Real-time 7-Day Forecast of Pollen Counts Using Deep Convolutional Neural Network

Yannic Lops[1], Yunsoo Choi[1] (ychoi6@uh.edu), Ebrahim Eslami[1], Alqamah Sayeed[1]

[1]Department of Earth and Atmospheric Sciences, University of Houston, Houston, TX 77204, USA

SUMMARY

This study implements a deep convolutional neural network with the great potential to recognize patterns of pollen phenomena that enable the prediction of pollen concentrations.

The model is trained using data from 2009 to 2015 from multiple meteorological data sets, satellite data, and processed data reflecting pollen flux as input for the model. The model forecasts pollen counts one to seven days ahead for the entire year of 2016. The performance of the algorithm for pollen prediction was evaluated using statistical parameters and categorical statistics evaluation by comparing 1-7 forecast to observation. The algorithm obtains a relatively high index of agreement (0.81-0.90) and Pearson correlation coefficient (0.88-0.75). Evaluation of categorical statistics based on defined threshold levels show satisfactory results.

Highlights

- Model developed and tested for real-time daily forecast
- Model can forecast daily pollen concentrations within minutes
- Good accuracy for daily weed and tree pollen forecast for 2016

STUDY AREA

Pollen data has been measured and acquired from the Houston Department of Health and Human Services (HDHHS) archives. Pollen data contains daily pollen (count/m^3) of 25 tree species, 15 weed species and one generalized grass species. We obtained meteorological data from the Texas Commission on Environmental Quality (TCEQ), which operates the Continuous Ambient Monitoring Sites (CAMS) in various metropolitan areas within the state of Texas. We selected data from CAMS station 695 (Moody Tower, near Downtown Houston) for its close proximity to the HDHHS pollen measurement station.

