Environmental Science & Technology

Article

Cite This: Environ. Sci. Technol. 2019, 53, 13238−13245

pubs.acs.org/est

Deep Learning Prediction of Polycyclic Aromatic Hydrocarbons in the High Arctic

Yuan Zhao,[†] Li Wang,[‡] Jinmu Luo,[†] Tao Huang,[§] Shu Tao,[†] Junfeng Liu,[†] Yong Yu,[∥] Yufei Huang,[†] Xinrui Liu,[†] and Jianmin Ma*[,†]

[†]Laboratory for Earth Surface Processes, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China

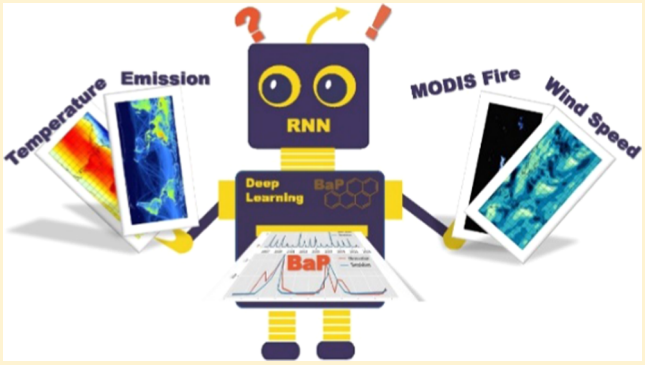
[‡]CAS Key Laboratory of Chemistry of Northwestern Plant Resources and Key Laboratory for Natural Medicine of Gansu Province, Lanzhou Institute of Chemical Physics, Chinese Academy of Sciences, Lanzhou 730000, China

[§]Key Laboratory for Environmental Pollution Prediction and Control, Gansu Province, College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China

[∥]Key Laboratory of Wetland Ecology and Environment, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China

Supporting Information

ABSTRACT: Given the lack of understanding of the complex physiochemical and environmental processes of persistent organic pollutants (POPs) in the Arctic and around the globe, atmospheric models often yield large errors in the predicted atmospheric concentrations of POPs. Here, we developed a recurrent neural network (RNN) method based on nonparametric deep learning algorithms. The RNN model was implemented to predict monthly air concentrations of polycyclic aromatic hydrocarbons (PAHs) at the high Arctic monitoring station Alert. To train the RNN system, we used MODIS satellite remotely sensed forest fire data, air emissions, meteorological data, sea ice cover area, and sampled PAH concentration data from 1996 to 2012. The system was applied to forecast monthly PAH concentrations from 2012 to 2014 at the Alert station. The results were compared with monitored PAHs and an atmospheric transport model (CanMETOP) for POPs. We show that the RNN significantly improved PHE and BaP predictions from 2012 to 2014 by 62.5 and 91.1%, respectively, compared to CanMETOP predictions. The sensitivity analysis using the Shapley value reveals that air emissions determined the magnitude of PAH levels in the high Arctic, whereas forest fires played a significant role in the changes in PAH concentrations in the high Arctic, followed by air temperature and meridional wind fields.



INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) have received extensive attention due to their carcinogenic, mutagenic, and teratogenic effects, posing exposure risks to human health. PAHs can travel in the atmosphere on regional and continental scales[1] and subsequently deposit on the underlying surfaces of soil, water, and ice/snow, which has been evidenced by their presence in the pristine Arctic environment.[2] Air provides a major pathway for the environmental cycling of PAHs on the Globe via emissions, diffusion, long-range transport, dry/wet deposition, gas−particle partitioning, and multimedia exchange. The Arctic PAH contamination has been primarily attributed to their emission in the Northern Hemisphere, such as from Asia, North America, Europe, and Russia.[3] Source apportionment studies showed that PAH sources in Asia (including the Russian Far East), Europe (including European Russia), and North America, respectively, accounted for 25, 45, and 27% of the PAH atmosphere levels in 2004 at Alert, a high Arctic sampling site for persistent organic pollutants (POPs).[4] Friedman and Selin (2012) used the global-scale atmospheric chemistry model GEOS-Chem to simulate phenanthrene (PHE), pyrene (PYR), and benzo[*a*]pyrene (BaP) in 2007 and compared their predicted PAH congeners with those measured at an arctic sampling site in Zeppelin, Norway. Their results revealed that Europe, Russia, and North America contributed 90% to the atmospheric levels of three congeners, among which European and Russian sources contributed more than 80%, while East and South Asia contributed 1% to PYR and 8% to BaP, respectively.[5]

Spatiotemporal changes in polycyclic aromatic hydrocarbons (PAHs) in the Arctic are, like other POPs, also subject to many

Received: August 19, 2019
Revised: October 16, 2019
Accepted: October 21, 2019
Published: October 21, 2019

DOI: 10.1021/acs.est.9b05000
Environ. Sci. Technol. 2019, 53, 13238−13245

complex environmental processes, in addition to emissions.[6] For example, although the air emissions of PAHs in the Northern Hemisphere have declined in the past decades due to worldwide emission regulation and control, the measured PAH concentrations have increased since the 2000s. This result applies particularly to the mid-2000s and is in line somewhat with the step changes in POPs occurring in 2004 and 2007, which have been attributed to Arctic climate warming.[7] The mid-2000s warming could have potentially promoted secondary emissions from PAH reservoirs in Arctic water, snow, and ice.[5,7] Friedman and Selin evaluated their GEOS-Chem-simulated seasonal concentrations for PHE, PYR, and BaP against the measured mean concentrations and reported that GEOS-Chem captured the sampled seasonal mean data only.[5] Yu et al. 2019 have recently employed the Canadian Model for Environmental Transport of Organochlorine Pesticides (CanMETOP) model to examine the unexpected increasing trend in PAHs observed at the Canadian high Arctic site Alert. Their modeled PAHs underestimated significantly the air concentrations at Alert and failed to predict increasing PAHs during the mid-2000s. The authors have partly attributed the disagreement between the modeled and measured PAH concentrations to the lack of local emission sources, such as forest fires, in the PAH emission inventory.[7] Forest fires and biomass burning have been considered an important source of PAHs.[8] Increasing incidents of forest fires in the Arctic have been reported and linked with Arctic warming,[8−10] perturbing the temporal changes in PAHs.[11] As a result, these complex environmental processes have created considerable difficulties in the prediction of Arctic PAH concentrations.

