latitudes from 30°S to 30°N and were saved as single-channel grayscale images with dimensions of 180 × 31 pixels (Fig. 2).

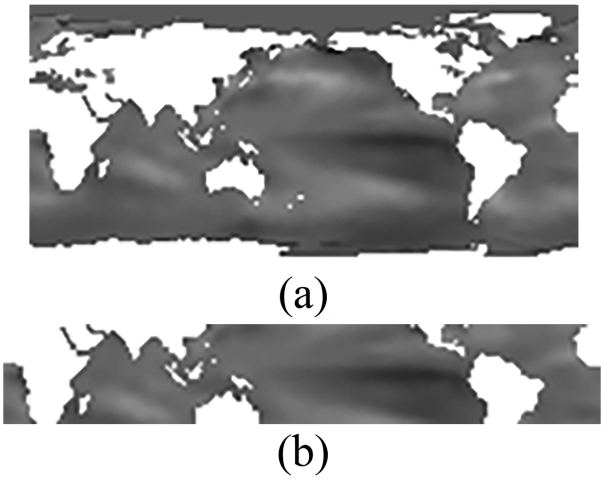

FIG. 2. Training image example for January 2022. (a) Full domain (88.75°S–88.75°N) and (b) tropical region (30°S–30°N).

We split all SST map sample images into three datasets: training, validation, and testing. The testing dataset comprised the most recent 15 years (2009–23) to evaluate our predictor. From the remaining samples, the validation dataset was selected to have five most recent high ACE years and five most recent low ACE years. The set was selected to be balanced between low and high years to result in an unbiased tuning of the learning. Validation set years were selected to be 2000, 2003–05, and 2008 (for high) and 1997, 2001, 2002, 2006, and 2007 (for low). Only the test set was considered for accuracy comparisons with NOAA's predictions.

2) OLR DATA

OLR data were used for our updated three-class prediction. OLR measures the energy emitted from Earth's surface, atmosphere, and oceans into space, providing a reliable indicator of deep convection (NCAR 2022). We downloaded OLR data in NetCDF format from the NOAA NCEP–NCAR CDAS-1 MONTHLY dataset under the variable name "diagnostic top upward longwave flux" (NOAA 2023b). This dataset provides monthly OLR averages indexed by a time variable (T, months since January 1960) and spans January 1949–November 2024. It covers longitudes from 0° to 360° and latitudes from 90°N to 90°S.

We note that the NCEP–NCAR dataset is a reanalysis of OLR data, which integrates historical observations into a numerical weather model to produce continuous and homogeneous estimates of atmospheric variables over extended periods. Consequently, the reanalysis allows us to extend the dataset to include OLR estimates from 1949 onward, even though direct satellite-based OLR observations are only available from 1979 (NOAA 2021b), which is crucial for expanding our training dataset, as the limited number of samples from 1979 to the present is insufficient for effectively training the CNN. We acknowledge that pre-1979 OLR estimates are

FIG. 3. July 2019 OLR map; longitude 0°–360°, measured eastward, and latitude ordered from 90°N to 90°S; data are measured in watts per meter square; white color represents high values of OLR; dark color represents low values.

model derived and may have greater uncertainty compared to satellite-based data. However, the augmentation relies on well-validated models, ensuring that the extended dataset remains suitable for our analysis.

Data processing was performed using Python's netCDF library. OLR data for July of each year from 1949 to 2023 were converted into single-channel (grayscale) images with 255 gray levels. The color scale was normalized to the maximum and minimum OLR values across all years as in Eq. (1).

Figure 3 displays a sample-generated OLR image for July 2019. The white regions represent high values of radiation, and the dark regions correspond with low values of radiation.

Recognizing that the Pacific and Atlantic basins, as well as the meridional gradient of OLR across Africa, are key predictor regions as described in the introduction, we performed a classification skill comparison for these respective regions. In this process, we generated new OLR maps for the corresponding latitude ranges: 85°W–35°E for the Atlantic basin, 130°E–85°W for the Pacific basin, and 35°W–45°E for the African continent region [capturing the meridional gradient of OLR across Africa considered in Karnauskas and Li (2016)]. Figure 4 depicts the corresponding geographic domains used for OLR in this study.

The OLR data were split into training and testing sets. Due to the limited amount of available data, no validation set was created. The testing set comprised the most recent 20 years (2004–23), while the training set included the remaining years (1949–2003).

b. *ACE index class labels*

The ACE of a hurricane season is calculated as the sum of the squares of the maximum sustained wind velocity of every tropical storm, measured every 6 h (CPC NOAA 2022). The ACE index incorporates storm genesis frequency, storm lifespan, and storm intensity (CPC NOAA 2022). We note that ACE measurements omit storms that technically undergo genesis by becoming TCs but never exceed tropical depression strength, thus only implicitly including storm genesis frequency. NOAA classifies hurricane season activity into three categories based on ACE levels (CPC NOAA 2022):

- High activity: ACE $> 111 \times 10^4$ kt^2.