Map of Study Area

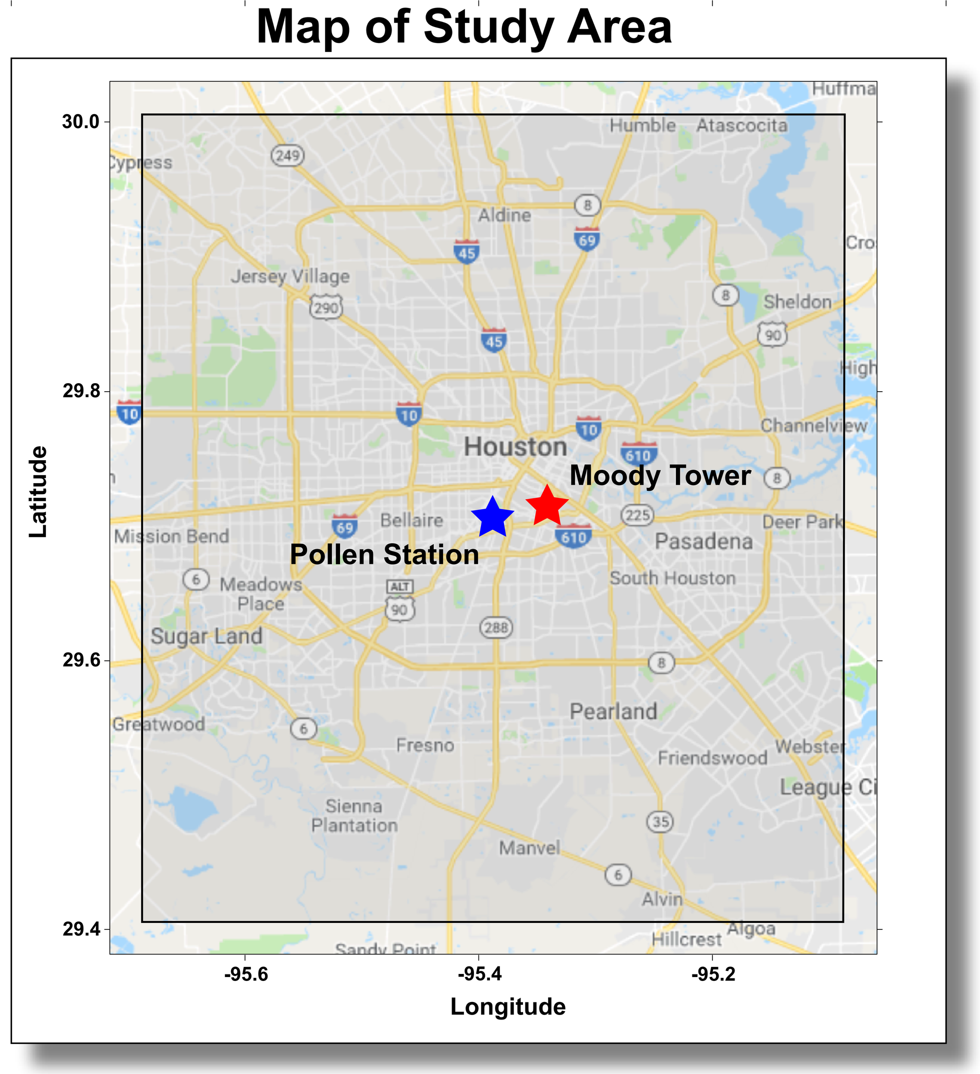

Fig. 1: Map of the study area in Houston. Moody Tower (Red) is the location of collected meteorology data. Pollen Station (Blue) is the location of collected pollen concentration data. A grey area (centered on the Pollen Station) represents the Leaf Area Index (LAI) based on MODIS data for pollen flux calculations.

METHODS

Deep CNN Forecasting System

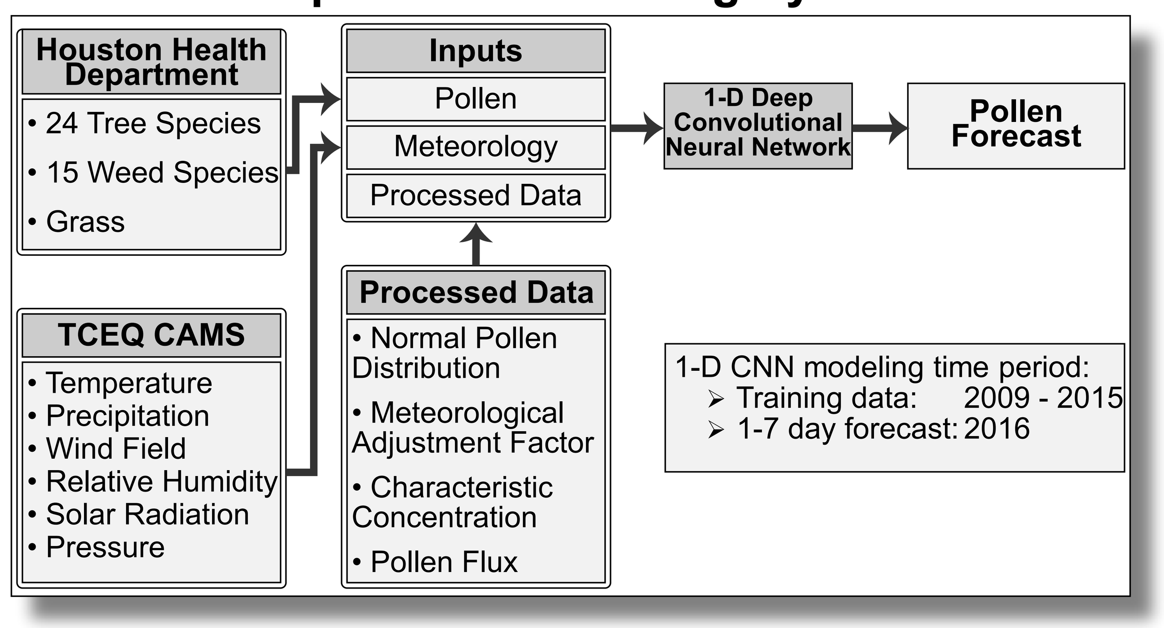

Fig. 2: Structure of the Deep CNN pollen forecasting system. Input data consist of pollen measurements, meteorological data, and processed data.

Processed data consist of:

- Meteorological adjustment factor (K_e)

$$K_e = 1 - \frac{3}{c1\frac{T}{T_{te}} + c2\frac{RH}{RH_{te}} + c3\frac{WS}{WS_{te}}}$$

T_{te}, RH_{te}, and WS_{te} represent the threshold values for temperature, relative humidity, and wind speed respectively. C_1, C_2, and C_3 are meteorological resistance factors.

- Normal pollen distribution (C_e)

$$C_e = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d-\mu)^2}{2\sigma^2}}$$

Where d is the number of consecutive days which pollen measurements meet or exceed the pollen count, μ is the mean distribution, and σ is the standard deviation.