Alternatively, machine learning (ML) might provide a tool for the prediction of PAH fluctuations in the Arctic. ML is not a fresh concept in geoscience and has been used in spatial data analysis and remote-sensing image processing.[12] ML has also been increasingly applied in the prediction of air pollutants, such as principal component analysis,[13] for source apportionment.[7] Random forests are also an important method of ML for estimating environmental pollution because of their interpretability.[14] In addition, deep learning, as a typical ML technology, has become popular in recent years. Deep learning, as a potentially powerful tool for Earth science, has achieved rapid development among machine learning technologies and big data in recent years.[15] Deep learning can customize the structure for different missions, such as classification, regression, anomaly detection, and state prediction. Based on computer and custom neural networks, deep learning can also be employed in a nonlinear system and filter out the demand information from the input data (also referred to as "feature") to any specific output. However, deep learning has not yet been applied toward POP data analysis, and this is partly owing to the limited measurement data due to the high cost of lab sample analysis. On the other hand, the long-term POP monitoring program in the Arctic has been implemented since the early 1990s under the umbrella of the Arctic Monitoring and Assessment Program (AMAP).[2,11,16] The in situ measurement of air concentrations collected from this program might allow us to perform deep learning forecasts of PAHs in the Arctic. In the present study, an effort was made to apply the ML technique to recover sampled daily PAH time series and to examine and predict temporal variations in PAHs that have been associated with different environmental processes in the Arctic and remote sources, such as emissions, biomass burning, meteorology, and sea ice, aiming to introduce a potentially useful tool for the quantitative assessment of POPs in the environment.

MATERIALS AND METHODS

The recent study by Yu et al. 2019 has addressed inconsistencies in the monitored increasing PAH concentrations at the high Arctic site Alert with their declining emissions worldwide, but their modeled data failed to capture the inclining trend in PAHs at Alert.[7] To compare with their model's outcomes, we adopted the same measurement dataset and developed a deep learning model to recover the sampled data.

Sampled PAH Concentrations. In the present study, the focus is on a high Arctic monitoring site for POPs: the Alert monitor station (82°30′N, 62°20′W).[2] Monthly averaged atmospheric concentrations of PAHs from 1992 to 2015 were collected from the EBAS database (ebas.nilu.no). We selected 14 PAHs as the learning objects of the deep learning model. These were acenaphthene, acenaphthylene (ACY), anthracene, BaP, benzo[*b*]fluoranthene (BbF), benzo[*ghi*]perylene (BghiP), benzo[*k*]fluoranthene, chrysene (CHR), dibenzo-[*ah*]anthracene, fluoranthene (FLA), fluorene (FLO), indene-[1,2,3-*cd*]pyrene, PHE, and PYR.

Biomass Burning. The moderate resolution imaging spectroradiometer (MODIS) active fire products provide global burned areas that are burning at the time of overpass under relatively cloud-free conditions using a contextual algorithm. MODIS Collection 6 (C6) burned area product contains 0.25° global-gridded monthly burning area data and burning land type. We collected the Northern Hemisphere burning area data from June 1995 to Dec 2016. C6 burned area products can be accessed at ba1.geog.umd.edu.[17] It is worth noting that global fire data (http://www.globalfiredata.org/data.html) provide the fire emission directly, which might be used to replace MODIS C6 burned area product to improve the model prediction in future. MODIS C6 also includes land-cover types. However, the land uses were not taken into consideration in our machine learning prediction because our modeling practice showed that the land uses did not make a significant contribution to model performance as compared to other features.

PAH Emission Inventory. Monthly global PAH emission data on a 0.1 × 0.1° lat/lon resolution were collected from the Peking University (PKU) emission inventory (PEK-FUEL).[18] The data were extrapolated into the CanMETOP grid with a 1 × 1° lat/lon spacing. The polygon cutout was applied to divide the global emissions into five regions based on a shape file of national boundaries, including Canada and the United States of America (CAUS), Russia, India, China, Europe (excluding Russia), and the rest of the land region in the Northern Hemisphere. Detailed information for the countries included in each region is presented in Table S1. Global PAH emission data can be found at http://inventory.pku.edu.cn.

Meteorology Data. The monthly air temperature (T) and the meridional wind speed (WS) on a 2.5 × 2.5° lat/lon scale at 17 pressure levels were collected from the NCEP/NCAR reanalysis.[19] To address the roles of long-range atmospheric transport and the mean atmospheric circulation in Arctic PAHs, we selected air temperature and meridional wind data for 700 and 70 hPa, respectively. The 700 hPa represents the top of the lower atmosphere. This pressure level is also a level where the long-range atmospheric transport of POPs is most prominent.[20,21] Dynamically, the air temperature at 700 hPa

DOI: 10.1021/acs.est.9b05000
Environ. Sci. Technol. 2019, 53, 13238−13245

often reflects the major temperature characteristics within the low- and mid-atmospheres and is not disturbed by underlying surface characteristics, which might overwhelm the monthly mean temperature characteristic. The meridional winds at 70 hPa, the low stratosphere, reflect the large- or continental-scale mean atmospheric circulation pattern, which is also not disturbed by the local underlying surface characteristics. The meteorological conditions at these two levels typically represent the atmospheric activities in the low troposphere and low stratosphere.[22] Figures S1 and S2 show the correlation coefficients for the meridional winds and temperatures at different pressure levels. The winds and air temperatures at 700 and 70 hPa were most strongly correlated with the winds and temperatures at other pressure levels, suggesting that the wind and temperatures at these two levels stand best for the wind and temperature fields in the low (700 hpa) and high (70 hpa) free atmosphere. The optimally interpolated sea surface temperatures (SSTs) at 1 × 1° lat/lon from late 1981 to 2018 were also collected and used in the present study because SSTs are often associated with climate variation (http://www.emc.ncep.noaa.gov/research/cmb/sst_analysis).