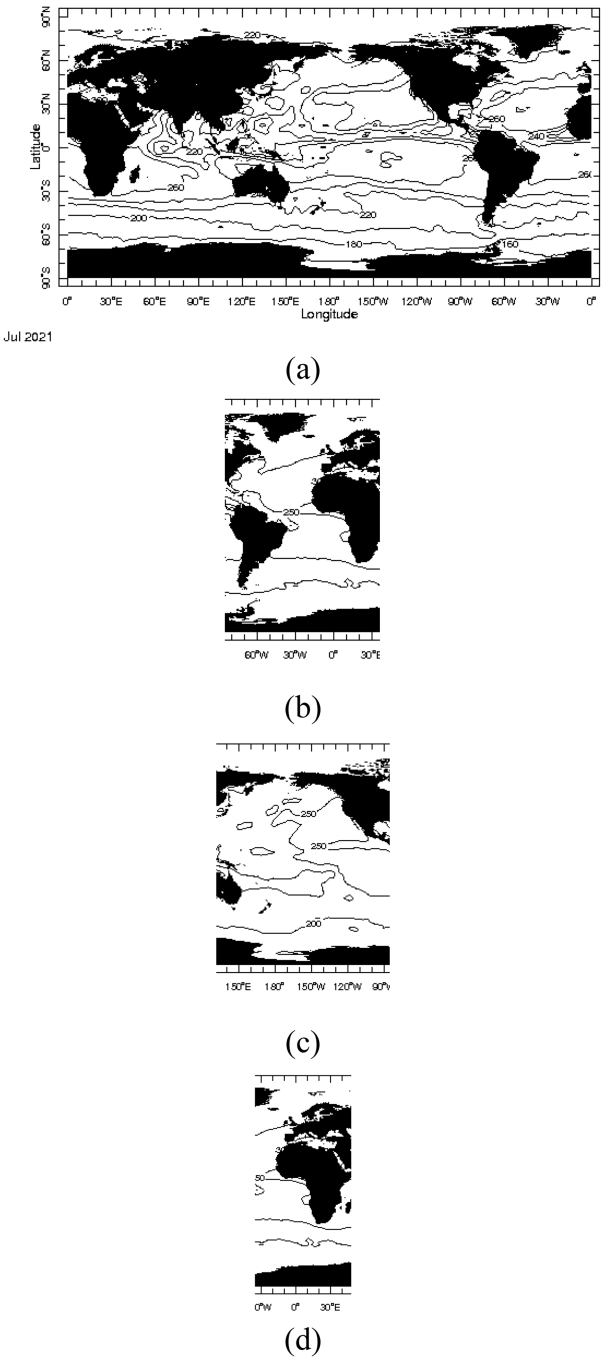

FIG. 4. (a) The full domain, (b) Atlantic basin (85°W–35°E), (c) Pacific basin (130°E–85°W), and (d) African continent (35°W–45°E) regions illustrated on a geographical map. This geographical map is a downloaded version of the OLR images showing just continents and OLR contours with labels of OLR values (NOAA 2023b).

- Medium activity: 66×10^4 kt$^2 \leq$ ACE $\leq 111 \times 10^4$ kt^2.
- Low activity: ACE $< 66 \times 10^4$ kt^2.

We constructed class labels for our data using NOAA's classification.

The SST images were paired with their corresponding class labels, which were determined by the ACE index for each respective year. ACE index data for the period 1856–2023 were downloaded from NOAA (Hurricane Research Division NOAA 2023). It is important to note that ACE values for the early twentieth and late nineteenth centuries are likely underestimates, as satellite measurements were not yet available before 1965 for detecting tropical storms (Vecchi and Knutson 2008). Consequently, some samples in our training set might have been classified as medium–low when they should have been high, potentially introducing a slight bias that reduces the detection accuracy for high-activity seasons. For our 4-month lead prediction based on SST, we simplified the classification into two categories: high and medium–low. Seasons with ACE $> 111 \times 10^4$ kt^2 were classified as high activity, while seasons with ACE $\leq 111 \times 10^4$ kt^2 were classified as medium–low activity. Using this classification, 47 samples were labeled as high activity, and 121 samples were labeled as medium–low activity.

In OLR's case, after we paired each OLR image to its year's ACE index, we assigned high, medium, and low activity class labels exactly following NOAA's ACE-based three-class categorizations.

c. CNN architecture

To classify the maps into activity categories, we used a CNN because it is best suited for extracting relevant features that characterize an image.

Given the fact that both SST and OLR datasets are similar in nature, i.e., both maps lack complex details, and we look for simple, local spatial patterns, we propose the same CNN architecture for both prediction methods. Hyperparameters, such as the number of filters and filter sizes, were designed using the validation set of the SST, and our choices were validated using both test sets of OLR and SST, respectively (see the appendix).

A challenge in designing the CNN was the limited amount of available training data for both of our prediction methods and the unbalanced datasets (much higher number of samples in the medium–low category for training set). To mitigate bias in learning and overfitting due to insufficient training examples (Li et al. 2021), we duplicated the most recent samples in each data class that was underrepresented, such that all classes will have the same number of training samples, and then, we enhanced the training SST and OLR datasets using image augmentation. Using the Python Keras library functions ImageDataGenerator and flow_from_directory, randomized versions of the initial image samples were generated in an infinite loop, ensuring diversity of samples across all epochs during each simulation (Chollet et al. 2015).

We acknowledge that using image augmentation is not a common approach for atmospheric science when data represent meteorological measurements. However, in our approach, we use a CNN to classify based on spatial gradient patterns and regions in the images rather than considering exact values. Consequently, image augmentation is appropriate in this context. Further, to minimize any potential negative impacts from augmentation, we use small perturbations of the images that are unlikely to disturb the localized spatial patterns that the network should capture. The following parameters were selected for the ImageDataGenerator function: rotation_range = 2, shear_range = 0.1, zoom_range = 0.2, width_shift_range = 3, height_shift_range = 3, and horizontal_flip = False. Furthermore, the validation dataset used for training will ensure that the model training is stopped once a good model is achieved, avoiding regression in learning due to potential detrimental samples.

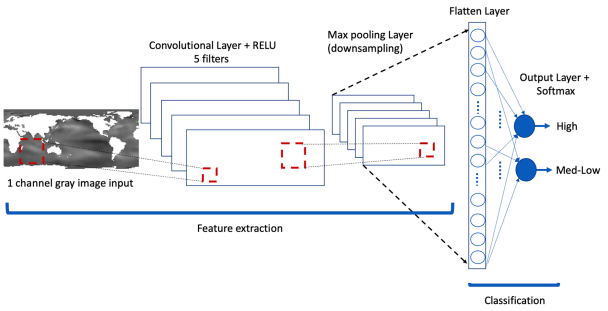

FIG. 5. CNN architecture for SST data input and two-class prediction.