- Characteristic concentration (C*)

$$C^* = \frac{p^q}{LAI \cdot h^c}$$

Where canopy height (h_c) is the mean canopy height of the vegetation species. *LAI* is the computed mean Leaf Area Index from MODIS satellite image data for the respective time period. p_q is set as 'Pollen Count +1'.

- Pollen flux (F_p)

$$F_p = C_e \cdot C^* \cdot K_e \cdot u_*$$

Where u_* is the averaged frictional velocity

Deep CNN Model

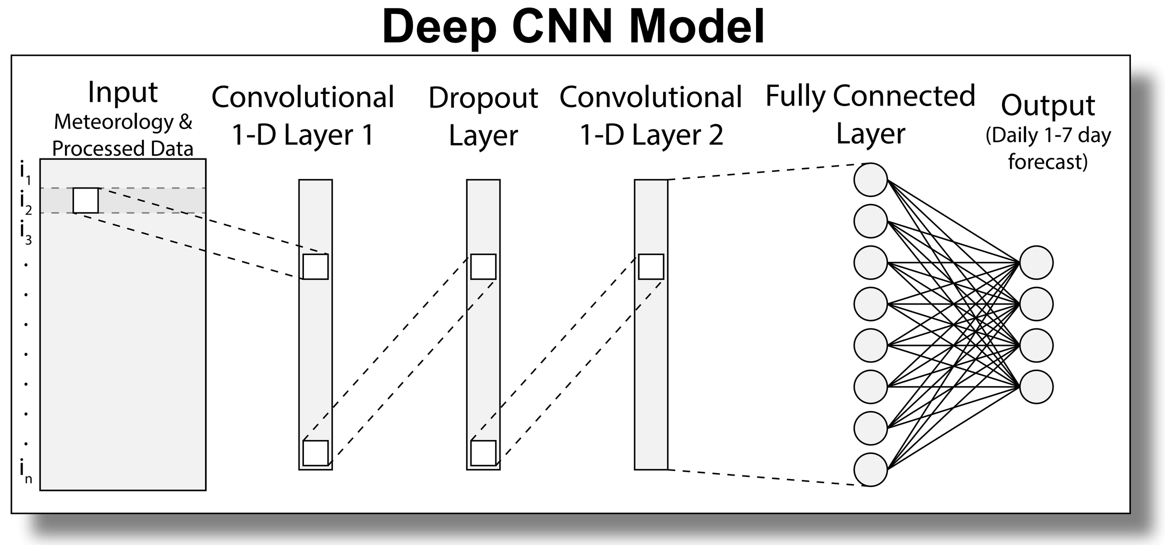

Fig. 3: Simple representation of the Deep CNN model.

RESULTS

For evaluation, the categorical statistics evaluation, and pearson correlation coefficient (r) and index of agreement (IOA) statistical evaluation methods are used. Each pollen category and days predicting ahead are evaluated.

The categorical statistics evaluation consist of:

- Hit Rate (HIT)
- Critical Success Index (CSI)
- False Alarm Rate (FAR)
- Equitable Threat Score (ETS)
- Proportion of Correct (POC)

Categorical statistic evaluation from 4 quadrants:

N_a - Predictions above and observations below threshold
N_b - Prediction and observation above threshold
N_c - Predictions and observations below threshold
N_d - Predictions below and observations above threshold

