scientific reports

OPEN

Influence of land-sea breeze on $PM_{2.5}$ prediction in central and southern Taiwan using composite neural network

George William Kibirige[1,2,3,4], Chiao Cheng Huang[1,4]✉, Chao Lin Liu[3] & Meng Chang Chen[1]

$PM_{2.5}$ prediction plays an important role for governments in establishing policies to control the emission of excessive atmospheric pollutants to protect the health of citizens. However, traditional machine learning methods that use data collected from ground-level monitoring stations have reached their limit with poor model generalization and insufficient data. We propose a composite neural network trained with aerosol optical depth (AOD) and weather data collected from satellites, as well as interpolated ocean wind features. We investigate the model outputs of different components of the composite neural network, concluding that the proposed composite neural network architecture yields significant improvements in overall performance compared to each component and the ensemble model benchmarks. The monthly analysis also demonstrates the superiority of the proposed architecture for stations where land-sea breezes frequently occur in the southern and central Taiwan in the months when land-sea breeze dominates the accumulation of $PM_{2.5}$.

Particulate matter (PM) is composed of air pollutants emitted into the atmosphere through human activities, urban development and industrialization. PM with an aerodynamic diameter smaller than or equal to 2.5 micrometers (μm) ($PM_{2.5}$) has been associated with cerebrovascular, cardiovascular, and pulmonary diseases[1–5]. In the Global Burden of Diseases study, $PM_{2.5}$ was ranked the sixth leading cause of human death[6]. One measure against $PM_{2.5}$ harm is to predict precise $PM_{2.5}$ concentrations; many governments have established ground monitoring stations to record $PM_{2.5}$ concentration to enact policies to control excessive atmospheric pollutants.

Taiwan's Environmental Protection Administration (EPA) has divided Taiwan into seven air quality zones according to geographical and meteorological conditions. Of these air quality zones, the middle and southern air quality zones suffer the most serious air pollution. The literature shows that the characteristics of weather and air pollution are widely considered and play important roles in $PM_{2.5}$ prediction[9]. In addition to $PM_{2.5}$ events caused by local emission, poor atmospheric diffusion conditions, and remote transport, $PM_{2.5}$ concentrations in central and southern Taiwan often reach the national warning threshold due to land-sea breezes[7,8,11]. Simulations from the literature have shown that land-sea breeze events occur with a northwest wind onshore formed during the day and east winds offshore at night[8]. However, this land-sea breeze effect is difficult to detect merely by monitoring station data.

The literature shows that the introduction of machine learning (ML) methods such as feedforward neural networks (FNNs)[12,13], convolutional neural networks (CNNs)[14], convolutional long short-term memory (ConvLSTM)[10,15] and random forests[16–18] improve the performance of $PM_{2.5}$ prediction. Recently, the development of deep neural network (DNN) approaches has overcome the weakness of other ML methods with their ability to capture complex interactions between datasets from different domains[19]. In our case, the introduction of DNN techniques facilitates the learning of spatio-temporal variation and the distribution of air pollutants from massive datasets. The presence of unknown factors also affects $PM_{2.5}$ prediction . For better prediction, the ensemble models (EMs) produce the softmax-weighted average of several ML model outputs to outperform DNNs[20]. AdaBoost (AD)[21], generalized additive models (GAM)[22,23], random forests (RF)[21,23], and extreme gradient boosting (XGBoost)[21,22,24] are popular EMs for $PM_{2.5}$ prediction. Recently, the composite neural network[25]

[1]Institute of Information Science, Academia Sinica, New Taipei, Taiwan. [2]Social Networks and Human Centered Computing Program, Taiwan International Graduate Program, New Taipei, Taiwan. [3]National Chengchi University, New Taipei, Taiwan. [4]These authors contributed equally: George William Kibirige and Chiao Cheng Huang. ✉email: neil1013@iis.sinica.edu.tw

has outperformed the EM methods in $PM_{2.5}$ prediction[10,26]. A composite neural network consists of individually pre-trained DNN components, each of which utilizes knowledge from datasets; component outputs are then connected as an acyclic tree. The leaf outputs are weighted by trained variables and collectively taken as an ensemble node, instead of being softmax weighted as in EM.

In this work, we build a remotely transported pollutants (RTP) model[10], a composite neural network consisting of two DNN components pre-trained by heterogeneous datasets from multiple sources to improve $PM_{2.5}$ prediction in southern and central Taiwan. We not only train the $PM_{2.5}$ prediction model using local meteorological and air pollution monitoring data, but we also introduce large-scale satellite images of East Asia to aid our model in capturing the spatiotemporal distribution of remotely transported $PM_{2.5}$. To capture the land-sea breezes that play an important role in $PM_{2.5}$ prediction in southern and central Taiwan, large-coverage wind features are also introduced. According to the observation in "Grouping of monitoring stations" and "Land-sea breeze", the proposed model should yield better $PM_{2.5}$ prediction results at stations where land-sea breeze frequently occur in months when land-sea breeze dominates $PM_{2.5}$.

Materials

Study region and air quality data.

The study region is located in the south and central part of Taiwan between latitude 21°25′ and 24°15′ north and longitude 120°12′ and 120°58′ east as shown in Supplementary Fig. S1. We created a grid area of $234\times80 = 18720$ km^2 that covers the study area for the subsequent data preprocessing. Each individual grid cell has a spatial resolution of 1 km.

The EPA monitoring stations detect air pollutants concentration values such as PM_{10} with a diameter of 10 μm, nitrogen dioxide (NO_2), other nitrogen oxides (NOx), ozone (O_3), carbon monoxide (CO) and sulfur dioxide (SO2). All of which strongly influence the formation and future status of $PM_{2.5}$. In this work, we collected hourly detected air pollutant data for 3 years (2014, 2015, 2016) from the Taiwan EPA (https://opendata.epa.gov.tw) as model input.