Following the rule of thumb in the artificial intelligence community to have at least 10 times the number of training data samples as the number of trainable weights in the model (Alwosheel et al. 2018), we aimed to create a simple CNN model that will lead to a lower number of trainable weights. Our model has only five layers: an input layer, a convolutional layer, a max pooling layer, a flatten (fully connected) layer, and an output layer with two output neurons for the early predictor and three output neurons for the updated predictor, each using softmax function activation. In Fig. 5, we illustrate the CNN architecture for the midseason OLR three-class predictor. Our choice of one convolution layer is supported by our type of image data, since the first convolutional layer learns basic patterns and features, such as local gradients and color contrasts, which are the features we want our CNN to detect in the maps (Molnar 2022). In addition, the fact that our data images will likely not contain many detectable patterns justifies our use of a low number of filters (five for both SST and OLR). We selected 5×5 filter dimensions for SST and 3×3 filter dimensions for OLR after experimenting with various filter dimensions, including 3×3, 5×5, and 7×7 (see the appendix).

To further alleviate overfitting, we used $L2$ regularization, with $L2 = 0.08$. Our neural network model uses cross entropy as a cost function and minibatch stochastic gradient with the Adam optimizer. A batch size of 64 for SST and 32 for OLR was selected based on the number of samples available for training (143 for SST and 55 for OLR).

3. Results

a. SST-based early prediction results

We trained the CNN network 10 times, obtaining 10 different models. Each obtained model can be different because of different random components, especially due to the random image augmentations and random weights initialization.

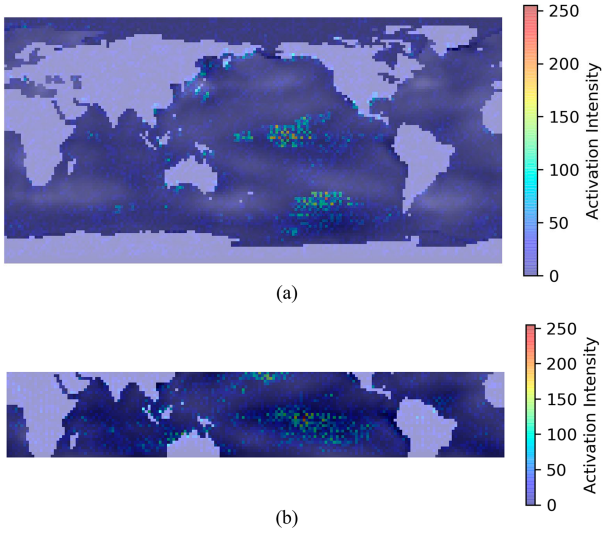

FIG. 6. Example of saliency map for January 2023 SST map: (a) full domain and (b) tropical region. Activation shown is for the high-class node. Highlighted regions represent regions of high activation. Activation intensity refers to how much changing one pixel will influence the output value.

Combining the outputs of multiple independently trained models is a well-established strategy for improving prediction performance. Opitz and Maclin (1999) found that most of the accuracy gains from such ensemble methods occur within the first 10–15 models, justifying our approach. The stopping criterion for training the full-domain-based network was for the network to reach at least 68% classification accuracy for the training set and at least 70% accuracy for the validation set. For the tropical-region-based network, we slightly reduced the training set and validation set accuracy stopping criteria, since the prediction skill of the model was lower. The tropical-region-based network had stopping criteria of 60% accuracy for both the training set and validation set.

Our two-class full-domain-based CNN predictor achieved an average accuracy of 67.83% ± 1.5% (95% confidence interval) over the 10 experiments, testing on the past 15 years and using a full-domain SST map as input. Ninety-five percent confidence intervals in this paper are calculated as ±1.96 times the standard error of the mean, assuming approximate normality.

Furthermore, we propose a predictor based on a majority voting over the 10 trained models to help alleviate errors from outlier models. The majority voting predictor achieved 73.33% prediction accuracy for the full domain.

We created saliency maps using the vis.visualize Python library (GitHub 2017) to understand which regions have the most prediction skill. Saliency maps highlight input regions that contribute the most changes in output, indicating areas with the greatest influence on the predicted class. The most salient parts of the image are highlighted in colors ranging from light yellow to red colors on the blue background. Figure 6 illustrates an example of saliency maps for the full domain and the tropical region models. Although we notice some low activation points (light blue) dispersed across land, these occur as artifacts of CNN learning based on noisy images. These low activation pixels have negligible impact on the classification. The concentrated high intensity activations in the Pacific basin suggest that the early prediction model relies heavily on SST signals in the tropical Pacific Ocean, indicating a potential influence from the ENSO phenomenon. This indication is consistent with the scientific consensus regarding the importance of ENSO as a factor in seasonal hurricane frequency (Gray 1984).

To further explore the impact of regional SSTs on our prediction skill, we also trained our classifier using maps for the tropical region only. We note that, although some of the regions of interest are contained within the tropical region and the region showed significant prediction skill, the average accuracy over 10 models decreased to 61% ± 2.26%. Furthermore, a majority voting classifier across 10 models did not improve the performance for the tropical-region-based predictor.

TABLE 1. Year-by-year prediction comparisons with NOAA's prediction for years 2000–22. Bold text signifies incorrect predictions.