Categorical Statistics Evaluation Overview

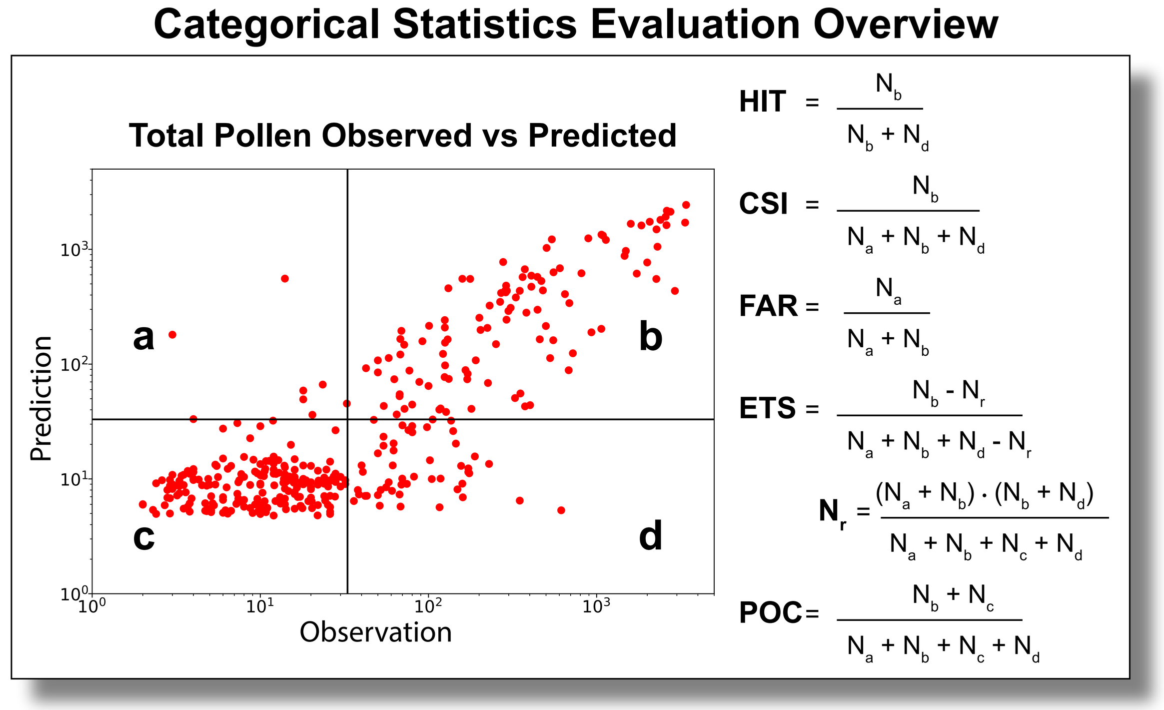

Fig 4: Overview on categorical statistics evaluation method. Threshold levels represent the concentration of pollen when allergic symptoms become significant. Threshold lines split the plots (right) into 4 quadrants, which are used to evaluate the model as an alarm system based on 5 categories (left).

Table 1: Statistical evaluation results of the deep convolutional neural network model based on threshold values for the respective pollen categories, their season ranges, and forecasting days ahead.

		Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Tree Pollen Threshold = 50 **Season:** Jan. 19 - May 12	*HIT*	0.667	0.679	0.603	0.667	0.628	0.641	0.615
	CSI	0.634	0.646	0.573	0.634	0.598	0.610	0.585
	FAR	0.071	0.070	0.078	0.071	0.075	0.074	0.077
	ETS	0.318	0.331	0.262	0.318	0.283	0.295	0.272
	POC	0.739	0.748	0.696	0.739	0.713	0.722	0.704
Grass Pollen Threshold = 30 **Season:** Jan. 21 - Jul. 13	*HIT*	0.313	0.250	0.250	0.125	0.125	0.125	0.125
	CSI	0.294	0.235	0.211	0.111	0.118	0.111	0.118
	FAR	0.167	0.200	0.429	0.500	0.333	0.500	0.333
	ETS	0.283	0.225	0.198	0.102	0.111	0.102	0.111
	POC	0.967	0.964	0.959	0.956	0.959	0.956	0.959
Weed Pollen Threshold = 20 **Season:** Sep. 6 - Nov. 11	*HIT*	0.885	0.904	0.904	0.923	0.923	0.923	0.904
	CSI	0.868	0.887	0.887	0.873	0.873	0.873	0.870
	FAR	0.021	0.021	0.021	0.059	0.059	0.059	0.041
	ETS	0.576	0.619	0.619	0.546	0.546	0.546	0.562
	POC	0.896	0.910	0.910	0.896	0.896	0.896	0.896
Total Pollen Threshold = 33 (Mean)	*HIT*	0.684	0.696	0.709	0.709	0.684	0.684	0.684
	CSI	0.667	0.675	0.685	0.671	0.642	0.657	0.663
	FAR	0.036	0.043	0.058	0.052	0.062	0.068	0.044
	ETS	0.525	0.532	0.541	0.526	0.492	0.506	0.518
	POC	0.850	0.855	0.844	0.850	0.847	0.861	0.836

