Long-term trends of ambient nitrate (NO$_3^-$) concentrations across China based on ensemble machine-learning models

Rui Li$^1$, Lulu Cui$^1$, Yilong Zhao$^1$, Wenhui Zhou$^1$, and Hongbo Fu$^{1,2,3}$

$^1$Shanghai Key Laboratory of Atmospheric Particle Pollution and Prevention, Department of Environmental Science & Engineering, Institute of Atmospheric Sciences, Fudan University, Shanghai, 200433, China
$^2$Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China
$^3$Shanghai Institute of Pollution Control and Ecological Security, Shanghai 200092, China

Correspondence: Hongbo Fu (fuhb@fudan.edu.cn) and Lulu Cui (15110740004@fudan.edu.cn)

Received: 18 August 2020 – Discussion started: 26 November 2020
Revised: 27 February 2021 – Accepted: 22 March 2021 – Published: 19 May 2021

Abstract. High loadings of nitrate (NO$_3^-$) in the aerosol over China significantly exacerbate the air quality and pose a great threat to ecosystem safety through dry–wet deposition. Unfortunately, limited ground-level observation data make it challenging to fully reflect the spatial pattern of NO$_3^-$ levels across China. Until now, long-term monthly particulate NO$_3^-$ datasets at a high resolution were still missing, which restricted the assessment of human health and ecosystem safety. Therefore, a unique monthly NO$_3^-$ dataset at 0.25° resolution over China during 2005–2015 was developed by assimilating surface observations, satellite products, meteorological data, land use types and other covariates using an ensemble model combining random forest (RF), gradient-boosting decision tree (GBDT), and extreme gradient-boosting (XGBoost) methods. The new developed product featured an excellent cross-validation $R^2$ value (0.78) and relatively lower root-mean-square error (RMSE: 1.19 µg N m$^{-3}$) and mean absolute error (MAE: 0.81 µg N m$^{-3}$). Besides, the dataset also exhibited relatively robust performance at the spatial and temporal scales. Moreover, the dataset displayed good agreement with ($R^2 = 0.85$, RMSE = 0.74 µg N m$^{-3}$, and MAE = 0.55 µg N m$^{-3}$) some unlearned data collected from previous studies. The spatiotemporal variations in the developed product were also shown. The estimated NO$_3^-$ concentration showed the highest value in the North China Plain (NCP) (3.55 ± 1.25 µg N m$^{-3}$); followed by the Yangtze River Delta (YRD) (2.56 ± 1.12 µg N m$^{-3}$), Pearl River Delta (PRD) (1.68 ± 0.81 µg N m$^{-3}$), and Sichuan Basin (1.53 ± 0.63 µg N m$^{-3}$), and the lowest one in the Tibetan Plateau (0.42 ± 0.25 µg N m$^{-3}$). The higher ambient NO$_3^-$ concentrations in the NCP, YRD, and PRD were closely linked to the dense anthropogenic emissions. Apart from the intensive human activities, poor terrain condition might be a key factor for the serious NO$_3^-$ pollution in the Sichuan Basin. The lowest ambient NO$_3^-$ concentration in the Tibetan Plateau was contributed by the scarce anthropogenic emission and favourable meteorological factors (e.g. high wind speed). In addition, the ambient NO$_3^-$ concentration showed a marked increasing tendency of 0.10 µg N m$^{-3}$ yr$^{-1}$ during 2005–2014 ($p < 0.05$), while it decreased sharply from 2014 to 2015 at a rate of $-0.40$ µg N m$^{-3}$ yr$^{-1}$ ($p < 0.05$). The ambient NO$_3^-$ levels in Beijing–Tianjin–Hebei (BTH), YRD, and PRD displayed gradual increases at a rate of 0.20, 0.11, and 0.05 µg N m$^{-3}$ yr$^{-1}$ ($p < 0.05$) during 2005–2013, respectively. The gradual increases in NO$_3^-$ concentrations in these regions from 2005 to 2013 were due to the fact that the emission reduction measures during this period focused on the reduction of SO$_2$ emission rather than NO$_x$ emission and the rapid increase in energy consumption. Afterwards, the government further strengthened these emission reduction measures and thus caused the dramatic decreases in NO$_3^-$ concentrations in these regions from 2013 to 2015 ($p < 0.05$). The long-term NO$_3^-$ dataset over China could greatly deepen the knowledge about the impacts of emission reduction measures on
1 Introduction