Year	NOAA prediction May 3 classes	NOAA prediction May 2 classes	CNN SST majority voting prediction (January)	Actual ACE class (actual ACE index)
2009	**Medium**: 50%, high 25%, low: 25%	Medium–low	**High**	Low (53)
2010	High: 85%, medium: 10%, low: 10%	High	High	High (165)
2011	High: 65%, medium: 25%, low: 10%	High	High	High (126)
2012	**Medium**: 50%, high: 25%, low: 25%	**Medium–low**	High	High (129)
2013	**High**: 70%, medium: 25%, low: 5%	**High**	**High**	Low (36)
2014	**Low**: 50%, medium: 40%, high: 10%	Medium–low	Medium–low	Medium (67)
2015	Low: 70%, medium; 20%, high: 10%	Medium–low	Medium–low	Low (63)
2016	**Medium**: 45%, high: 30%, low: 25%	**Medium–low**	High	High (141)
2017	High: 45%, medium: 35%, low: 20%	High	High	High (223)
2018	**Medium**: 40%, high: 35%, low: 25%	**Medium–low**	**Medium–low**	High (132)
2019	**Medium**: 40%, low: 30%, high: 30%	**Medium–low**	**Medium–low**	High (130)
2020	High: 60%, medium: 30%, low: 10%	High	High	High (180)
2021	High: 60%, medium: 30%, low: 10%	High	High	High (146)
2022	**High**: 65%, medium: 25%, low: 10%	**High**	Medium–low	Medium (95.1)
2023	**Medium**: 40%, high: 30%, low: 30%	**Medium–low**	High	High (139)
Prediction accuracy		53.00%	73.33%	

TABLE 2. Confusion matrix table for our proposed CNN SST-based early predictions from the years 2000 to 2022.

	Actual high	Actual medium–low
Predicted high	8	2
Predicted medium–low	2	3

In Table 1, we compare the year-by-year predictions for the full-domain majority voting SST-based CNN model, which achieved an overall accuracy rate of 73.33%, to NOAA's May predictions for the most recent 15 years, 2009–23 (NOAA 2023a). For a fair comparison, we recalculated NOAA's prediction accuracy for two classes: high and medium–low. Our early majority voting CNN-based predictor is 20% more accurate than NOAA's predictions when evaluating the past 15 years of seasonal hurricane activity.

To further compare our prediction skill with NOAA's, we used Table 1 to build confusion matrices, which are seen in Tables 2 and 3 for our CNN predictor and NOAA's predictor, respectively. Our CNN performs better in detecting high seasons, missing only two seasons compared to four for NOAA. In addition, we calculated the bias scores for both our predictor and NOAA's predictor over the same prediction period. The bias score is defined as the number of hits plus the number of false alarms divided by the number of hits plus the number of misses (World Weather Research Programme and WCRP 2015). The bias score for our majority voting SST predictor is 1.0, while NOAA's bias score is 0.8, indicating a balanced model for our predictor and a tendency for NOAA's model to underpredict high seasons.

b. *OLR-based updated predictor results*

We trained the CNN network 10 times, resulting in 10 different models. Each model can vary due to random components in the training process, such as different weights, different initialization, and image augmentation (random transformations). To account for these variations and eliminate outlier models, we applied majority voting across the 10 models to determine the final predictions. The training stopping criterion was set to achieve at least 75% accuracy for the training dataset.

Our approach achieved an average accuracy of 80% when using full-domain OLR maps as input, with majority voting applied across 10 models. We created saliency maps to identify the areas that most strongly influence our predictions. The regions of activations seen in the saliency maps in Fig. 7 indicate that both the Pacific and Atlantic basins, which are known as key predictor regions, contribute significantly to the prediction skill. Additionally, our model detected some signals across Africa, consistent with the findings of Karnauskas and Li (2016), although these signals were relatively weak.

TABLE 3. Confusion matrix table for NOAA's May predictions from the years 2000 to 2022.

	Actual high	Actual medium–low
Predicted high	6	2
Predicted medium–low	4	3

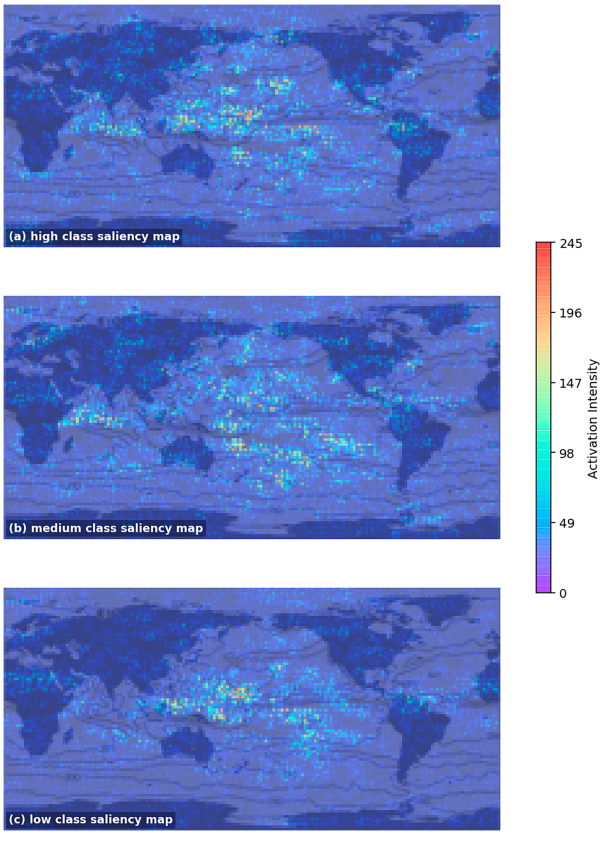

FIG. 7. July 2020 OLR saliency map example for low-, medium-, and high-class nodes superimposed on the geographical continents map for easy identification of regions of interest in prediction. Highlighted regions represent regions of high activation. Activation intensity refers to how much changing one pixel will influence the output value.