Aerosol optical depth data from MAIAC algorithm.

Aerosol optical depth (AOD) products are typically generated by dark target (DT) and deep blue (DB) algorithms at spatial resolutions of 3 to 10 km. However, AOD retrieval is challenging, especially when thick smoke is observed by satellite-based monitoring devices, which view the smoke as clouds. This makes the retrieved AOD data unreliable.

Multiangle atmospheric correlation implementation (MAIAC) is an advanced AOD retrieval algorithm based on time series analysis that has been proven reliable for predicting $PM_{2.5}$[27]. The accuracy of MAIAC AOD in China and East Asia has been validated by the AErosol RObotic NETwork (AERONET) ground measurement network[28]. Given MAIAC's strong performance and global coverage, we use these data to capture information on remote $PM_{2.5}$ transported long distances, for example, from one country to another[10].

In this work, we collected 3 years (2014, 2015, and 2016) of MAIAC AOD data at a 1×1 km^2 spatial resolution from NASA.(https://ladsweb.modaps.eosdis.nasa.gov) The AOD products cover two tiles from the investigation area (h28v06 and h29v06). The coordinates of the four corner points for h28v06 are (19°56′N, 106°3′E), (30°3′N, 115°5′E), (29°59′N, 127°2′E) and (19°53′N, 117°2′E). The coordinates of four corner points for h29v06 are (19°56′N, 116°41′E), (30°3′N, 126°37′E), (29°59′N, 138°34′E) and (19°52′N, 127°41′E). AOD preprocessing is described in "Data preprocessing".

Remote meteorological data.

$PM_{2.5}$ can float in the air for 4 to 7 days[29] and can be transported from one place to another with the help of meteorological features. Meteorological features are also involved in the formation of $PM_{2.5}$[29].

We used 3 years (2014, 2015, and 2016) of meteorological data from two different sources available in the remote area to capture more remote pollutants. The first source is data on temperature, pressure, vertical velocity (VVEL), absolute vorticity (ABSV), lifted index (LFTX), wind speed (ws) and wind direction (θ) at pressure levels from 10 mb (millibars) to 1000 mb (total 148 features) from the National Center for Environmental Prediction Final (NCEP FNL) Operational Global Analysis data.(https://rda.ucar.edu/datasets/) NCEP FNL data is provided in 2×7 grids which covers 27°N to 29°N in latitude and 120°E to 127°E in longitude at six-hour intervals. We pre-processed the data and converted them to hourly intervals, as explained in "Data preprocessing".

The second source is buoy monitoring stations that record the hourly wind speed and direction over the oceans. The ocean wind (OW) influences the diurnal variation of $PM_{2.5}$ in central and southern Taiwan[11]. Therefore, we create a grid area that covers Each grid cell has a spatial coverage of 1×1 km^2. We constructed wind direction and wind speed feature maps by filling non-observed grid cells using kriging interpolation based on wind direction and wind speed features Center Weather Bureau (CWB) stations and buoy weather monitoring devices. Another preprocessing is described in "Data preprocessing".

By interpolating non-observed grid cells with CWB stations on land and buoy weather monitoring devices on the ocean, we assemble wind feature maps that are reliable within our research area, which is encircled by buoy monitoring devices.

Local meteorological data.

The dispersion and transportation of $PM_{2.5}$ is strongly influenced by meteorological features (rainfall, pressure, temperature, humidity, wind speed, and wind direction)[27]. In this work, we downloaded these features from the CWB website,(http://opendata.cwb.gov.tw/index) which included hourly weather and weather forecast data from 337 monitoring stations. We preprocessed the data as explained in "Data preprocessing" using spatial interpolation to populate all non-observed grid cells and vectorize the wind speed and wind direction data as described in "Wind feature vectorization".

Methods

Wind feature vectorization. Our wind feature maps derive from wind features composed of speed and direction. Wind direction data are usually represented in polar coordinates, which must be converted to vector form. We vectorized the wind feature from wind speed at a particular angle into meridional (v-wind) and zonal (u-wind) components. To isolate the wind speed feature from the direction features, we then normalized the u-wind and v-wind by dividing them by the wind speed to yield the meridional and zonal components of the unit wind direction vector.

Data preprocessing. Data preprocessing includes conversion from monitoring station-based areas to a grid, linear interpolation, spatial interpolation to populate empty grid cells, data cleaning, and spatial downscaling.

For AOD, NCEP meteorological data and ocean wind which are input to STRI model in "Modeling methods", we vectorized the wind direction into zonal and meridian components of the meteorological dataset (NCEP) as described above. We also used linear interpolation to convert the meteorological dataset (NCEP) to hourly intervals from a six-hour interval.

We cleaned the MAIAC AOD data at 550 nm by filtering out poor quality grid values, after which we interpolated using the remaining grid cells. We also downscaled the spatial dimension of each remote tile (h28v06 and h29v06) to 300×300 km^2 from 1200×1200 km^2 using maximum pooling[15] to fit the available memory of the GPU. Then, we repeated the daily reading of each grid cell 24 times to match the hourly interval of other datasets.

To capture the spatio-temporal characteristics of the speed and direction of the ocean wind over the sea, we created a grid area (492×396 = 194,832 km^2) inside the remote area with each grid cell covering 1×1 km^2. We created a feature map by populating the dataset in the grid area according to the latitude and longitude coordinates of the monitoring stations (CWB and buoys). We used kriging interpolation to populate the remaining grid cells that did not match the station coordinates. Shown in Fig. 1 is an example of the results after kriging interpolation on the CWB and buoy dataset. Maximum pooling was applied to the kriging interpolated feature map to reduce the spatial dimensions to 246×198 km^2 to match the memory of the computing resource.