Sea Ice Index. Sea ice area data from 1978 to 2018 were collected from the National Snow and Ice Data Center (https://nsidc.org/). The purpose of selecting the sea ice area is that changes in the sea ice area provide some of the most important signals of the seasonal, interannual, and long-term aspects of climate change that will otherwise affect the phase partitioning between environmental reservoirs and the secondary emission of POPs in the Arctic.[6]

Input Data Preprocessing. All input grid data or features, including the MODIS burned area, PAH emission inventory, and meteorological variables from June 1995 to Nov 2014, were summed (PAH emission) or averaged over the Northern Hemisphere, converted to a time-series dataset, as shown in the data flow from Figure 1a,b. The sea ice data themselves are

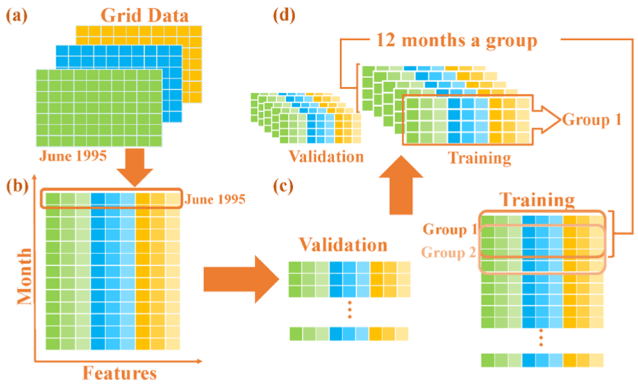


Figure 1. Flow chart for input data preprocessing. (a) Grid data format of the burned area, PAH emission inventory, and meteorology variables, (b) time-series data format, (c) splitting all data into the training set and validation set, and (d) input data every 12 months as a group. In the figure, the different colors denote the data flow in the different processes of the deep learning model.

a time-series dataset and hence did not need to be further processed. These time-series data were separated into two datasets: the training dataset and the validation dataset (Figure 1b,c). Since the time-series datasets' input into the model is not independent, cross validation cannot be taken in the present model.[23] The training dataset, extending from June 1995 to May 2012, was applied during the self-learning phase of the deep learning model. The validation dataset, covering the period from June 2012 to Nov 2014, was used to verify the deep learning predicted results. For the data training process, we assembled input data for every 12 months as a group of input (from Figure 1c to 1d) to predict the monthly PAHs in the last month of the group. For example, by inputting a group of features from June 1995 to May 1996 into the deep learning model, the model generated the monthly PAH concentration for May 1996.

Recurrent Neural Network. Neural networks deal with the problems of a nonlinear system by an inner activation function, especially when theoretical models cannot be constructed.[24] After extensive comparative evaluations and tests for different deep learning algorithms, we selected recurrent neural networks (RNNs), a class of neural networks in deep learning, to construct our deep learning model. The RNN extracts the information of a time series through iterating on each time step in the sequence of neurons. Long short-term memory (LSTM) is an RNN architecture constructed with an LSTM unit. An RNN model with LSTM is capable of handling a nonlinear time series[25] with multiple inputs. In our case, the RNN was applied to estimate temporally evolved PAH concentrations using multiple input variables, including the forest burned area, PAH emission inventory, meteorological variables, and sea ice area. These data with different dimensions and magnitudes were scaled by the RNN to predict and construct future PAH concentration time series. Figure 2 illustrates the LSTM unit and major RNN process.

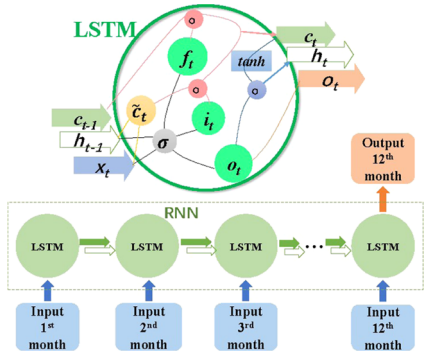


Figure 2. Flow chart for the RNN with an LSTM unit.

Supporting Information (SI) presents the corresponding equation for LSTM. There are three inputs in LSTM going from the left to the right of Figure 2: c_{t-1}, h_{t-1}, and x_t. x_t is the input of the current timestep, while c_{t-1} and h_{t-1} are hidden states that carry data information from the last time step. After the multiple operations and time steps in the LSTM run, the unit generates new c_t and h_t and outputs and achieves o_t in the last time step of our RNN. In the LSTM units, the forget gate f determines the information in previous data state c_{t-1} that will be forgotten, as shown in eqs S1 and S5. The c_{t-1} in the last time step, combined with new input i_t and the candidate hidden state $\tilde{c}$, is then iterated to yield the new cell state c_t, as described in eqs S2, S4, and S5 in SI. The output o_t is calculated from the current input i_t and its previous state h_{t-1}, as shown in eq S3. The predicted o_t and the state h_{t-1} from the last time step can then also be used to update the current state h_t (eq S6). In this way, the RNN operates 12 recurrences with successive 12 month PAH concentrations, and finally, generates the 12th month PAH concentration at the last

DOI: 10.1021/acs.est.9b05000
Environ. Sci. Technol. 2019, 53, 13238−13245

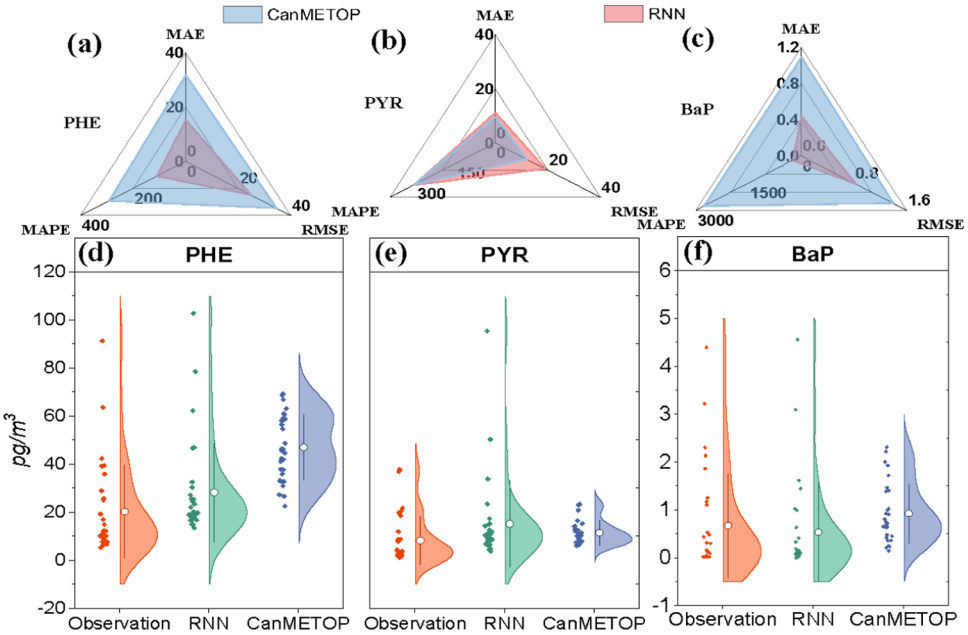


Figure 3. Performance indices, statistical distributions, means, and standard deviations for the modeled concentrations of PHE, PYR, and BaP by the RNN and CanMETOP against their respective measured concentrations.