To further understand the contribution of the Atlantic (85°W–35°E) and Pacific (130°E–85°W) basin regions to prediction skill, a classification skill comparison was performed for the respective regions. Table 4 depicts the accuracy results for the predictions based on the Atlantic and Pacific basin regions of the OLR map. Both regions demonstrated high prediction skill: The Atlantic basin–based predictor and the Pacific basin–based predictor achieved similar accuracies with majority voting, 50% and 55%, respectively. However, the entire map region still provided the best prediction performance. This result is intuitively appealing, as the full-domain image enables the CNN to exploit the spatial distribution of information from all regions of the map. For the remainder of this paper, the OLR three-class prediction refers to predictions based on the full-domain OLR map.

In addition to the Atlantic and Pacific basins, we tested the African continent region. While some signals over the African continent are evident in our saliency maps, the region is too small to be used effectively for prediction. The training process did not produce reliable models, resulting in high-class predictions for all test samples.

TABLE 4. Comparison between average accuracies over 10 experiments and 95% confidence interval when using different input regions of OLR maps for our CNN and majority voting accuracy over the 10 models.

	Atlantic basin (stopping criteria: learning accuracy ≥ 65%)	Pacific basin (stopping criteria: learning accuracy ≥ 60%)	Full domain
Average accuracy	39%	43.50%	51.50%
95% CI	[23.38%, 54.62%]	[25.73%, 61.27%]	[41.59%, 61.41%]
Majority voting accuracy	50%	55%	80%

We compared our prediction accuracy for the most recent 20 years with NOAA's prediction accuracy for the same period. We note that testing on the most recent years is customary for hurricane prediction studies, as documented in Karnauskas and Li (2016), which used the most recent 15 years for testing. The prediction accuracy comparison, shown in Table 5, was obtained using majority voting across the 10 trained models for our CNN OLR–updated prediction, which achieved an accuracy of 80%. NOAA's predictions had an average accuracy of 50%, indicating that our proposed method with majority voting is 30% more accurate compared to NOAA's August prediction update for the past 20 years. It is noteworthy that three of the four hurricane seasons (i.e., 2007, 2013, and 2014) misclassified by our model were also misclassified by NOAA's predictions. This overlap suggests the presence of an interfering factor that is neither captured by our CNN model nor considered in NOAA's predictions. Specifically, although the 2013 hurricane season was expected to be active due to observed and predicted positive SST anomalies in the tropical North Atlantic, hurricane activity was suppressed by Rossby wave breaking over the extratropical North Atlantic, which inhibited hurricane development due to increased vertical wind shear over the tropical North Atlantic (Zhang et al. 2016). Such atmospheric forcing from the midlatitudes is a missing component in both our CNN models and NOAA's predictions. However, predicting these atmospheric forces based on large-scale conditions from previous months remains a challenge.

To further compare our prediction skill with NOAA, we used Table 5 to build confusion matrices, which are illustrated in Tables 6 and 7 for our CNN-updated predictor and NOAA's updated predictor, respectively. The confusion matrices show that both NOAA's model and our CNN OLR predictor classify high-class samples more accurately than those from the low and medium classes.

4. Conclusions

In this study, we developed a prediction system based on CNNs that leverages spatial information from data image maps to predict the upcoming Atlantic hurricane season's activity class (high, medium, low). Our models generally achieved higher accuracy compared with existing approaches.

Our approach consists of two main components. First, we developed an early prediction CNN model for two ACE-based activity classes, high and medium–low, which uses generated

TABLE 5. Year-by-year prediction comparisons with NOAA's updated August prediction for years 2004–23, using majority voting prediction over 10 CNN OLR models. Bold text signifies incorrect predictions.

Year	NOAA updated prediction August	OLR CNN prediction August	Actual ACE class (actual ACE index)
2004	High: 45%, medium: 45%, low: 10%	High: 39%, medium: 34% low: 27%	High (227)
2005	High: 95%–100%	High: 39%, medium: 32%, low: 29%	High (250)
2006	**High**: 75%, medium: 20%, low: 5%	Medium: 36%, high: 32%, low: 32%	Medium (81)
2007	**High**: 85%, medium: 10%, low: 5%	**High**: 39%, medium: 35%, low: 26%	Medium (74)
2008	High: 85%, medium: 10%, low: 5%	High: 39%, medium: 32%, low: 29%	High (146)
2009	**Medium: 50%**, low: 40%, high: 10%	Low: 37%, high: 34%, medium: 29%	Low (53)
2010	High: 90%, medium: 10%, low: 0%	High: 40%, medium; 34%, low: 26%	High (165)
2011	High: 85%, medium: 15%, low: 0%	High: 38%, medium: 32%, low: 30%	High (126)
2012	**Medium: 55%**, high: 35%, low: 15%	High: 40%, medium: 31%, low: 29%	High (129)
2013	**High: 70%**, medium: 25%, low: 5%	**High**: 38%, medium: 34%, low: 28%	Low (36)
2014	**Low: 70%,** medium: 25%, high: 5%	**Low**: 39%, high: 31%, medium: 30%	Medium (67)
2015	Low: 90%, medium: 10%	Low: 37%, high: 32%, medium: 31%	Low (63)
2016	**Medium: 50%**, high: 35%, low: 15%	High: 40%, medium: 32%, low: 28%	High (141)
2017	High: 60%, medium: 30%, low: 10%	High: 38%, medium: 31%, low: 31%	High (223)
2018	**Low: 60%**, medium: 30%, high: 10%	High: 38%, medium: 31%, low: 31%	High (132)
2019	High: 45%, medium: 35%, low: 20%	High: 40%, medium: 33%, low: 27%	High (130)
2020	High: 85%, medium: 10%, low: 5%	High: 42%, medium: 35%, low: 23%	High (180)
2021	High: 65%, medium: 25%, low: 10%	High: 40%, medium: 31%, low: 29%	High (146)
2022	**High**: 60%, medium: 30%; low: 10%	Medium: 37%, high: 33%, low: 30%	Medium (95.1)
2023	**Medium**: 40%, high: 30%, low: 30%	**Low**: 37%, medium: 32%, high: 31%	High (139)
Prediction accuracy	50.00%	80%	

TABLE 6. Confusion matrix for our proposed CNN OLR-based early predictions from the past 20 years.