Reactive nitrogen (N\textsubscript{r}) emissions displayed remarkable increases in the past decades owing to high-speed industrial development and urbanisation (Cui et al., 2016; Singh et al., 2017). Ambient reactive N emissions were mainly characterised with nitrogen oxides (NO\textsubscript{x}), accounting for about 30% of the gross N\textsubscript{r} emissions (Chen et al., 2015; Liu et al., 2011). These important N-bearing precursors could be transformed into the nitrate (NO\textsubscript{3}\textsuperscript{-}) via multiple chemical pathways (e.g., heterogeneous or liquid phase reaction) and finally deposited in the terrestrial or aquatic ecosystem (Jia et al., 2016; Qiao et al., 2015; Zhao et al., 2017). On the one hand, heavy loadings of NO\textsubscript{3}\textsuperscript{-} greatly degraded the atmospheric visibility and cool the surface of the Earth system because particulate NO\textsubscript{3} significantly scattered solar radiation (Fu and Chen, 2017). Moreover, enhanced N deposition might pose a negative effect on the ecosystem health such as biodiversity losses, freshwater eutrophication, and oceanic acidification (Compton et al., 2011; Erisman et al., 2013). Hence, deepening knowledge about the spatial patterns and long-term trends of particulate NO\textsubscript{3} in the atmosphere is beneficial to accurately evaluate the ecological and environmental effects of N deposition.

Ground-level observation is often acknowledged to be an effective means to explore the spatial patterns of ambient NO\textsubscript{3}\textsuperscript{-} concentrations. Many long-term monitoring networks including the Clean Air Status and Trends Network (CASTNET) and the Canadian Air and Precipitation Monitoring Network (CAPMoN) were established to quantify the ambient NO\textsubscript{3}\textsuperscript{-} concentration and inorganic N deposition. Du et al. (2014) revealed that the NO\textsubscript{3} deposition showed significant decrease across the United States during 1985–2012 based on these observation data. To date, most of these observation networks focused on North America and Europe, whereas few monitoring sites were located in East Asia, especially in China. Fortunately, China has constructed some ground-level observation networks such as the CARE-China observation network in recent years. On the basis of these observation networks, the overall spatiotemporal trend of particulate NO\textsubscript{3} concentration has been clarified (Y. Wang et al., 2019b; Xu et al., 2018a). Xu et al. (2018a) observed that the particulate NO\textsubscript{3} concentration (< 4.5 µm) over China did not show significantly temporal variation during 2011–2015. Very recently, Y. Wang et al. (2019a) found that the NO\textsubscript{3} level in fine particulate matter (PM\textsubscript{2.5}) decreased by 34% during 2015–2017. Although the overall spatial patterns have been preliminarily revealed based on these isolated sites, these sparse ground-observed sites did not accurately reflect the high-resolution NO\textsubscript{3} pollution, especially the regions far away from these sites, because each station only possessed limited spatial representation, and NO\textsubscript{3} concentration was often highly variable in space and time (L. Liu et al., 2017). More importantly, the current studies only investigated the ambient NO\textsubscript{3} concentrations in recent years, while the long-term variation in NO\textsubscript{3} level remained unknown. It was well known that the energy consumption in China has displayed a remarkable increase in recent decades (Zhan et al., 2018a). Meanwhile, the Chinese government has also proposed pollutant emission reduction policies since 2005 to ensure the coordinated development of economic growth and environmental protection (Ma et al., 2019). However, the synergistic effects of air pollution control policies and increased energy consumption on the long-term evolution trend of NO\textsubscript{3} pollution over China, which were extremely critical for the implementation of emission control measures, were not assessed yet.