recurrence. Here, we adopted an RNN featuring a deeper structure, powerful in computation skill, and requiring less training time. The custom RNN adopted in the present study has 7 113 985 parameters in total. The model includes seven layers, in which the last layer is a fully connected layer. The other layers are LSTM hidden layers with 256 units. More details of the custom RNN structure are presented in Table S2. It should be noted that relatively small data size of in situ PAH measurements to be trained in RNN could yield overfitting.[26] Efforts were made to minimize overfitting, including optimizing the split of training and validation data sets, performing model validation, and properly setting hyperparameters. We also adopted an early stopping approach to avoid overfitting.[27] These efforts reduced model overfitting considerably.

RNN Model Performance Evaluation. A total of four performance indices were adopted to evaluate the performance of the proposed RNN. These are the mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination R^2. These statistics are often recommended for the evaluation and validation of RNN forecasts.[28] The details of these statistical performance indices are presented in SI. From the expression of R^2 (eq S12), it is noted that this statistical quantity can be either positive or negative for a nonlinear variable.[29] We also compared the RNN model results for BaP, PHE, and PYR with the CanMETOP-simulated concentrations.

CanMETOP. CanMETOP is a three-dimensional atmospheric transport and multiple-environment exchange model (details in SI). The spatial resolution of CanMETOP is 1 × 1° lat/lon with 14 vertical levels, which have been used to simulate the transport of PAHs[30] and pesticides.[20,21] The simulated monthly concentrations of PHE, PYR, and BaP were used to compare with the RNN results.

Sensitivity Analysis. While a nonlinear RNN is a very useful tool for the prediction of a complex phenomenon, it often fails to discern the causes behind the phenomenon. In our case, the PAH concentrations were forecasted through the cooperation of multiple features and the changes in concentration. These features interact with each other in every LSTM unit. The details of the units are difficult to understand because of the enormous parameters and the confusing nonlinear process involved in the RNN algorithm. The sensitivity analysis aims at analyzing the relationship between input and output parameters in a black box or complex model, thereby revealing how and to what extent the input variables contribute to the output ones. Here, the input of the RNN is multidimensional, ordered, and interactive. The method chosen for the sensitivity analysis should possess the capabilities to deal with these input parameter characteristics. In the present study, we adopted the Shapley value (SHAP value, SI eq S13) approach derived from a solution concept in coalitional game theory, which quantifies the contribution from every feature of the input parameters toward the output variables.[31] The detailed SHAP values are presented in SI.

RESULTS

RNN Prediction. The RNN-predicted 14 PAHs were evaluated using four performance indices. The details are presented in Table S3, Figures S3 and S4. Among the 14 PAH congeners, R^2 for BaP, BbF, BghiP, CHR, FLA, and FLO are positive and greater than that of the other 8 PAHs. For these six congeners, smaller MAPE values (<200) were found for BbF, BghiP, FLO, and FLA. The smaller MAPE and larger R^2 suggest that the RNN exhibits a better performance for BbF, BghiP, FLO, and FLA. On the other hand, PHE has the smallest MAPE at 109.8 but its R^2 is −0.69 (an explanation for the negative R^2 is described in SI), suggesting that the RNN-forecasted PHE might include an outlier, located away from the measured monthly mean concentration. In a contrasting case, the predicted ACY has an R^2 value of −0.05, considerably smaller than that of PHE, but its MAPE value is 171.17. These characteristics suggest that the mean lies between the prediction and observation, but the predicted monthly ACY might contain a large deviation from the mean observation (Figure S5).

DOI: 10.1021/acs.est.9b05000
Environ. Sci. Technol. 2019, 53, 13238−13245

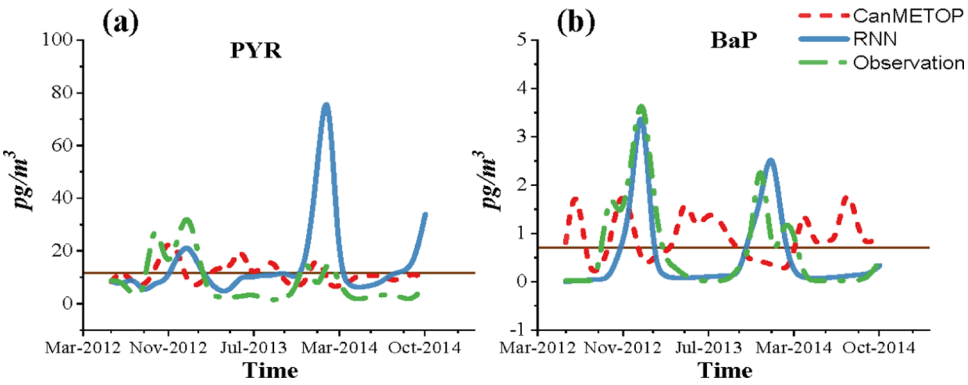


Figure 4. CanMETOP (dashed red line) and RNN (solid blue line) predicted PYR (a) and BaP (b) monthly mean air concentrations. The dashed green line stands for the concentrations measured at the Alert site. The solid brown lines are the mean observed PYR and BaP concentrations at the Alert site.

The performance indices can also be used to elucidate the temporal trend and variation in the modeled PAH concentrations. An $R^2 \approx 0$ suggests that the forecasting result approaches the observed mean, with weak upward and downward trends. An $R^2 < 0$ indicates that the sum of the forecasted concentrations has a large deviation from the observed mean, but such a large deviation does not occur in every individual month. The other two performance indices, MAE and RMSE, depend on different forecasting estimation methods but not among different PAHs because the concentrations of these PAHs exhibit different magnitudes when they are used to evaluate the model performance (Table S4).