For air quality, weather, and weather forecast feature maps that are input to the base model in "Modeling methods", we converted the study regions to the grid area (234×80 cells) and created the feature map by populating the grid cells with the observed air quality and meteorological data according to the coordinates of the monitoring stations (37 EPA, 174 CWB) and using four nearest neighbors (4-NN) to populate grid cells outside these coordinates.

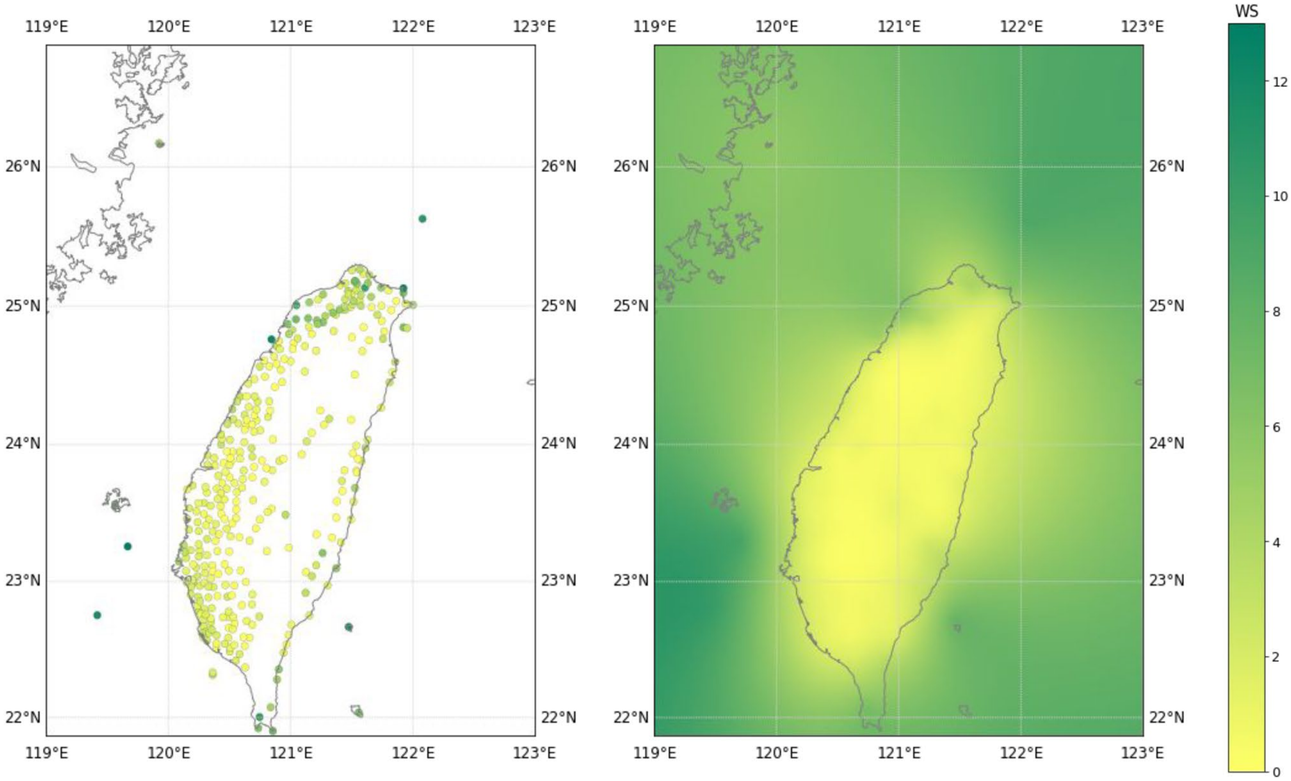

Figure 1. Left side: CWB and buoy monitoring stations. Right side: distribution of ocean wind dataset after kriging interpolation.

Modeling methods. The proposed composite neural network models—RTP with DNN components (base, STRI)—were trained over 2 years (2014, 2015) of data and tested on one year (2016). All models were constructed using Keras with a TensorFlow backend and trained on an NVIDIA GPU with 11 GB of memory.

STRI component. The spatiotemporal remote information neural network (STRI)[10] is a component of the RTP model that captures remotely transported $PM_{2.5}$ and predicts local $PM_{2.5}$ concentration. We added ML layers (CNN, ConvLSTM) to the STRI model to capture the spatiotemporal characteristics of the new heterogeneous dataset (AOD, meteorology, ocean wind). We included the AOD data to provide more spatial-temporal information on $PM_{2.5}$ remotely transported towards Taiwan.

In this work, the large STRI model with multiple layers of ML predicts the local $PM_{2.5}$ concentration of 37 EPA stations . The model uses large and heterogeneous datasets (AOD, meteorology, ocean wind) with local $PM_{2.5}$ as input. In Fig. 2, STRI_fe inputs 300×300 sized AOD satellite image, 2×7×148 sized NCEP meteorological grid data and the local $PM_{2.5}$ value; STRI_p inputs Kriging interpolated ocean wind grid data (246×198) and the embedding generated from STRI_fe. The idea is to capture spatiotemporal characteristics of heterogeneous datasets in different spatial scales, concatenate these, and then merge them with local features ($PM_{2.5}$) to predict local $PM_{2.5}$ concentration.

Furthermore, to fit the large STRI model into the GPU memory, we divided the model into two components, as shown in Fig. 2. STRI_fe, the first component[10], is used for the extraction of remote pollutants (ERP) given the AOD input from two tiles with their meteorology dataset. STRI_p, the second component, is used for the prediction given the ERP input, local features, and spatiotemporal features of ocean wind (Fig. 2). The detailed configuration of STRI model is described in detail in Supplementary Table S1.