To complement the gaps of ground-level observations, the satellite product of NO\textsubscript{2} is regarded as a welcome addition to investigate the long-term trends of N-bearing components in the atmosphere. Ozone Monitoring Instrument (OMI) was regarded as the typical satellite product applied to simulate the ambient NO\textsubscript{3}\textsuperscript{-} concentration (X. Liu et al., 2017; Vrekoussis et al., 2013). Jia et al. (2016) first used the linear regression method to predict the NO\textsubscript{3}\textsuperscript{-} levels and dry deposition fluxes at the global scale based on OMI-derived NO\textsubscript{2} column amount. However, the dry deposition fluxes of NO\textsubscript{3} modelled by Jia et al. (2016) showed weak correlation with the measured value ($R = 0.47$), which might be attributable to the simple linear assumption between NO\textsubscript{2} column amount and NO\textsubscript{3}\textsuperscript{-} deposition flux. It was well documented that the nonlinearity relationship between multiple predictors and NO\textsubscript{3}\textsuperscript{-} concentration was hard to reveal on the basis of the simple linear model (Zhan et al., 2018a, b). To enhance the predictive performance of NO\textsubscript{3}\textsuperscript{-} concentration, L. Liu et al. (2017) used the chemical transport models (CTMs) to estimate the dry deposition fluxes of N-bearing species recently based on the remotely sensed NO\textsubscript{2} column amount. However, CTMs often suffered from high uncertainty because of the limited knowledge about the generation pathways for particulate NO\textsubscript{3} in the atmosphere (Zhan et al., 2018a). Recently, the emergence of machine learning models provided unprecedented opportunities to estimate the concentrations of N-bearing components (Z. Y. Chen et al., 2019; Zhan et al., 2018b). It was well known that the machine-learning models generally showed better predictive accuracy than CTMs and traditional statistical models when the training samples were sufficient (Zang et al., 2019; Zhan et al., 2017). Zhan...
et al. (2018b) employed random forest (RF) coupled with a spatiotemporal kriging model to simulate the ambient NO$_3^-$ levels over China and achieved moderate modelling performance ($R^2 = 0.62$). Afterwards, J. Chen et al. (2019) used the extreme gradient-boosting (XGBoost) model combined with the kriging-calibrated satellite method to estimate the national NO$_3^-$ concentration and significantly improved the predictive performance ($R^2 = 0.85$). Until now, no study has utilised machine-learning models to significantly improve predictive accuracy of NO$_3^-$ concentration. Moreover, nearly all of the current studies only focused on the spatial pattern of particulate NO$_3^-$ level in China (L. Liu et al., 2017; Jia et al., 2016), while they cannot establish a long-term NO$_3^-$ dataset across China.

Here, we first developed a high-resolution (0.25°) monthly particulate NO$_3^-$ dataset across China during 2005–2015 based on an ensemble model including RF, XGBoost, and gradient-boosting decision tree (GBDT) algorithms. At first, the modelling performance and improvement of this newly developed product compared with previous datasets were evaluated. Afterwards, we analysed the spatial variation and long-term evolution trend of estimated NO$_3^-$ concentration over China and explored the potential impacts of air pollution control measures on NO$_3^-$ variation. The long-term NO$_3^-$ datasets could supply scientific evidence for policy makers to mitigate the severe nitrate pollution in China.

2 Input data

2.1 Ground-level NO$_3^-$ data

The monthly NO$_3^-$ monitoring data during 2010–2015 were collected from the Nationwide Nitrogen Deposition Monitoring Network (NNDMN) including 32 sites (Figs. 1, S1, and S2 in the Supplement), and these sites could be divided into three types including urban, rural, and background sites (Xu et al., 2018a). Ambient concentrations of particulate NO$_3^-$ were determined on the basis of an active DEnduer for Long-Term Atmospheric sampling system (DELTA). The system comprises a pump, a filter sampling instrument, and a dry gas meter with high sensitivity. Two sets of filters in a two-stage filter pack were applied to sample the aerosol particles, first with a K$_2$CO$_3$/glycerol-impregnated filter to obtain NO$_3^-$ particles in PM$_{10}$. All of the monitoring sites kept the same sampling frequency at the month scale, and these samples were continuously collected over a month. The detailed sampling and analysis procedures have been described by Xu et al. (2018a, 2019). The detection limit of particulate NO$_3^-$ concentration over China is 0.01 mg NL$^{-1}$.

2.2 Satellite product of NO$_2$ column density

The OMI-NO$_2$ level-3 tropospheric column densities (0.25° resolution) were used to predict the NO$_3^-$ concentration (Fig. S3 in the Supplement). The OMI aboard the Aura satellite has been available since September 2004, which displayed global coverage and crossed the entire Earth each day. OMI possessed three spectral channels ranging from 270 to 500 nm and thus was often applied to monitor gaseous pollutants such as NO$_2$, SO$_2$, and O$_3$.