Model Comparison and Evaluation. The RNN-predicted PHE, PYR, and BaP were further compared with the CanMETOP-simulated results. The statistics from the two models for the three PAH congeners against the measurements are shown in Figure 3a–c, in which a smaller triangle area indicates the better performance of a model. There is a small difference between the CanMETOP- and RNN-modeled PYR results. The performance indices for the CanMETOP-modeled PYR concentrations are smaller than those predicted by the RNN. Among these indices, the RMSE in the CanMETOP results is approximately half that of the RNN-predicted concentrations. The higher RMSE of the RNN-predicted PYR (Figure 3b) was caused by an outlier PYR concentration, which does not match well with the observation (Figure 4a). However, the mean concentrations predicted by the two models do not differ significantly in terms of their similar MAE and MAPE values (Figure 3b). For PHE and BaP, the RNN improves significantly the predicted concentrations compared with those from CanMETOP, as demonstrated by the smaller values of the performance indices shown in Figure 3 and Table S5. As can be seen, the MAPE values of the RNN-predicted PHE and BaP concentrations show 62.5 and 91.1% improvements over the CanMETOP-simulated results, respectively.

Figure 3d–f compares the distributions of the RNN- and CanMETOP-modeled and sampled concentrations of PHE, PYR, and BaP at the Alert site. The measured concentrations present a long tail and right-skewed distribution (also shown in Table S4). The right-skewed distribution is primarily attributed to the rapid decline in PAH air concentrations at Alert during the past decades. The long tail predicts an abnormal increase or decrease in PAH concentrations or a large departure from the mean concentrations (outlier). As shown, the RNN simulated successfully both the right-skewed distribution and long tail for PHE, matching well with the observations. Likewise, the RNN-predicted and the measured monthly mean concentrations of BaP are almost identical, as shown by their respective right-skewed distributions and long tails. However, the RNN appears not to yield a significant improvement in the modeling of PYR compared with that of CanMETOP because of the high RMSE value in the RNN forecasting (Figure 3b), likely attributed to an outlier in the RNN-forecasted PYR concentrations shown as the long tail of the concentration distribution (Figure 3e).

In contrast, the three CanMETOP-modeled PAH congeners show bimodal distribution patterns, particularly for PHE. The model failed to predict both right-skewed PHE and BaP concentration distributions and their long tails, suggesting a relatively low predictability for PAHs in the high Arctic compared with the RNN. This conclusion can be further illustrated by the triangles of three performance indices (Figure 3a–c). The CanMETOP-modeled PHE and BaP concentrations yield large statistical errors compared with their respective sampled ambient concentrations, as shown by the large shaded area in the performance index triangle. Overall, the RNN yields more accurate forecasts for the monthly concentrations of PHE and BaP.

Figure 4 compares the modeled and measured monthly time series and means (solid brown line) for the PYR and BaP air concentrations at Alert. The CanMETOP-simulated PYR concentrations oscillate around the mean of the sampled concentrations (solid brown line) with a small amplitude, whereas the RNN-predicted PYR concentrations exhibit large fluctuations, with a maximum value occurring in January 2014 (Figure 4a). This observation is a reason for the RNN-predicted PYR long tail shown in Figure 3e and the larger RMSE value shown in Figure 3b. In this sense, CanMETOP provides a better performance than RNN in forecasting monthly PYR concentrations. However, with a close look at the modeled monthly concentrations, one can identify that the RNN-modeled PYR concentrations display a similar monthly variation to that of the sampled concentration fluctuations. The Spearman ranking correlation coefficients between concentrations by the modeled RNN and CanMETOP and the sampling time series are −0.220 (p = 0.242) for CanMETOP and 0.301 (p = 0.106) for the RNN, implying that the RNN predicts the temporal variation in the monthly PYR concentrations better than CanMETOP. For BaP, the statistically significant Spearman correlations are 0.651 and −0.310 (p = 0.096) for the RNN and CanMETOP,

DOI: 10.1021/acs.est.9b05000
Environ. Sci. Technol. 2019, 53, 13238−13245

respectively. The RNN-predicted peaks and fluctuations in the monthly BaP time series are identical to those of the observation. Rather, CanMETOP does not catch these peaks and yields fluctuations opposite to those of the BaP monthly concentrations sampled at Alert. The Spearman correlation between the RNN- and CanMETOP-simulated and the measured PHE monthly concentrations is also positive; however, a greater value of the Spearman correlation is found with the RNN modeling result, confirming that the RNN-predicted monthly PHE time series agrees better with the measurement than does that by CanMETOP (Figure S6).

Sensitivity Analysis. Considering BaP as a marker of carcinogenic PAHs[32] and the excellent agreement between the RNN-simulated results and the in situ measurements, we selected BaP as a representative PAH congener to conduct the sensitivity analysis of the RNN modeling results of the input data (or feature) using the SHAP values and to examine potential contributions of the input data to temporal changes in the RNN-simulated BaP time series. Here, we calculated two SHAP values in the sensitivity analysis. The SHAP mean values indicate the positive and negative contributions of the input data (feature or independent variable) to the RNN-predicted results from 2012 to 2014 (output or dependent variables). The SHAP absolute value indicates the net contribution of the input variables to the output (namely, the predicted BaP air concentrations from 2012 to 2014). It is noted that all SHAP mean values and absolute values were summed over their all-month values and all-model grids. For example, as can be seen in Figure 5a, the temperature and wind speed at 700 hPa in the

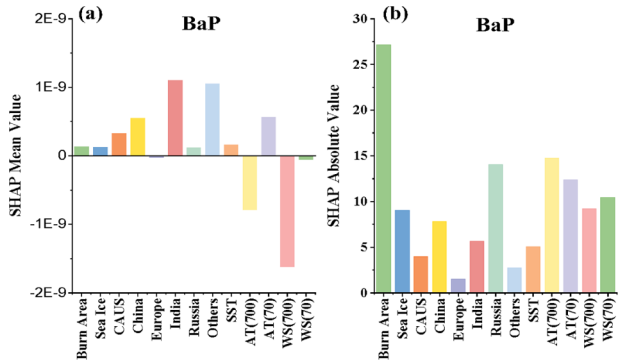


Figure 5. SHAP mean values (a) and SHAP absolute values (b) for input variables (Burn Area is the burned area, Sea Ice is the arctic sea ice area, CAUS is the denoted emissions for China, India, Russia, Europe and others, AT(700) and WS(700) stand for air temperature and wind speed at 700 hPa, and AT(70) and WS(70) stand for air temperature and wind speed at 70 hPa).