After dividing the model into two components, we borrowed techniques from previous work[10] to fine-tune the individual components with fewer training parameters to improve the final prediction results.

Base component. The base model[10] is a component of the RTP model that predicts $PM_{2.5}$ concentration using local features only. The input to the base model is the air quality feature maps interpolated from EPA monitoring stations , the weather and weather forecast feature maps interpolated from CWB monitoring stations that covers the study area, and the prediction hour value to predict $PM_{2.5}$ of 37 EPA stations. . The model is described in detail in Supplementary Fig. S2.

RTP model. Given the prediction output of its pre-trained components (STRI and the base model), the RTP model outputs the final $PM_{2.5}$ predictions for the 37 EPA stations by hour. The RTP model is described in detail in Supplementary Fig. S2.

Evaluation

Metrics. We evaluated the proposed models using the root mean square error (RMSE), which measures the difference between the predicted $PM_{2.5}$ and its true value. In this work, the RMSE is the squared mean of the error between the ground truth and the predicted value at every hour among the monitoring stations of interest:

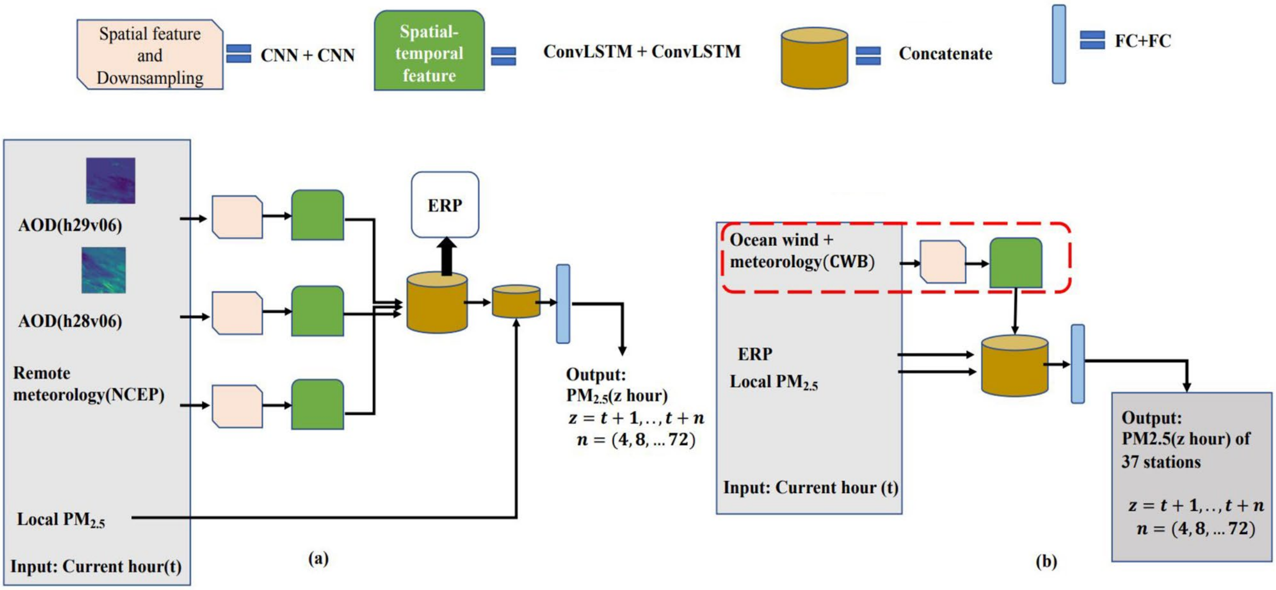

Figure 2. STRI model components STRI_fe (**a**) and STRI_p (**b**) with modifications indicated by red dashed line.

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{t=1}^{T}\sum\nolimits_{i=1}^{n}(y_{t,i} - \hat{y}_{t,i})^2}, \tag{1}$$

where $y_{t,i}$ and $\hat{y_{t,i}}$ are the true and predicted value of monitoring station i at hour t respectively, T is the length of the prediction sequence and n is the total number of monitoring stations.

Evaluation of proposed architecture.

We conducted experiments to show the the model performance for the next 3 days (72 h) $PM_{2.5}$ prediction at 4-hour intervals by comparing them with benchmarks and also to evaluate the contribution of each input feature to the prediction performance. In our architecture, each model was trained with the corresponding data from 2014 to 2015 and evaluated with the data from 2016.

1. We first compared the prediction performance of RTP_ow with its components (STRI pre-trained with the ocean wind and the base model) to evaluate the improvement of the composite neural network architecture with respect to $PM_{2.5}$ prediction.
2. After comparing the $PM_{2.5}$ prediction performance of the RTP model with its components, we compared the RTP model with other ensemble models (ADA, GAM, RF, XGB). RTP and the ensemble models use the same inputs: the prediction output of STRI and the base model. The main objective of these comparisons is to show that RTP outperforms its pre-trained components and other ensemble techniques.
3. Furthermore, We compared the $PM_{2.5}$ prediction performance of RTP models composed of STRI pre-trained with ocean wind (RTP_ow) and RTP models composed of the STRI pre-trained without ocean wind (RTP_no_ow) components to evaluate the effect of pre-training with ocean wind data on $PM_{2.5}$ prediction. Before this experiment, we grouped the 37 EPA stations of interests into two groups by ranking the frequency of land-sea breeze occurrences (The detailed grouping method is described in "Grouping of monitoring stations"). In this comparison, we averaged the RMSE of each group of stations during prediction hours to investigate the effect of ocean wind data on $PM_{2.5}$ prediction performance at stations where land-sea breeze frequently occur.
4. Finally, we present the monthly $PM_{2.5}$ prediction performance of models trained under the proposed architecture by averaging the RMSE of the 37 stations at each prediction hour in each month to compare the model performance during months when land-sea breezes affect southern and central Taiwan to the performance during the rest of the year. In addition to the RMSE of the 37 stations, we also averaged the RMSE of each group of stations grouped in "Grouping of monitoring stations" at each prediction hour to investigate the effect of ocean wind data on $PM_{2.5}$ prediction performance at stations where land-sea breeze frequently occur in each month.