	Predicted high	Predicted medium	Predicted low
Actual high	12	0	1
Actual medium	1	2	1
Actual low	1	0	2

January SST maps as input. The early prediction model provides forecasts in February, based on January data, offering a 4-month lead time before the hurricane season and 3 months ahead of NOAA's early prediction. Our early predictor, which applies majority voting across 10 trained models, achieves an average accuracy of 73.33%, 20% higher than NOAA's early predictions for the period tested (2009–22). Furthermore, our CNN model exhibits high accuracy in predicting high activity seasons, with a lower miss rate compared to NOAA's predictions. This capability makes it particularly effective in prompting preparedness for hurricane seasons with the potential for high activity.

In the second part of our study, we designed a midseason, high accuracy, three-class ACE-based prediction model. This model uses July OLR data maps as input to a similar CNN architecture, except with three output classes: high, medium, and low. Our OLR-based prediction becomes available in August, coinciding with NOAA's updated prediction. The majority voting model, applied across 10 trained predictors, achieved an accuracy of 80% when tested on data from the past 20 years. This represents a 30% improvement over NOAA's average August prediction accuracy during the same period.

Our results underscore the effectiveness of employing a CNN for prediction using encoded visualization image data, since CNNs excel at extracting image features and detecting relevant complex patterns within the data's spatial information. This capability allows them to uncover nonlinear relationships between inputs and outputs, thereby enhancing their predictive power.

For both types of maps, we analyzed the impact of various regions on our CNN's prediction skill. Saliency maps were implemented to identify the key regions contributing most to our prediction. The saliency maps highlighted that both predictors rely heavily on known regions of interest: the tropical region for SST and the Atlantic and Pacific basin regions for OLR, with especially strong signals over the Pacific basins. To further evaluate these regions, we conducted experiments using cropped images containing only the respective regions of interest. For SST maps, we found that the full domain achieved the best prediction accuracy. For OLR maps, the full domain also performed the best, while both the Atlantic basin and Pacific basin regions demonstrated good prediction skill. We also tested the African continent region as input to our OLR-based CNN, but the model performed poorly due to the weak signals detected over this region and very small training images. Overall, the full-domain map provided the best performance for both SST and OLR, despite the presence of isolated regions of interest. We suggest that this is because the CNN architecture on a full-domain image integrates the relationship among key prediction regions by leveraging the spatial distribution information across the entire map.

TABLE 7. Confusion matrix for NOAA's updated August predictions from the past 20 years.

	Predicted high	Predicted medium	Predicted low
Actual high	9	3	1
Actual medium	3	0	1
Actual low	1	1	1

Our proposed model demonstrated reasonable predictive capabilities using only SST and OLR data, providing an alternate method to regression, which is a difficult task, and instead classifying into broader activity classes. One of the main challenges in this study is accounting for the dynamic and nonlinear interactions among various atmospheric and oceanic factors, which are not fully captured by current CNN models. Future work could explore incorporating additional data sources, such as wind shear or atmospheric pressure fields, and using advanced architectures like attention mechanisms to better model these interactions (Vaswani et al. 2017). Furthermore, one interesting direction for future research would also be to compare the prediction skill of the image maps, which relies heavily on Pacific basin signals with predictors based on traditional climate indices, such as Niño-3.4 and AMO.

Acknowledgments. The statements, findings, and conclusions presented in this study are those of the authors and do not necessarily reflect the views of the National Oceanic and Atmospheric Administration, the U.S. Department of Commerce, or the U.S. Army Corps of Engineers. We gratefully acknowledge Dr. Zachary Labe for his valuable comments and suggestions. The authors declare no competing financial interests. Correspondence and requests for materials should be directed to Antonia Comaniciu.

Data availability statement. Our SST maps are generated based on the Extended Reconstructed SST (ERSST) dataset v5, which provides data from January 1854 to the present and is updated monthly (NOAA 2021a). Data are available to download at https://www.ncei.noaa.gov/products/extended-reconstructed-sst. Our OLR map images were encoded based on NOAA NCEP–NCAR CDAS-1 MONTHLY dataset, available using the variable name, diagnostic top upward longwave flux, at https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.MONTHLY/. ACE index data are available to download from NOAA at https://www.aoml.noaa.gov/hrd/hurdat/comparison_table.html.

APPENDIX

Testing Results for Hyperparameter Tuning

We conducted various experiments to determine the optimal number of filters (Tables A2 and A4) and the optimal filter dimensions (Tables A1 and A3) for our CNN network. Our results in Table A1 showed that 5×5 and 3×3 filter dimensions led to the highest accuracy for SST,

TABLE A1. Filter dimension testing for our SST-based CNN with five filters. The accuracy was obtained by averaging over 10 different simulations. Margin of error is displayed, calculated from 95% confidence interval.

CNN with five filters	3 × 3 filter dimensions	5 × 5 filter dimensions	7 × 7 filter dimensions
Average training accuracy	67.8% ± 1.5%	67.03% ± 0.67%	66.12% ± 0.97%
Average test accuracy	67.33% ± 5.04%	61.33% ± 4.82%	52.67% ± 6.52%

TABLE A2. Number of filters testing for our SST-based CNN, when filters selected have 3 × 3 dimensions. The accuracy was obtained by averaging over 10 different simulations. Margin of error is displayed, calculated from 95% confidence interval.