entire model domain show negative SHAP values, suggesting that these two summed input meteorological parameters at 700 hPa yield a negative contribution to the BaP concentrations at Alert, as featured by decreasing BaP concentrations. Based on the definition of the SHAP mean value, in some of the months, the temperature and wind speed might make positive contributions to the BaP levels, as defined by positive SHAP mean values, and in other months, these two meteorological input variables make negative contributions to the BaP levels, as indicated by the negative SHAP values. After summing these positive and negative SHAP values, we obtained the net SHAP value, as shown in Figure 5a. The positive net SHAP mean values occur in forest (biomass) burning, air emissions from different regions, sea ice, SST, and the wind speed at 70 hPa, indicating that these input variables contributed positively to BaP concentrations over the entire model, as featured by increasing BaP levels at Alert. In particular, the BaP air emissions from India and the other region (the rest of the land area in the Northern Hemisphere) make the largest positive contribution to the RNN-predicted BaP. As shown in Figure 5b, forest fires present the largest net contribution to BaP concentrations, followed by air emissions and 700 hPa temperatures. In this case, the contributions of the MODIS burned areas to the Alert BaP concentrations are either positive in some months or negative in others, and the summed net SHAP mean value for the burned area is small, as shown in Figure 5a. However, from summing all absolute SHAP values of the positive and negative SHAP mean values, the SHAP value for the burned area associated with high Arctic BaP is the largest of the input variables (Figure 5b).

DISCUSSION

With nonlinear data processing, unlimited input (independent or feature) variables, and no prerequisite assumptions in the relationship between input variables and output results, the RNN shows a better performance than CanMETOP in the prediction of PAH anomalies, means, and trends at the high Arctic site Alert. This performance is demonstrated by the statistical analysis between the modeled and ambient measurement data. We show smaller MAPE and larger R^2 values for the predicted PAHs against the measured values using the RNN, manifesting in the smaller differences (eq S11) and similar fluctuations (eq S12) between the modeled and sampled concentration time series. We also reveal that the RNN detects the large concentration anomalies and outliers in the sampled time series poorly. It is likely that these anomalies and outliers were caused by unknown environmental processes that were not associated with the input (independent) variables or training data. On the other hand, if an environmental event occurred in the training data but not in the validation data, an outlier would occur in the predicted concentration time series but not in the observation. Large anomalies and outliers in the sampled concentration time series might be induced by the accidental release of PAHs from local sources, step changes in climate, and human activities near the sampling site. These uncertainties can be overcome by introducing more training data that include as many environmental processes as possible.

Based on both the SHAP mean and absolute values in the sensitivity analysis, the emissions from different regions all together made the most significant and positive contribution to BaP levels, whereas meteorological conditions and forest fires overwhelmed the fluctuations. In particular, although the net SHAP mean value for forest fires (burned area) is small (Figure 5a), after summing its positive and negative values, the largest absolute SHAP value is found for forest fires (burned area, Figure 5b), indicating the significant influence of forest fires on the BaP concentrations at Alert. It has been known that forest fires are important natural sources of PAHs. Recently, boreal forest fires have been increasingly occurring with warming climate across the Arctic. On the other hand, forest fires can also contribute to climate warming by increasing greenhouse gas and black carbon emissions, decreasing surface albedo, and accelerating snow/ice melt,[8] which indirectly affect the environmental fate of PAHs.[33]

Figures S7 and S8 illustrate the summer and winter mean air temperatures and winds at 700 and 70 hPa, respectively,

DOI: 10.1021/acs.est.9b05000
Environ. Sci. Technol. 2019, 53, 13238−13245

averaged over 1980 to 2010. The most significant differences between the 700 hPa temperatures and winds and the values at 70 hPa are that the 700 hPa temperatures and winds illustrate clear atmospheric circulation patterns subject to large-scale underlying surface conditions and monsoons, whereas these two meteorological fields at 70 hPa exhibit a general atmospheric circulation pattern, featured by their more uniform distributions. In this sense, the air temperatures and winds at 70 hPa provide a large-scale background atmospheric circulation pattern that can be applied to predict a long-term change in BaP concentrations in the high Arctic. Accordingly, the mean air temperatures and winds at the low troposphere (700 hPa) reflect their mean characteristics as featured by both higher level atmospheric circulation and large-scale ground surface types. Among the selected meteorological variables, the SHAP absolute value indicates that the air temperature at 700 hPa is the most important factor contributing to monthly BaP fluctuations (Figure 5b); however, it makes a negative contribution to the variation in BaP concentration in terms of the SHAP mean value for the air temperature (Figure 5a), indicating that increasing temperature at 700 hPa can be associated with a declining BaP concentration near the surface of the Alert site. The low temperatures in the atmosphere favor the downward atmospheric motion and the dry deposition of particle-phase organic chemicals.[6,34] In our case, BaP is mostly in the particle phase and hence tends to be removed from the air by the atmospheric deposition. Figure S9 shows the vertical cross section of the air temperatures at 82°N from 0°E to 360°E. Relatively higher temperatures can be identified from 65°W to 45°W near the surface, corresponding to relatively lower temperatures at 700 hPa over this longitude range. This vertical temperature profile in the high Arctic near Alert indicates a typical stable atmospheric condition that favors descending atmospheric motion and the atmospheric removal of BaP. Likewise, the 700 hPa winds also provide a net negative contribution to BaP concentrations at the high Arctic Alert site (Figure 5a), indicating that stronger winds can decrease BaP concentrations. Although 700 hPa provides a more efficient atmospheric transport pathway for POPs,[21] it is common knowledge that stronger winds are always linked to the declining concentration of an air pollutant because the strong winds can more efficiently out-diffuse the air pollutant. It should be noted that stronger winds at the free troposphere (approximately 700 hPa) often predict stronger wind speeds near the surface due to the momentum downdraft.[35]

Both the sea ice and SSTs over the model domain make net positive contributions to BaP levels at the high Arctic (Figure 5a), but the sea ice exerts a stronger influence on BaP concentration (Figure 5b). This observation is expected because the sea ice is mostly accumulated in the Arctic, and organic chemicals show strong responses to the changes in Arctic sea ice.[36] Compared with the BaP emissions in the different regions, the wind and air temperature fields at the low and high atmospheres play significant roles in the changes in BaP air concentrations, as can be seen in Figure 5b. It is interesting to note that although Russia has the lowest BaP emission among the five emission regions examined in the present study (Figures S10–S12) and its net SHAP mean value after summing all positive and negative SHAP mean values is also low (Figure 5a), the BaP emission from Russia makes the largest contribution to the change in BaP concentration, as shown by its SHAP absolute value (Figure 5b). This result is likely a consequence of the atmospheric circulation over Russia providing a more efficient atmospheric transport route from the BaP sources to the high Arctic[37] (Figure S8).