In this study, we downloaded the daily NO$_2$ columns during 2005–2015 from https://earthdata.nasa.gov/ (last access: 31 July 2020). The tropospheric NO$_2$ column density data of poor quality (e.g. cloud radiance fraction > 0.5, solar zenith angles > 85°, and terrain reflectivity > 30 %) should be removed. Additionally, the cross-track pixels sensitive to significant row anomalies must also be deleted. Finally, the monthly NO$_2$ columns were estimated by averaging the daily NO$_2$ columns.

2.3 Meteorological factors, land use types, and other variables

These independent variables for particulate NO$_3^-$ estimates were gained from multiple sources. The daily meteorological data (European Centre for Medium-Range Weather Forecasts reanalysis (ECMWF ERA-Interim) datasets (0.25° resolution)) were downloaded from the website of http://www.ecmwf.int/ (last access: 31 July 2020) (Table S1 in the Supplement). Among all of the daily meteorological data in ECMWF, 2 m temperature ($T_{2m}$), 2 m dew point temperature ($D_{2m}$), 10 m latitudinal wind component ($U_{10}$), 10 m meridional wind component ($V_{10}$), sunshine duration (Sund), surface pressure (Sp), boundary layer height (BLH), and total precipitation (Tp) were applied to estimate national NO$_3^-$ levels. The elevation, gross domestic production (GDP), and population density (PD) data over China were downloaded from the website of http://www.resdc.cn/ (last access: 31 July 2020). PD and GDP in 1995, 2000, 2005, 2010, and 2015 were linearly interpolated to calculate PD and GDP in each year. Then, the yearly GDP data were divided by 12
to estimate the monthly GDP. Afterwards, these data were incorporated into the sub-model to predict the particulate NO$_x$ concentration over China. In addition, the land use data (e.g. grassland, forest, urban, and agricultural land) were also downloaded from the website of http://www.resdc.cn/.

These independent variables collected from various sources were uniformly resampled to $0.25^\circ \times 0.25^\circ$ grids. For instance, the land use area, GDP, and PD in the $0.25^\circ$ grid were calculated based on the area-weighted average algorithm. To ensure better predictive performance, it was necessary to employ the appropriate variable selection method to remove some redundant predictors. The basic principle of the variable choice was to remove the variables with the lower importance values. The variables could be regarded as the redundant ones when the $R^2$ value of the final model showed a dramatic decrease after removing them. Based on this method, in the final sub-model, all of the variables except GDP, PD, and grassland have been applied to estimate the ambient NO$_x$ concentrations across China.

3 Methods

3.1 Ensemble model development

In the previous studies concerning air pollution prediction, RF, gradient-boosting decision tree (GBDT) and extreme gradient-boosting (XGBoost) showed good predictive performance (Li et al., 2020a). The RF model possesses a large gradient-boosting (XGBoost) showed good predictive performance (Li et al., 2019), XGBoost generally shows an excellent performance (Breiman, 2001). This model generally shows the higher prediction accuracy due to the injected randomness. The model performance mainly relies on the number of trees, the variable group, and the splitting features. The detailed algorithms are shown as follows.

$$ f(x) = \sum_{i=1}^{Z} c_i I(x \in M_i) $$ (1)

$$ c_i = \text{mean}(y_i | x_i \in M_i) $$ (2)

$$ L_1(m, n) = \{X | X_j \leq n\} \text{ and } L_2(m, n) = \{X | X_j > n\} $$ (3)

$$ \Delta c_1 = \text{mean}(y_i | x_i \in M_1(m, n)) \quad \text{and} \quad \Delta c_2 = \text{mean}(y_i | x_i \in M_2(m, n)) $$ (5)

Here $(x_i, y_i)$ denotes the sample for $i = 1, 2, \ldots, N$ in $M$ regions $(M_1, M_2, \ldots, M_z)$; $I$ denotes the weight of each branch; $L$ denotes the branch of decision tree; $c_m$ represents the response to the model; $\Delta c_2$ denotes the best value, $m$ represents the feature variable; $c_1$ denotes the mean value of the left branch; $c_2$ denotes the mean value of the right branch and $n$ is the split point.