Grouping of monitoring stations.

To evaluate whether the introduce of ocean wind data improves the $PM_{2.5}$ prediction performance at stations where land-sea breeze frequently occur, we selected the 28 stations land-sea breeze frequently occur and annotated these as LS stations, as listed in Supplementary Table S2. The processes we determine the LS sites are listed below. First, calculate the average daily wind direction (from 7am to 6pm) and the night wind direction of each site within 2014 to 2016; Second, calculate the counts of days when the daytime average wind direction lies between 157.5° and 337.5° (this step aims at counting the days when the daytime wind comes from the sea), and the difference of average wind direction in the daytime and nighttime is greater than 135° (this threshold indicates the significant diurnal wind direction variation); Third, select sites of top 28 counts as LS sites. The remaining nine stations (Xianxi, Lunbei, Mailiao, Taixi, Xingang, Puzi, Xinying, Annan and Hengchun) are annotated as normal stations.

Land-sea breeze.

Many studies present land-sea breeze with backward trajectory simulation for few hours period. To provide a synoptic observation of land-sea breeze in different season, a quantified metric, Jensen-Shannon divergence (JS divergence), is used to measure land-sea breeze through statistics from monitoring stations' observation. JS divergence is a method of measuring the similarity between two probability distributions. The lower JS divergence of two distributions is, the closer the two distributions are. As shown in Supplementary Fig. S3, we present the daytime and nighttime wind directions of each month in two discrete probability distributions. In practice, we present each probability as an array of 8 elements (the elements represent the probability of 8 principal wind directions within a month). Then, we calculate the similarity of the two distribution with the following JS divergence formula:

$$JSD(P||Q) = \frac{1}{2}\sum_{i} P_i \log\frac{P_i}{\frac{P_i+Q_i}{2}} + \frac{1}{2}\sum_{i} Q_i \log\frac{Q_i}{\frac{P_i+Q_i}{2}}$$

where P and Q represent the discrete probability distribution in the 8 principal wind directions in day and night, respectively; i represents each principal wind direction.

Thus, JS divergence is able to represent the diurnal wind direction variation in monthly probability.

In Supplementary Fig. S4, we presents four cases (Annan, Mailiao, Nanzi and Linyuan) of stacked bar plot which represent the distribution that $PM_{2.5}$ events of different AQI index level occur under eight principal wind direction every months in 2016. For each case, top row is the daytime distribution and the bottom row is the nighttime. JS divergence value is shown in top diagram. When land-sea breeze dominates, the diurnal wind direction distribution shift is recognizable, and JS divergence value is greater than 0.3.

From the four stations, we found JS divergences are relatively higher from April to August. This indicates that land-sea breeze dominates from April to August. For Nanzi and Linyuan, which are LS stations, the occurrences of $PM_{2.5}$ concentration at *Unhealthy for sensitive groups* and *Unhealthy* AQI index level grow at nighttime from March to May. This indicates that the $PM_{2.5}$ concentration at LS stations is dominated by land-sea breeze from March to May. However, $PM_{2.5}$ concentration is mostly under *Good* and *Moderate* AQI index level in June, July and August. In summer, strong vertical convection also influenced $PM_{2.5}$ and attributes to low $PM_{2.5}$ concentration. Hence, the proposed RTP is expected to perform well especially during March and May.

In January, February, November and December, the distributions between daytime and nighttime look similar and the JS divergences are small. The bars at northeast, north and northwest are relatively high. This indicates northeast monsoon dominates the whole day during the four months. However, for Nanzi and Linyuan stations, JS divergence values are relatively higher than the other two stations during winter. JS divergences of Linyuan station even keep above 0.3 for the whole year. As we have known that Linyuan and Nanzi are selected as the LS stations, this quantified metric clearly demonstrate that the LS stations are prone to be influenced by land-sea breeze.

The observations above help explaining the RTP performance in each month in "Monthly analysis".

Results

RTP and its components.

Figure 3 (left) shows that RTP and STRI both significantly outperform the base model from prediction hour 4 to 32. However, for the prediction hour after 32, STRI is worse than the base model, while RTP exhibits the best $PM_{2.5}$ prediction performance. This experiment shows that composite neural network architecture significantly improves $PM_{2.5}$ prediction compare to its components.

RTP model and other ensemble models.

As ensemble models such as AdaBoost, generalized additive models, random forests, and XGBoost have been widely used for $PM_{2.5}$ prediction, we further compared the RTP model trained under the proposed architecture with these models. In this experiment, we input the prediction output from both the STRI and the base model components into the RTP and the ensemble models. In Fig. 3 (right), the RTP model outperforms the ensemble models (ADA, GAM, RF, XGBoost) at every prediction hour. This shows that the proposed composite neural network architecture has the best overall $PM_{2.5}$ prediction performance in southern and central Taiwan with components pre-trained using large-scale AOD, weather, and ocean wind data.

Effect of pre-trained components on RTP model.