In summary, we demonstrate that deep learning might serve as a very useful tool in the prediction of air concentrations of organic chemicals. Although the present study focuses on PAHs in the high Arctic, this method can be extended to any chemicals in any region as far as the sufficient long-term ambient concentration time series are available. Our results reveal that the RNN provides a better performance in forecasting high Arctic PAHs than does the complex atmospheric transport model CanMETOP. Using sampled PAH data from 1996 to 2012, the RNN successfully predicts PAH concentrations from 2012 to 2014, and the forecasting result agrees well with the sampled data. However, unlike an atmospheric transport and multimedia exchange model (which can be used to interpret physical, chemical, and dynamic processes in the environmental cycling of an organic chemical), the massive input data in machine learning lead to difficulties in facilitating the understanding of complex nonlinear systems and processes in data training. The heuristic knowledge about the relationships between input and output (forecasted) data and the selection of the right sensitivity analysis approaches can help to improve our understanding of the machine leaning results.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.9b05000.

Regions of selected PAH emission inventory; input meteorological data; plot of training data for 14 PAHs; recurrent neural networks; structure of custom RNN; shapely value; total BaP emissions from five regions; monthly grid emission of PHE and BaP (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jmma@pku.edu.cn. Tel/Fax: +86-10-62759639.

ORCID

Shu Tao: 0000-0002-7374-7063

Junfeng Liu: 0000-0002-7199-6357

Jianmin Ma: 0000-0002-6593-570X

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China through grant U1806207 and the National Key R&D Program of China through grant 2017YFC0212002.

■ REFERENCES

(1) National Research Council. *Global Sources of Local Pollution: An Assessment of Long-Range Transport of Key Air Pollutants to and from the United States*; National Academies Press, 2010.

(2) Hung, H.; Blanchard, P.; Halsall, C. J.; Bidleman, T. F.; Stern, G. A.; Fellin, P.; Muir, D. C. G.; Barrie, L. A.; Jantunen, L. M.; Helm, P. A.; Ma, J.; Konoplev, A. Temporal and Spatial Variabilities of Atmospheric Polychlorinated Biphenyls (PCBs), Organochlorine (OC) Pesticides and Polycyclic Aromatic Hydrocarbons (PAHs) in