RESULTS

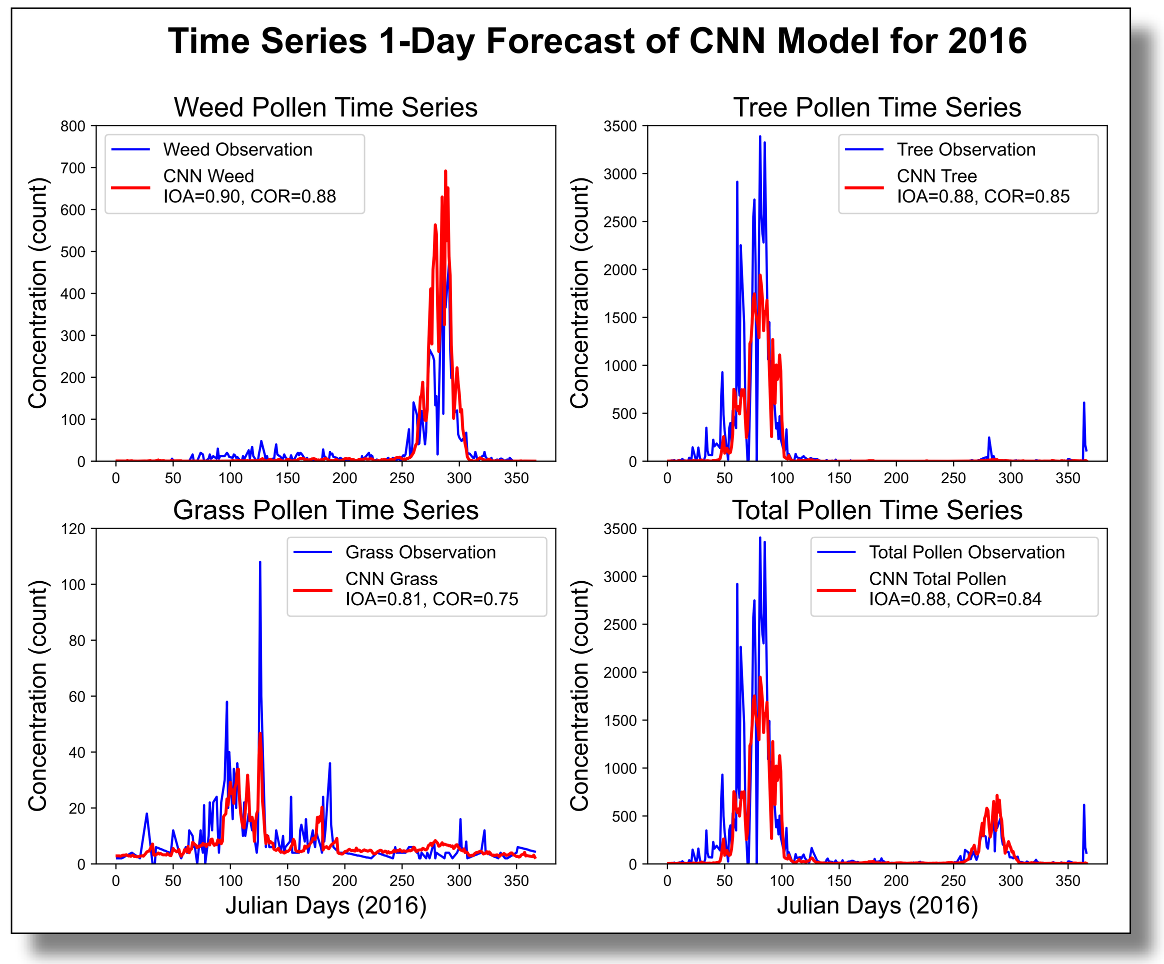

Fig 5: Time series of the deep convolutional neural network (CNN) model forecasting of Weed, Tree, Grass, and Total (sum of the tree, grass, and weed pollen) pollen concentrations one day ahead for 2016. Blue indicates the observed pollen concentration, and red indicates forecasted pollen concentration. Respective statistical scores are based on the entire year of 2016.