CNN with 3 × 3 filter dimension	3 filters	5 filters	10 filters	32 filters
Average training accuracy	67.61% ± 1.23%	67.80% ± 1.5%	66.36% ± 1.11%	68.61% ± 1.62%
Average test accuracy	64.67% ± 5.56%	67.33% ± 5.04%	60.00% ± 5.23%	61.33% ± 4.82%

TABLE A3. Filter dimension testing for our OLR-based CNN with five filters. The accuracy was obtained by averaging over 10 different simulations. Margin of error is displayed, calculated from 95% confidence interval.

CNN with five filters	3 × 3 filter dimensions	5 × 5 filter dimensions	7 × 7 filter dimensions
Average training accuracy	76.61% ± 1.96%	71.78% ± 1.32%	76.53% ± 5.14%
Average test accuracy	51.67% ± 10.64%	54.00% ± 12.14%	41.50% ± 12.40%

TABLE A4. Number of filters testing for our OLR-based CNN, when filters selected have 3 × 3 dimensions. The accuracy was obtained by averaging over 10 different simulations. Margin of error is displayed, calculated from 95% confidence interval.

CNN with 3 × 3 filter dimension	3 filters	5 filters	10 filters
Average training accuracy	73.84% ± 1.96%	76.61% ± 1.96%	72.42% ± 1.64%
Average test accuracy	51.67% ± 10.64%	51.67% ± 10.64%	58.50% ± 9.55%

and the differences between the two were not statistically significant. The 7 × 7 filter dimension performed significantly worse (the confidence intervals did not overlap with 3 × 3 and barely overlapped with 5 × 5). In Table A2, we show that three or five filters perform best for SST, while the performance deteriorates with increasing number of filters. This result is likely due to overfitting, as the differences between the accuracy of the training set and testing set widens. In Table A3, we see that the confidence intervals for all amounts of filters overlap, as the ranges are very large.

REFERENCES

Alaka, G. J., and E. D. Maloney, 2017: The intraseasonal variability of tropical cyclogenesis in the West African monsoon region. *J. Climate*, **30**, 5531–5553, https://doi.org/10.1175/JCLI-D-16-0750.1.

Alwosheel, A., S. van Cranenburgh, and C. G. Chorus, 2018: Is your dataset big enough? Sample size requirements when using Artificial Neural Networks for discrete choice analysis. *J. Choice Modell.*, **28**, 167–182, https://doi.org/10.1016/j.jocm.2018.07.002.

Asthana, T., H. Krim, X. Sun, S. Roheda, and L. Xie, 2021: Atlantic hurricane activity prediction: A machine learning approach. *Atmosphere*, **12**, 455, https://doi.org/10.3390/atmos12040455.

Bell, G. D., and Coauthors, 2000: The 1999 North Atlantic and eastern North Pacific hurricane season [in "Climate Assessment for 1999"]. *Bull. Amer. Meteor. Soc.*, **81**, S19–S22.

Caron, L.-P., F. Massonnet, P. J. Klotzbach, T. J. Philp, and J. Stroeve, 2020: Making seasonal outlooks of Arctic sea ice and Atlantic hurricanes valuable—Not just skillful. *Bull. Amer. Meteor. Soc.*, **101**, E36–E42, https://doi.org/10.1175/BAMS-D-18-0314.1.

Chen, B., B.-F. Chen, and H.-T. Lin, 2018: Rotation-blended CNNs on a new open dataset for tropical cyclone image-to-intensity regression. *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, London, United Kingdom, Association for Computing Machinery, 90–99, https://doi.org/10.1145/3219819.3219926.

Chollet, F., and Coauthors, 2015: A superpower for ML developers. Keras, accessed 22 June 2025, https://keras.io.

Comaniciu, A., and H. Murakami, 2022: Early prediction of seasonal Atlantic hurricanes using sea surface temperature maps and neural networks. *Proc. OCEANS 2022*, Hampton Roads, VA, Institute of Electrical and Electronics Engineers, 1–5, https://doi.org/10.1109/OCEANS47191.2022.9977003.

CPC NOAA, 2022: Background information: North Atlantic hurricane season. Accessed 28 December 2023, https://www.cpc.ncep.noaa.gov/products/outlooks/Background.html.

CSU, 2022: Forecast archive. Accessed 28 December 2023, https://tropical.colostate.edu/archive.html.

——, 2024: Seasonal hurricane forecasting. Accessed 6 April 2024, https://tropical.colostate.edu/forecasting.html.

Fu, D., P. Chang, and X. Liu, 2023: Using convolutional neural network to emulate seasonal tropical cyclone activity. *J. Adv. Model. Earth Syst.*, **15**, e2022MS003596, https://doi.org/10.1029/2022MS003596.

GitHub, 2017: What is saliency? Accessed 17 March 2024, https://raghakot.github.io/keras-vis/visualizations/saliency/.

Gray, W. M., 1984: Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences. *Mon. Wea. Rev.*, **112**, 1649–1668, https://doi.org/10.1175/1520-0493(1984)112<1649:ASHFPI>2.0.CO;2.

Griffin, S. M., A. Wimmers, and C. S. Velden, 2022: Predicting rapid intensification in North Atlantic and eastern North Pacific tropical cyclones using a convolutional neural network. *Wea. Forecasting*, **37**, 1333–1355, https://doi.org/10.1175/WAF-D-21-0194.1.

Hurricane Research Division NOAA, 2023: North Atlantic hurricane basin (1851–2022): Comparison of original and revised HURDAT. Accessed 28 December 2023, https://www.aoml.noaa.gov/hrd/hurdat/comparison_table.html.

Karnauskas, K. B., and L. Li, 2016: Predicting Atlantic seasonal hurricane activity using outgoing longwave radiation over Africa. *Geophys. Res. Lett.*, **43**, 7152–7159, https://doi.org/10.1002/2016GL069792.