To evaluate the effect of ocean wind data on RTP model with respect to $PM_{2.5}$ prediction, we compared RTP composed of two different pre-trained STRI models: STRI pre-trained with $PM_{2.5}$ and AOD data (RTP_no_ow), and STRI pre-trained with $PM_{2.5}$, AOD, and ocean wind data (RTP_ow). In Fig. 4, comparing RTP_ow to RTP_no_ow shows that ocean wind features does help $PM_{2.5}$ prediction performance from prediction hour 4 to 28; however, Table 1 shows that in terms of average RMSE during the prediction hours, RTP_ow outperforms RTP_no_ow. This shows that ocean wind helps pre-trained components of composite neural network models to improve the overall $PM_{2.5}$ prediction performance during 72-hour prediction.

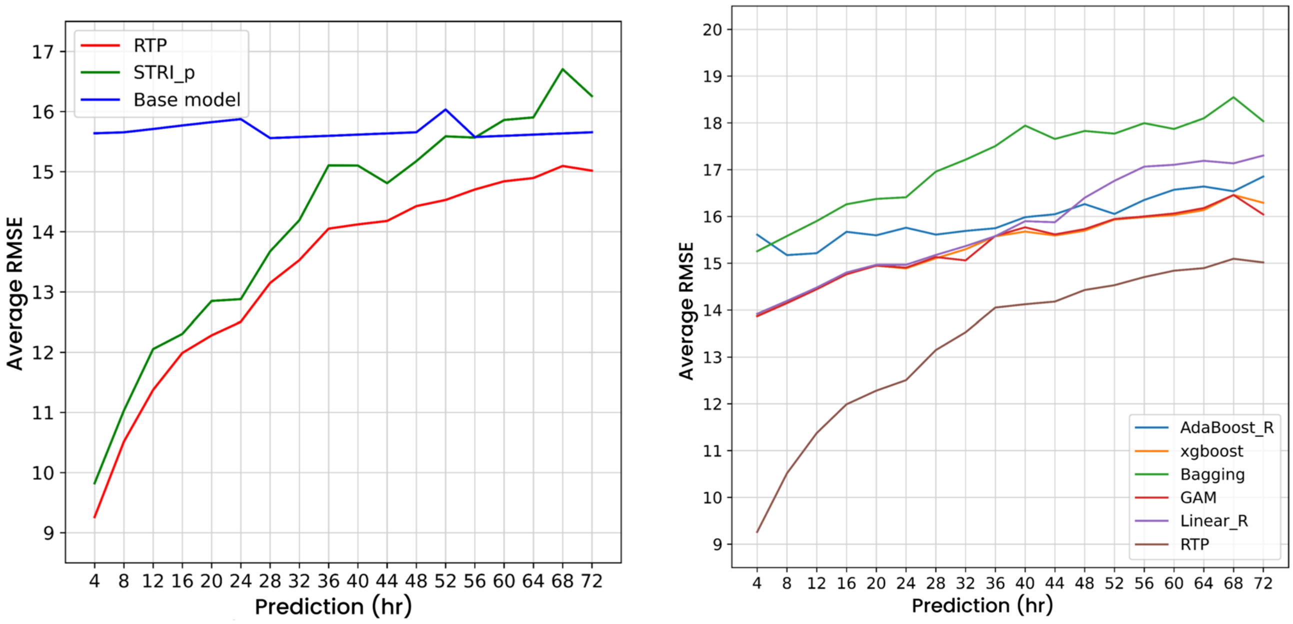

Figure 3. Left: average RMSE for RTP and its components. Right: average RMSE of RTP and other ensemble models.

	RTP_no_ow	RTP_ow
LS stations	13.4834	13.4037
Normal stations	13.2319	13.2190
All stations	13.4222	13.3588

Table 1. Average RMSE for land-sea (LS) and normal stations for RTP pre-trained with (RTP_ow) and without (RTP_no_ow) ocean wind.

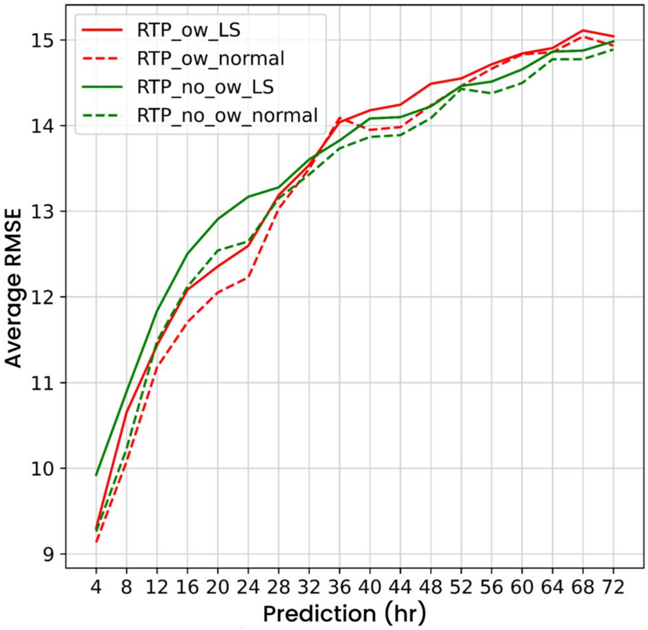

Figure 4. Average RMSE of LS stations (solid line) or normal stations (dashed line) in every prediction hour for RTP pre-trained with (RTP_ow) or without (RTP_no_ow) ocean wind.