The GBDT model is often considered to be a typical boosting method. Compared with the RF model, each classifier is applied to decrease the residual of the last round. The detailed equations are as follows.

$$ c_{tj} = \argmin_{y \in \mathcal{R}_{tj}} \sum_{i \in \mathcal{R}_{tj}} L(y_i, f_{t-1}(x_i) + c) $$ (6)

$$ f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj} I $$ (7)

$c_{tj}$ denotes the predicted estimation error in the last round, $R_{tj}$ denotes each leaf node for the decision trees, $y_i$ represents the observed value, and $f_{t-1}(x_i)$ is the predicted value in the last round. $c$ was regarded as the optimal value when $c_{tj}$ reaches the lowest value.

The XGBoost method is an updated version of the GBDT model, and loss functions are expanded to the second-order function. On the basis of the pioneering studies (J. Chen et al., 2019), XGBoost generally shows an excellent performance because of its high efficiency and impressive accuracy. The detailed XGBoost algorithm is shown as the following formula (Zhai and Chen, 2018):

$$ L^{(i)} = \sum_{i=1}^{N} \left[ l\left(y_i, f_t(x_i)\right) + \frac{1}{2} \sum_{j=1}^{J} \left( \frac{y_i - f_t(x_i)}{\frac{y_i - f_t(x_i)}{1}} \right)^2 \right] + \Omega(f_t) $$ (8)

where $L^{(i)}$ represents the cost function at the $i$th period, $\partial$ denotes the derivative of the function, $\frac{\partial^2}{\partial y^{(i-1)}}$ denotes the second derivative of the function, $l$ is the differentiable convex loss function that reveals the difference of the predicted value $(\hat{y})$ of the $i$th instance at the $i$th period and the target value $(y_i)$, $f_t(x)$ denotes the increment, and $\Omega(f_t)$ represents the regulariser.

However, each model still shows some disadvantages in the prediction accuracy. Consequently, it was proposed to combine these models with a multiple linear regression (MLR) model to further estimate monthly NO$_x$ concentration in the atmosphere over China. As shown in Fig. 2, three submodels including RF, GBDT, and XGBoost were stacked through the MLR model to estimate the monthly NO$_x$ concentration over China. At first, a fivefold cross-validation method was adopted to train each submodel to determine the appropriate parameter. Afterwards, the MLR model was trained with the final simulated concentrations of three submodels and observations. Finally, the high-resolution ambient NO$_x$ levels over China were estimated based on the optimal ensemble model. The detailed algorithms are shown as
3.1 The error estimation and uncertainty assessment

The estimation performance of the ensemble model was evaluated based on a 10-fold cross-validation algorithm. The principle of this method meant that the entire datasets were divided into 10 groups with the same capacity randomly. Nine groups were applied to develop the model, and the remaining one was used to predict the NO$_3^-$ level. After 10 rounds, every observed NO$_3^-$ concentration showed a corresponding predicted value. Some key indices such as determination coefficient ($R^2$), root-mean-square error (RMSE), and mean absolute prediction error (MAE) were selected as the key indicators to identify the optimal modelling method.

The uncertainty of the ensemble model was mainly derived from input ancillary variables. For instance, both the satellite data and meteorological data often suffered from some uncertainties. To quantify the uncertainties derived from meteorological data, the meteorological data at 0.25$^\circ$ across China were validated using ground-measured meteorological data downloaded from the website of the Chinese Meteorology Bureau (http://data.cma.cn/), last access: 31 July 2020). Additionally, NO$_2$ columns generally suffered from some uncertainties, whereas the uncertainties of these NO$_2$ columns cannot be determined because the data about the ground-level NO$_2$ columns were not open access. In our study, we only estimated the missing ratio of the NO$_2$ column, thereby evaluating the uncertainty of the NO$_3^-$ dataset.

3.3 Trend analysis

The trend analysis of particulate NO$_3^-$ concentration was performed using the Mann–Kendall nonparametric test. This method has been widely applied to analyse the historical trends of carbon fluxes (Tang et al., 2019) and air quality (Kong et al., 2019), which could reflect whether these data suffered from significant changes at a significance level of 0.05. The detailed calculation process is summarised in Mann (1945) and Kendall (1975).