Klotzbach, P. J., M. M. Bell, and A. J. DesRosiers, 2023: Forecast of Atlantic seasonal hurricane activity and landfall strike probability for 2023. Colorado State University Discussion, 46 pp., https://tropical.colostate.edu/Forecast/2023-07.pdf.

Li, Q., M. Yan, and J. Xu, 2021: Optimizing Convolutional Neural Network performance by mitigating underfitting and overfitting. *Proc. IEEE/ACIS 19th Int. Conf. on Computer and Information Science (ICIS)*, Shanghai, China, Institute of Electrical and Electronics Engineers, 126–131, https://doi.org/10.1109/ICIS51600.2021.9516868.

Lian, J., P. Dong, Y. Zhang, J. Pan, and K. Liu, 2020: A novel data-driven tropical cyclone track prediction model based on CNN and GRU with multi-dimensional feature selection. *IEEE Access*, **8**, 97 114–97 128, https://doi.org/10.1109/ACCESS.2020.2992083.

McNeely, T., P. Khokhlov, N. Dalmasso, K. M. Wood, and A. B. Lee, 2023: Structural forecasting for short-term tropical cyclone intensity guidance. *Wea. Forecasting*, **38**, 985–998, https://doi.org/10.1175/WAF-D-22-0111.1.

Molnar, C., 2022: *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable.* 2nd ed. Molnar, 317 pp.

NCAR, 2022: Outgoing Longwave Radiation (OLR): HIRS. Accessed 28 December 2023, https://climatedataguide.ucar.edu/climate-data/outgoing-longwave-radiation-olr-hirs.

NOAA, 2021a: Extended Reconstructed Sea Surface Temperature (ERSST). NOAA National Centers for Environmental Information, accessed 10 September 2024, https://www.ncei.noaa.gov/products/extended-reconstructed-sst.

——, 2021b: Outgoing Longwave Radiation—Monthly CDR. NOAA National Centers for Environmental Information, accessed 15 October 2024, https://www.ncei.noaa.gov/products/climate-data-records/outgoing-longwave-radiation-monthly.

——, 2023a: Atlantic hurricane outlook and summary archive. NOAA Climate Prediction Center, accessed 28 December 2023, https://www.cpc.ncep.noaa.gov/products/outlooks/hurricane-archive.shtml.

——, 2023b: NCEP–NCAR CDAS-1 MONTHLY. NOAA Physical Sciences Laboratory, accessed 28 December 2023, https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.MONTHLY/.

——, 2025: Hurricane costs. NOAA Office for Coastal Management, accessed 14 March 2024, https://coast.noaa.gov/states/fast-facts/hurricane-costs.html.

NOAA CPC, 2023: NOAA 2025 Atlantic hurricane season outlook. Accessed 6 April 2024, https://www.cpc.ncep.noaa.gov/products/outlooks/hurricane.shtml.

——, 2025: NOAA 2025 eastern Pacific hurricane season outlook. NOAA National Centers for Environmental Prediction, accessed 22 June 2025, https://www.cpc.ncep.noaa.gov/products/Epac_hurr/index.shtml.

Opitz, D., and R. Maclin, 1999: Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.*, **11**, 169–198, https://doi.org/10.1613/jair.614.

Pradhan, R., R. S. Aygun, M. Maskey, R. Ramachandran, and D. J. Cecil, 2018: Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Trans. Image Process.*, **27**, 692–702, https://doi.org/10.1109/TIP.2017.2766358.

Takaya, Y., and Coauthors, 2023: Recent advances in seasonal and multi-annual tropical cyclone forecasting. *Trop. Cyclone Res. Rev.*, **12**, 182–199, https://doi.org/10.1016/j.tcrr.2023.09.003.

Tan, J., Q. Yang, J. Hu, Q. Huang, and S. Chen, 2022: Tropical cyclone intensity estimation using Himawari-8 satellite cloud products and deep learning. *Remote Sens.*, **14**, 812, https://doi.org/10.3390/rs14040812.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, 2017: Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., 6000–6010, https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vecchi, G. A., and T. R. Knutson, 2008: On estimates of historical North Atlantic tropical cyclone activity. *J. Climate*, **21**, 3580–3600, https://doi.org/10.1175/2008JCLI2178.1.

Villarini, G., and G. A. Vecchi, 2013: Multiseason lead forecast of the North Atlantic power dissipation index (PDI) and accumulated cyclone energy (ACE). *J. Climate*, **26**, 3631–3643, https://doi.org/10.1175/JCLI-D-12-00448.1.

Wang, C., and X. Li, 2023: Deep learning in extracting tropical cyclone intensity and wind radius information from satellite infrared images—A review. *Atmos. Oceanic Sci. Lett.*, **16**, 100373, https://doi.org/10.1016/j.aosl.2023.100373.

World Weather Research Programme, and WCRP, 2015: Forecast verification methods across time and space scales. Accessed 15 March 2023, https://www.cawcr.gov.au/projects/verification/.

Xu, R., Z. Wu, J. Wang, and H. Li, 2023: Estimating hurricane intensity from satellite imagery using deep CNNs networks. *IEEE Second Int. Conf. on Electrical Engineering, Big Data and Algorithms (EEBDA)*, Changchun, China, Institute of Electrical and Electronics Engineers, 1271–1275, https://doi.org/10.1109/EEBDA56825.2023.10090706.

Zhang, G., Z. Wang, T. J. Dunkerton, M. S. Peng, and G. Magnusdottir, 2016: Extratropical impacts on Atlantic tropical cyclone activity. *J. Atmos. Sci.*, **73**, 1401–1418, https://doi.org/10.1175/JAS-D-15-0154.1.