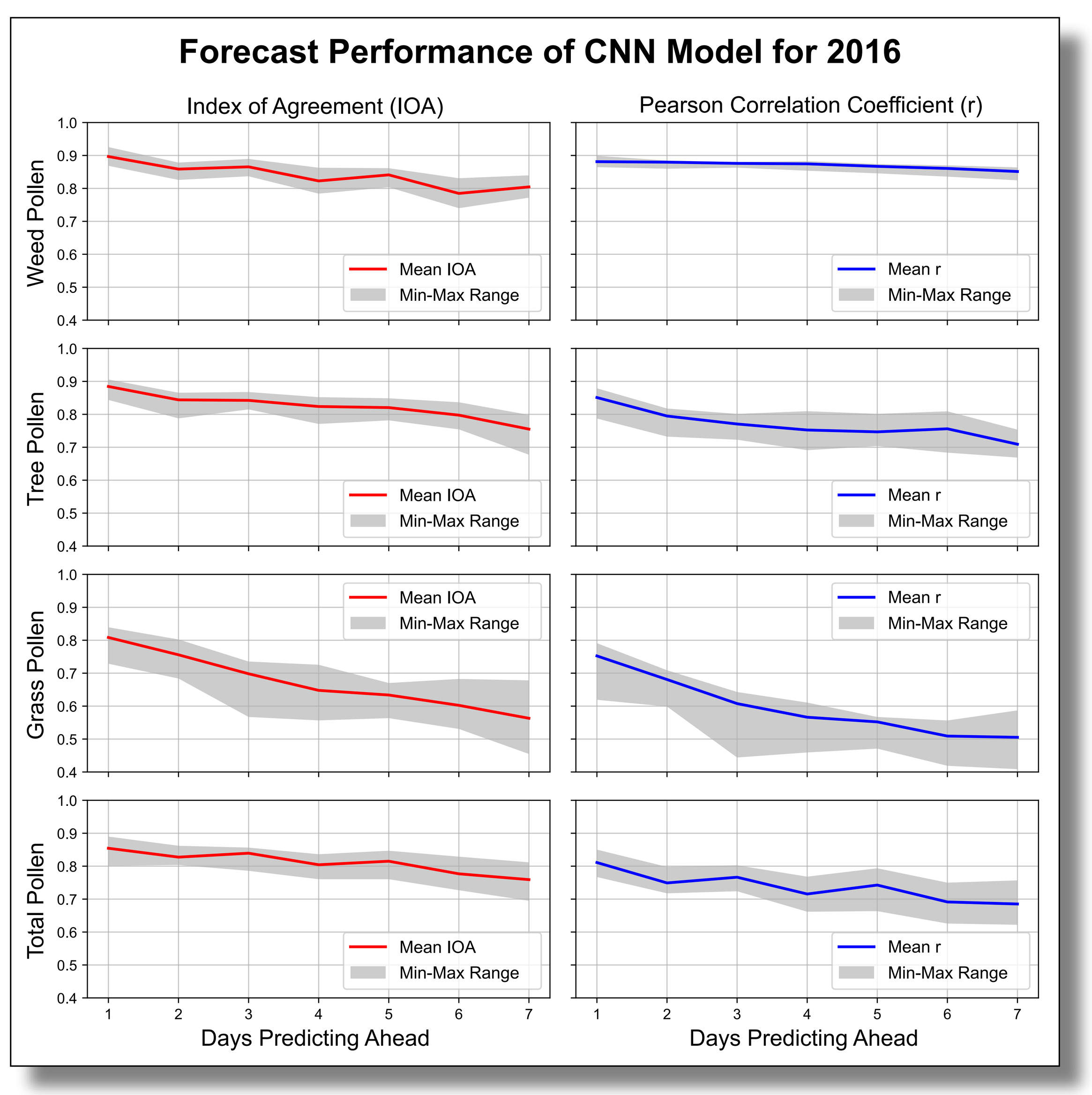

Fig 6: Deep convolutional neural network model (CNN) performance in forecasting tree, grass, weed, and total pollen 1-7 days ahead for the entire year of 2016. Red (IOA) and blue (Pearson correlation coefficient) lines represent the mean performance of the model over multiple runs. Shaded areas represent the maximum and minimum performance of the multiple runs for each category.

Conclusion: Our Deep CNN model forecasted real-time concentrations of pollen with favorable statistics within minutes of initializing the model. Thus, the computational efficiency of the deep CNN algorithm could supplement deterministic and regression models to more accurately and rapidly forecast pollen concentrations - offering a more reliable warning system for populations at high risk of pollen-related allergies.

Acknowledgement: We acknowledge the Texas Commission on Environmental Quality (TCEQ) and Houston Health Department for sharing their data.