4 Results and discussion

4.1 Descriptive statistics of observed NO$_3^-$ concentrations

The ensemble model was applied to fit the NO$_3^-$ estimation model based on 1636 matched samples across China during 2010–2015. In general, the ground-observed NO$_3^-$ concentration over China ranged from 0.3 µg N m$^{-3}$ in Bayinbrook of Xinjiang Province to 7.1 µg N m$^{-3}$ in Zhengzhou of Henan Province with the mean value of 2.7 ± 1.7 µg N m$^{-3}$. The monthly particulate NO$_3^-$ concentrations displayed the highest and lowest values in the North China Plain (NCP) and Tibetan Plateau, respectively. In addition, the monthly NO$_3^-$ level exhibited significant temporal variation during 2010–2015. The ambient NO$_3^-$ concentrations in most sites displayed a gradual increase during 2010–2014, while they decreased sharply from 2014 to 2015. The spatiotemporal variation in ambient NO$_3^-$ concentration over China shared similar characteristics with NO$_2$ column amount (Fig. S3). The Pearson correlation analysis revealed that the monthly particulate NO$_3^-$ level showed a significantly positive relationship with NO$_2$ column amount ($r = 0.57$, $p < 0.01$) and urban land area ($r = 0.35$, $p < 0.05$) (Fig. S4 in the Supplement). However, $D_{200}$ showed the remarkably negative correlation with ambient NO$_3^-$ concentration ($r = -0.31$, $p < 0.05$).

4.2 The validation of newly developed NO$_3^-$ dataset and comparison with previous products

In our study, the ensemble model was applied to develop a monthly particulate NO$_3^-$ dataset over China based on various predictors. In addition, three other individual models were also trained to compare with their predictive performances. The cross-validation result indicated that the $R^2$ value of the new product developed by an ensemble decision tree model reached 0.78, significantly higher
than those developed by RF (0.57), GBDT (0.73), and XGBoost (0.45). Nonetheless, both RMSE and MAE exhibited the opposite trends. The RMSE value was in the order of XGBoost (1.98 µg N m⁻³) > RF (1.67 µg N m⁻³) > GBDT (1.35 µg N m⁻³) > ensemble model (1.19 µg N m⁻³). The MAE value followed a similar characteristic with the order of XGBoost (1.29 µg N m⁻³) > RF (0.99 µg N m⁻³) > GBDT (0.95 µg N m⁻³) > ensemble model (0.81 µg N m⁻³). In some previous studies (Xiao et al., 2018), XGBoost often showed the better performance compared with RF, which seemed to be in contrast to our study. It was assumed that XGBoost showed the better performance for big data samples. However, the size of training samples in our study was relatively smaller than that in previous studies. Xiao et al. (2018) also verified that the XGBoost showed better accuracy than RF in some developed regions such as east China, while RF showed better performance than XGBoost in northwest China because monitoring sites in northwest China were relatively scarce. Wolpert (1992) suggested the combination of various machine-learning models can significantly strengthen the transferability of models. J. Chen et al. (2019) demonstrated that the ensemble model significantly outperformed the individual machine-learning model because the ensemble model can overcome the weaknesses of individual model. In addition, we also assessed the annual modelling performance of NO₃⁻ estimation. Figure S5 in the Supplement shows that the R² value of annual NO₃⁻ estimation reached 0.81, slightly higher than monthly NO₃⁻ prediction (0.78). However, both RMSE (1.23 µg N m⁻³) and MAE (0.85 µg N m⁻³) for annual NO₃⁻ estimation were slightly higher than those of monthly NO₃⁻ prediction.