Monthly analysis. In Fig. 4, although RTP_ow significantly improves the $PM_{2.5}$ prediction performance from prediction hour 4 to 28, RTP_ow_normal is obviously better than RTP_ow_LS through out every prediction hours, which means that the proposed architecture performs worse at LS stations during the whole testing period (2016). However, land-sea breeze does not dominate $PM_{2.5}$ throughout the year according to the observations in "Land-sea breeze". To evaluate whether the proposed composite neural network architecture that introduces ocean wind yields improved the $PM_{2.5}$ prediction performance for stations where land-sea breeze frequently occur, we separated the $PM_{2.5}$ prediction performance of RTP_ow for all 2016 into intervals of one-month for both LS and normal stations, as shown in Fig. 5. These monthly prediction results show that, in terms of average RMSE, RTP_ow for LS stations outperform RTP_ow for normal stations during prediction hours in March, April,and May. Importantly, RTP_ow for LS stations shows better performance compared to RTP_no_ow for LS stations from prediction hour 4 to 32 in March, April,and May. In June and August, when $PM_{2.5}$ pollution is the lowest in the year, both RTP_ow and RTP_no_ow have similar performance no matter for LS stations or normal stations.

In Supplementary Fig. S3, we present the monthly prediction results for autumn and winter: clearly, RTP_ow exhibits superior prediction performance for normal stations in September, October, December, and January. Although LS stations show higher JS divergence throughout the whole year according to Supplementary Fig. S4, northeast monsoon mainly dominates $PM_{2.5}$ events in *Unhealthy* level, which weaken the effect of land-sea breeze. In summary, the proposed composite neural network architecture that introduces ocean wind data, RTP, produces an improved $PM_{2.5}$ prediction performance for stations where land-sea breeze frequently occur in southern and central Taiwan in months when land-sea breeze dominates $PM_{2.5}$.

Conclusion

We propose a composite neural network architecture that uses components pre-trained with large-scale weather features and ocean wind to predict $PM_{2.5}$ in southern and central Taiwan. The neural network RTP_ow, which uses STRI, pre-trained with $PM_{2.5}$, AOD, large-scale weather features and ocean wind features as components, achieved the best overall $PM_{2.5}$ prediction performance compared to its individual components and other

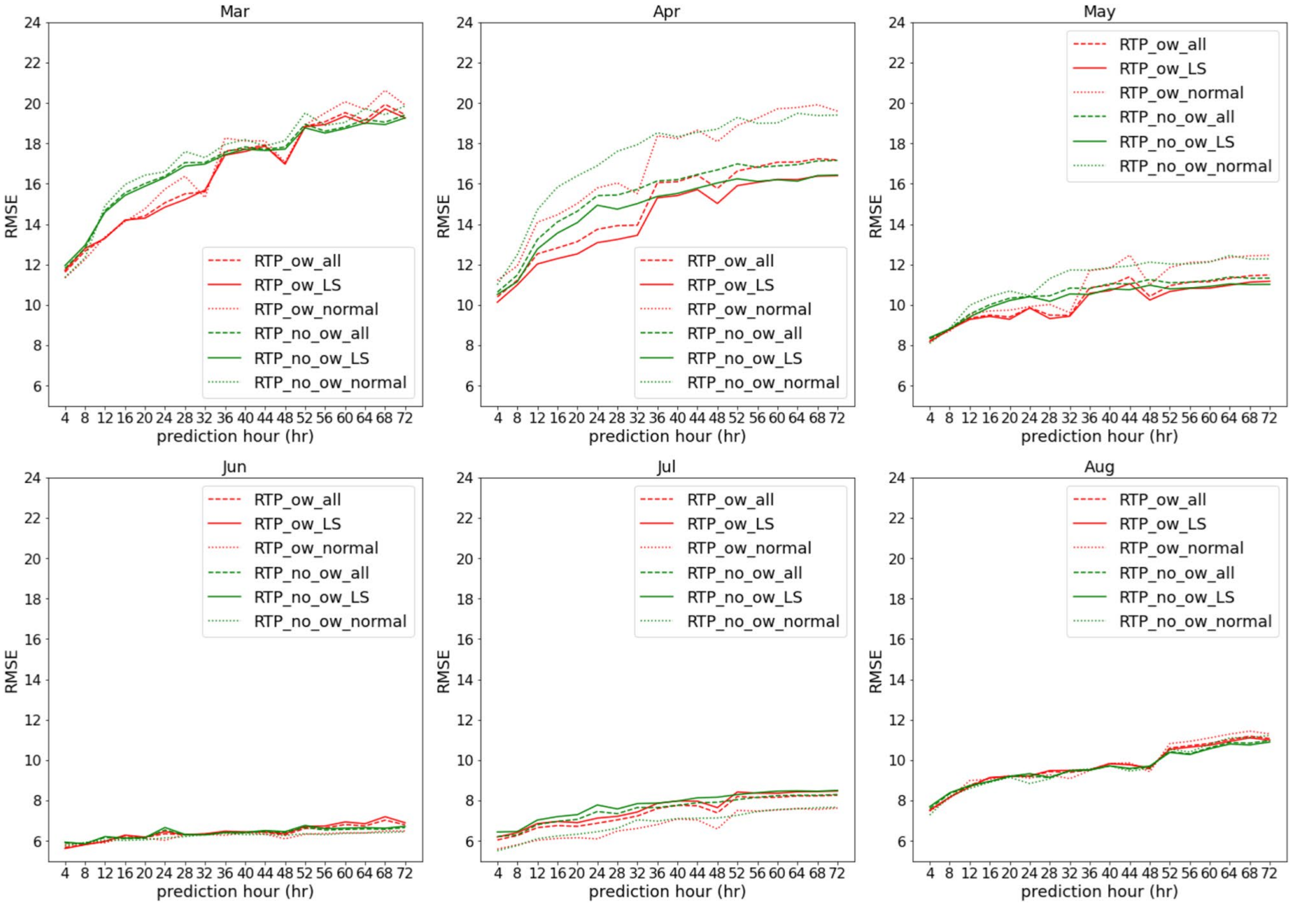

Figure 5. Monthly average in RMSE.