the Canadian Arctic: Results from a Decade of Monitoring. *Sci. Total Environ.* **2005**, *342*, 119−144.
(3) Balmer, J. E.; Hung, H.; Yu, Y.; Letcher, R. J.; Muir, D. C. G. Sources and Environmental Fate of Pyrogenic Polycyclic Aromatic Hydrocarbons (PAHs) in the Arctic. *Emerging Contam.* **2019**, *5*, 128−142.
(4) Li, Y. F.; Macdonald, R. W. Sources and Pathways of Selected Organochlorine Pesticides to the Arctic and the Effect of Pathway Divergence on HCH Trends in Biota: A Review. *Sci. Total Environ.* **2005**, *342*, 87−106.
(5) Friedman, C. L.; Selin, N. E. Long-Range Atmospheric Transport of Polycyclic Aromatic Hydrocarbons: A Global 3-D Model Analysis Including Evaluation of Arctic Sources. *Environ. Sci. Technol.* **2012**, *46*, 9501−9510.
(6) Ma, J.; Hung, H.; Macdonald, R. W. The Influence of Global Climate Change on the Environmental Fate of Persistent Organic Pollutants: A Review with Emphasis on the Northern Hemisphere and the Arctic as a receptor. *Global Planet. Change* **2016**, *146*, 89−108.
(7) Yu, Y.; Katsoyiannis, A.; Bohlin-Nizzetto, P.; Brorström-Lundén, E.; Ma, J.; Zhao, Y.; Wu, Z.; Tych, W.; Mindham, D.; Sverko, E.; Barresi, E.; Dryfhout-Clark, H.; Fellin, P.; Hung, H. Polycyclic Aromatic Hydrocarbons Not Declining in Arctic Air Despite Global Emission Reduction. *Environ. Sci. Technol.* **2019**, *53*, 2375−2382.
(8) Randerson, J. T.; Liu, H.; Flanner, M. G.; Chambers, S. D.; Jin, Y.; Hess, P. G.; Pfister, G.; Mack, M. C.; Treseder, K. K.; Welp, L. R.; Chapin, F. S.; Harden, J. W.; Goulden, M. L.; Lyons, E.; Neff, J. C.; Schuur, E. A. G.; Zender, C. S. The Impact of Boreal Forest Fire on Climate Warming. *Science* **2006**, *314*, 1130−1132.
(9) Rogers, B. M.; Soja, A. J.; Goulden, M. L.; Randerson, J. T. Influence of Tree Species on Continental Differences in Boreal Fires and Climate Feedbacks. *Nat. Geosci.* **2015**, *8*, 228−234.
(10) Bonan, G. B. Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests. *Science* **2008**, *320*, 1444−1449.
(11) Hung, H.; Katsoyiannis, A. A.; Brorström-Lundén, E.; Olafsdottir, K.; Aas, W.; Breivik, K.; Bohlin-Nizzetto, P.; Sigurdsson, A.; Hakola, H.; Bossi, R.; Skov, H.; Sverko, E.; Barresi, E.; Fellin, P.; Wilson, S. Temporal Trends of Persistent Organic Pollutants (POPs) in Arctic Air: 20 Years of Monitoring under the Arctic Monitoring and Assessment Programme (AMAP). *Environ. Pollut.* **2016**, *217*, 52−61.
(12) Kanevski, M.; Parkin, R.; Pozdnukhov, A.; Timonin, V.; Maignan, M.; Demyanov, V.; Canu, S. Environmental Data Mining and Modeling Based on Machine Learning Algorithms and Geostatistics. *Environ. Model. Softw.* **2004**, *19*, 845−855.
(13) Robert, C. Machine Learning, a Probabilistic Perspective. *CHANCE* **2014**, *27*, 62−63.
(14) Hu, X.; Belle, J. H.; Meng, X.; Wildani, A.; Waller, L. A.; Strickland, M. J.; Liu, Y. Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environ. Sci. Technol.* **2017**, *51*, 6936−6944.
(15) Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep Learning and Process Understanding for Data-Driven Earth System Science. *Nature* **2019**, *566*, 195−204.
(16) Kallenborn, R.; Hung, H.; Harner, T.; Bohlin-Nizzetto, P.; Nash, S. B. Research and Monitoring of Atmospheric Persistent Organic Pollutants (POPs) in the Polar Atmosphere. In *Implications and Consequences of Anthropogenic Pollution in Polar*;Kallenborn, R., Ed., From Pole to Pole; Springer: Berlin, 2016; pp 5−19.
(17) Giglio, L.; Boschetti, L.; Roy, D.; Hoffmann, A. A.; Humber, M.; Hall, J. V. *Collection 6 MODIS Burned Area Product User's Guide*, version 1.2.; NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, 2016.
(18) Shen, H. Global Atmospheric Emissions of PAH Compounds. In *Polycyclic Aromatic Hydrocarbons: Their Global Atmospheric Emissions, Transport, and Lung Cancer Risk*; Shen, H., Ed.; Springer: Berlin, 2016; pp 85−119.
(19) Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J.; Zhu, Y.; Leetmaa, A.; Reynolds, R. The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Am. Meteorol. Soc.* **1996**, *77*, 437−471.
(20) Zhang, L.; Ma, J.; Venkatesh, S.; Li, Y.-F.; Cheung, P. Modeling Evidence of Episodic Intercontinental Long-Range Transport of Lindane. *Environ. Sci. Technol.* **2008**, *42*, 8791−8797.
(21) Zhang, L.; Ma, J.; Tian, C.; Li, Y.; Hung, H. Atmospheric Transport of Persistent Semi-Volatile Organic Chemicals to the Arctic and Cold Condensation in the Mid-Troposphere − Part 2: 3-D Modeling of Episodic Atmospheric Transport. *Atmos. Chem. Phys.* **2010**, *10*, 7315−7324.
(22) Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J.; Zhu, Y.; Chelliah, M.; Ebisuzaki, W.; Higgins, W.; Janowiak, J.; Mo, K. C.; Ropelewski, C.; Wang, J.; Leetmaa, A.; Reynolds, R.; Jenne, R.; Joseph, D. The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Am. Meteorol. Soc.* **1996**, *77*, 437−471.
(23) Arlot, S.; Celisse, A. A Survey of Cross-Validation Procedures for Model Selection. *Stat. Surv.* **2010**, *4*, 40−79.
(24) Li, R.; Mao, H.; Wu, L.; He, J.; Ren, P.; Li, X. The Evaluation of Emission Control to PM Concentration during Beijing APEC in 2014. *Atmos. Pollut. Res.* **2016**, *7*, 363−369.
(25) Gers, F. A.; Schmidhuber, J.; Cummins, F. *Learning to Forget: Continual Prediction with LSTM*, 9th International Conference on Artificial Neural Networks: ICANN '99, 1999; pp 850−855.
(26) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.
(27) Prechelt, L. Early Stopping - But When?. In *Neural Networks: Tricks of the Trade*; Springer, 1998; pp 55−69.
(28) Barbounis, T. G.; Theocharis, J. B.; Alexiadis, M. C.; Dokopoulos, P. S. Long-Term Wind Speed and Power Forecasting Using Local Recurrent Neural Network Models. *IEEE Trans. Energy Convers.* **2006**, *21*, 273−284.
(29) Colin Cameron, A.; Windmeijer, F. A. G. An R-Squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models. *J. Econom.* **1997**, *77*, 329−342.
(30) Zhang, Y.; Shen, H.; Tao, S.; Ma, J. Modeling the Atmospheric Transport and Outflow of Polycyclic Aromatic Hydrocarbons Emitted from China. *Atmos. Environ.* **2011**, *45*, 2820−2827.
(31) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions, ArXiv1705.07874 Cs Stat, 2017.
(32) Li, C.; Kang, S.; Chen, P.; Zhang, Q.; Fang, G. C. Characterizations of Particle-Bound Trace Metals and Polycyclic Aromatic Hydrocarbons (PAHs) within Tibetan Tents of South Tibetan Plateau, China. *Environ. Sci. Pollut. Res.* **2012**, *19*, 1620−1628.
(33) Halsall, C. J.; Barrie, L. A.; Fellin, P.; Muir, D. C. G.; Billeck, B. N.; Lockhart, L.; Rovinsky, F. Y.; Kononov, E. Y.; Pastukhov, B. Spatial and Temporal Variation of Polycyclic Aromatic Hydrocarbons in the Arctic Atmosphere. *Environ. Sci. Technol.* **1997**, *31*, 3593−3599.
(34) Wania, F.; Mackay, D. Global Fractionation and Cold Condensation of Low Volatility Organochlorine Compounds in Polar Regions. *Ambio* **1993**, *22*, 10−18.
(35) Booth, J. F.; Thompson, L. A.; Patoux, J.; Kelly, K. A.; Dickinson, S. The Signature of the Midlatitude Tropospheric Storm Tracks in the Surface Winds. *J. Clim.* **2010**, *23*, 1160−1174.
(36) Zhao, Y.; Huang, T.; Wang, L.; Gao, H.; Ma, J. Step Changes in Persistent Organic Pollutants over the Arctic and Their Implications. *Atmos. Chem. Phys.* **2015**, *15*, 3479−3495.
(37) Halsall, C. J.; Sweetman, A. J.; Barrie, L. A.; Jones, K. C. Modelling the Behaviour of PAHs during Atmospheric Transport from the UK to the Arctic. *Atmos. Environ.* **2001**, *35*, 255−267.

DOI: 10.1021/acs.est.9b05000
Environ. Sci. Technol. 2019, 53, 13238−13245