The new developed NO₃⁻ dataset showed the marked temporal discrepancy. The R² values of NO₃⁻ estimates during 2011–2015 (0.88, 0.89, 0.83, 0.74, and 0.78) were notably higher than those during 2010 (0.62) (Table 1 and Fig. 3). The relatively lower R² value in 2010 attested to the dominant role of sampling size in the predictive accuracy for machine-learning models. The training samples in 2010 (135 samples) were notably less than those in other years due to the lack of observation data in spring. However, both RMSE and MAE were not sensitive to the sampling size. The higher RMSE and MAE were observed in 2010, 2014, and 2015. The higher RMSE and MAE observed in 2010 might be contributed by the relatively scarce training samples, while the higher RMSE and MAE likely pertained to the higher NO₃⁻ levels during other years. In addition, the performance of the NO₃⁻ dataset varied greatly at the seasonal scale. The R² value was in the order of summer (0.85) > spring (0.80) = autumn (0.80) > winter (0.75) across China (Table 2). The seasonal variation in NO₃⁻ concentration was in contrast to the results of fine particles modelled by previous studies (Li et al., 2020a; Qin et al., 2018). It was supposed that aerosol optical depth (AOD) was sensitive to the precipitation and relative humidity and thus showed the worse performance in summer. However, the predictive accuracy of NO₃⁻ estimation based on NO₂ column amount was closely linked with the chemical transformation from NO₂ to NO₃⁻.

The performance of the NO₃⁻ dataset also displayed marked spatial variation. The highest R² value was observed in the NCP (0.70), followed by southwest China (0.60), southeast China (0.59), and northwest China (0.55), and the lowest one was in northeast China (0.44) (Table 3). The highest R² value occurring in the NCP was mainly attributable to the largest training samples (> 400) compared with other regions. Southeast China and southwest China showed satisfactory cross-validation R² values because the valid training samples in both of these regions were higher than 300. Although both northeast China and northwest China possessed limited training samples (< 200), the predictive performances of these regions showed a significant discrepancy. It was assumed that the sampling sites in northeast China were very centralised, while the sampling sites in northwest China were uniformly distributed across the whole region. Geng et al. (2018) revealed that the modelling accuracy based on statistical models was significantly affected by the distribution characteristics of sampling sites. However, both RMSE and MAE showed different spatial distributions with the R² value and slope of fitting curve. Note that the higher values of RMSE and MAE were concentrated in southwest China (2.08 and 1.41 µg N m⁻³) and northwest China (2.06 and 1.38 µg N m⁻³) rather than the NCP (1.74 and 1.06 µg N m⁻³). There are two reasons responsible for the result. At first, the predictive performances of southwest China and northwest China were significantly worse than that of the NCP, thereby leading to the higher RMSE and MAE. Moreover, most of the sampling sites in southwest China were focused on the Sichuan Basin, which often showed severe NO₃⁻ pollution all year round. Meanwhile, the annual mean NO₃⁻ concentrations in Yangling and Wuwei reached 4.1 and 4.5 µg N m⁻³, respectively. The higher loadings of NO₃⁻ concentrations for training samples led to the higher RMSE and MAE for northwest China.

Although the cross-validation result suggested the newly developed dataset achieved better modelling accuracy, the cross-validation algorithm could not test the transferability and agreement of this dataset in past years. Hence, the unlearned data (annual mean NO₃⁻ concentration in 10 cities) collected from previous references were employed to validate the transferability of this product. As shown in Fig. 4 and Table S2 in the Supplement, we found that the R² value of the newly developed NO₃⁻ product and historical data reached 0.85 (Fig. 4), and the out-of-range R² value was even slightly higher than the cross-validation R² value. Moreover, the out-of-bag slope based on these unlearning data reached 0.81 and equaled the slope of the cross-validation database. In addition, the site-based cross-validation was also applied to validate the transferability of this dataset. The basic principle is that all of the sites were evenly classified into 10 clusters based on the geographical locations. Afterwards, 9 of 10 were used to train the model and then test the model...
Table 1. The cross-validation result of NO$_3^-$ estimation over China during 2010–2015.