ensemble models. Monthly analysis reveals that the proposed model yields improved $PM_{2.5}$ prediction for LS stations in southern and central Taiwan in months when land-sea breeze dominates $PM_{2.5}$.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 24 August 2022; Accepted: 10 February 2023
Published online: 07 March 2023

References

1. Lippmann, M. Toxicological and epidemiological studies of cardiovascular effects of ambient air fine particulate matter (PM2.5) and its chemical components: Coherence and public health implications. *Crit. Rev. Toxicol.* **44**, 299–347 (2014).
2. Liang, R. *et al.* Effect of exposure to PM2.5 on blood pressure: A systematic review and meta-analysis. *J. Hypertens.* **32**, 2130–2141 (2014).
3. Stafoggia, M. *et al.* Long-term exposure to ambient air pollution and incidence of cerebrovascular events: Results from 11 European cohorts within the ESCAPE project. *Environ. Health Perspect.* **122**, 919–925 (2014).
4. Puett, R. C. *et al.* Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses Health Study. *Environ. Health Perspect.* **117**, 1697–1701 (2009).
5. Wu, C.-F. *et al.* Association of short-term exposure to fine particulate matter and nitrogen dioxide with acute cardiovascular effects. *Sci. Total Environ.* **569**, 300–305 (2016).
6. Gakidou, E. *et al.* Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**, 1345–1422 (2017).
7. Tsai, H.-H. *et al.* Physicochemical properties of PM2.5 and PM2.5–10 at inland and offshore sites over southeastern coastal region of Taiwan Strait. *Aerosol Air Qual. Res.* **11**, 664–678 (2011).
8. Cheng, F.-Y., Chin, S.-C. & Liu, T.-H. The role of boundary layer schemes in meteorological and air quality simulations of the Taiwan area. *Atmos. Environ.* **54**, 714–727 (2012).
9. Fang, X., Li, S., Xiong, L. & Zou, B. Analysis of pm2.5 variations based on observed, satellite-derived, and population-weighted concentrations. *Remote Sens.*https://doi.org/10.3390/rs14143381 *(2022).*
10. Kibirige, G., Yang, M.-C., Liu, C.-L. & Chen, M. C. *Using Satellite Data on Remote Transportation of Air Pollutants for PM2.5 Prediction in Northern Taiwan* (2021).
11. Hsu, C.-H. *et al.* Synoptic weather patterns and associated air pollution in Taiwan. *Aerosol Air Qual. Res.* **19**, 1139–1151 (2019).

12. Feng, X. *et al.* Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmos. Environ.* **107**, 118–128 (2015).
13. Biancofiore, F. *et al.* Recursive neural network model for analysis and forecast of PM10 and PM2.5. *Atmos. Pollut. Res.* **8**, 652–659 (2017).
14. Di, Q. *et al.* Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environ. Sci. Technol.* **50**, 4712–4721 (2016).
15. Sønderby, C. K. *et al. MetNet: A Neural Weather Model for Precipitation Forecasting*. arXiv preprint arXiv:2003.12140 (2020).
16. Huang, K. *et al.* Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. *Environ. Pollut.* **242**, 675–683 (2018).
17. Brokamp, C., Jandarov, R., Rao, M., LeMasters, G. & Ryan, P. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmos. Environ.* **151**, 1–11 (2017).
18. Wei, J. *et al.* Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. *Remote Sens. Environ.* **231**, 111221 (2019).
19. Yi, X., Zhang, J., Wang, Z., Li, T. & Zheng, Y. Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 965–973 (2018).
20. Li, L. *et al.* Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke. *Environ. Int.* **145**, 106143 (2020).
21. Zhai, B. & Chen, J. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. *Sci. Total Environ.* **635**, 644–658 (2018).
22. Shtein, A. *et al.* Estimating daily PM2.5 and PM10 over Italy using an ensemble model. *Environ. Sci. Technol.* **54**, 120–128 (2019).
23. Xiao, Q., Chang, H. H., Geng, G. & Liu, Y. An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. *Environ. Sci. Technol.* **52**, 13260–13269 (2018).
24. Li, L. *et al.* Estimation of PM2.5 concentrations at a high spatiotemporal resolution using constrained mixed-effect bagging models with MAIAC aerosol optical depth. *Remote Sens. Environ.* **217**, 573–586 (2018).
25. Meng, X. & Karniadakis, G. E. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. *J. Comput. Phys.* **401**, 109020 (2020).
26. Yang, M.-C. & Chen, M. C. PM2.5 forecasting using pre-trained components. In *2018 IEEE International Conference on Big Data (Big Data)* 4488–4491 (organizationIEEE, 2018).
27. Chu, Y. *et al.* A review on predicting ground PM2.5 concentration using satellite aerosol optical depth. *Atmosphere* **7**, 129 (2016).
28. Su, X., Wang, L., Zhang, M., Qin, W. & Bilal, M. A High-Precision Aerosol Retrieval Algorithm (HiPARA) for Advanced Himawari Imager (AHI) data: Development and verification. *Remote Sens. Environ.* **253**, 112221 (2021).
29. Cobourn, W. G. An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. *Atmos. Environ.* **44**, 3015–3023 (2010).

Author contributions

G.K. and C.H. contributed to data collection, the main manuscript text writing and literature survey. G.K. contributed to the main architecture establishment and prepared Figs. 2, S1 and Table S1. C.H. contributed to the the architecture perfomance analysis and interpretation of data, and prepared Figs. 1, 3, 4, 5, S2 and Table 1. M.C. and C.L. administrated and supervised this preojects. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-29845-w.

Correspondence and requests for materials should be addressed to C.C.H.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.