<table>
<thead>
<tr>
<th>Year</th>
<th>Sample size</th>
<th>$R^2$ value</th>
<th>Slope</th>
<th>RMSE (µg N m$^{-3}$)</th>
<th>MAE (µg N m$^{-3}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>135</td>
<td>0.62</td>
<td>0.60</td>
<td>1.39</td>
<td>0.90</td>
</tr>
<tr>
<td>2011</td>
<td>291</td>
<td>0.88</td>
<td>0.85</td>
<td>0.32</td>
<td>0.24</td>
</tr>
<tr>
<td>2012</td>
<td>274</td>
<td>0.89</td>
<td>0.86</td>
<td>0.33</td>
<td>0.28</td>
</tr>
<tr>
<td>2013</td>
<td>312</td>
<td>0.83</td>
<td>0.82</td>
<td>0.64</td>
<td>0.43</td>
</tr>
<tr>
<td>2014</td>
<td>306</td>
<td>0.74</td>
<td>0.76</td>
<td>1.50</td>
<td>1.04</td>
</tr>
<tr>
<td>2015</td>
<td>318</td>
<td>0.78</td>
<td>0.78</td>
<td>1.35</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Figure 3. Density scatterplots of 10-fold cross-validation results for monthly NO$_3^-$ estimation (unit: µg N m$^{-3}$) across China for the ensemble decision tree model (a), RF (b), GBDT (c), and XGBoost (d). The colour bar reflects the sampling size of each model. The red solid line denotes the best-fit line through the data points (1636 points). The black dashed line denotes the diagonal, which could be used to reflect the deviation of data points.

based on the remaining one. After round 10, all of the observed values versus estimate values were considered to be the final result to validate the spatial transferability of this model. As depicted in Fig. S6 in the Supplement, the site-based cross-validation $R^2$ value reached 0.73, which was slightly lower than the cross-validation $R^2$ value of the training model (0.78). The result suggested the newly developed dataset showed excellent performance in the past decade.

Owing to the severe air pollution issue frequently observed in recent years, especially nitrogen-bearing haze events, many studies have tried to predict the NO$_3^-$ concentrations in China. Most of these studies employed CTMs to simulate the ambient NO$_3^-$ concentrations over China. Huang et al. (2015) employed WRF-CMAQ to estimate the inorganic nitrogen deposition over the Pearl River Delta (PRD) and confirmed that the $R$ value only reached 0.54. Afterwards, Han et al. (2017) used RAMS-GMAQ to predict the dry deposition flux of reactive nitrogen and significantly underestimated the NO$_3^-$ concentration in the atmosphere. Very recently, Geng et al. (2019) used CMAQ to estimate the NO$_3^-$ concentrations over east China, and the predictive performance ($R = 0.53$) showed a similar result to Huang et al. (2015). Apart from these CTMs, statistical models have also been applied to estimate the ambient NO$_3^-$ concentration over China. Unfortunately, the predictive accuracy was not good based on traditional statistical models (e.g. linear regression) ($R = 0.47$) (Jia et al., 2016). In terms of model performance, the developed NO$_3^-$ product in our study was much better than those developed by pioneering studies. Furthermore, this product showed many extra advantages than those obtained by CTMs, especially for the estimates of air pollutants. For instance, CTMs generally required continuous emission
Table 2. The cross-validation result of NO$_3^-$ estimation over China in four seasons.

<table>
<thead>
<tr>
<th>Season</th>
<th>Sample size</th>
<th>$R^2$ value</th>
<th>Slope</th>
<th>RMSE ($\mu$g N m$^{-3}$)</th>
<th>MAE ($\mu$g N m$^{-3}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring</td>
<td>395</td>
<td>0.80</td>
<td>0.80</td>
<td>0.71</td>
<td>0.48</td>
</tr>
<tr>
<td>Summer</td>
<td>418</td>
<td>0.85</td>
<td>0.84</td>
<td>0.29</td>
<td>0.20</td>
</tr>
<tr>
<td>Autumn</td>
<td>437</td>
<td>0.80</td>
<td>0.78</td>
<td>1.10</td>
<td>0.70</td>
</tr>
<tr>
<td>Winter</td>
<td>386</td>
<td>0.75</td>
<td>0.73</td>
<td>1.85</td>
<td>1.23</td>
</tr>
</tbody>
</table>

Table 3. The cross-validation result of NO$_3^-$ estimation over China in different regions (northeast China includes Heilongjiang, Jilin, and Liaoning provinces; NCP includes Beijing, Tianjin, Hebei, Henan, Shandong, and Shanxi provinces; southeast China includes Jiangsu, Zhejiang, Fujian, Guangdong, Jiangxi, Anhui, Hunan, Shanghai, and Hubei provinces; southwest China includes Yunnan, Guangxi, Sichuan, Tibet, Chongqing, and Guizhou provinces; northwest China includes Inner Mongolia, Xinjiang, Gansu, Qinghai, Ningxia, and Shaanxi).

<table>
<thead>
<tr>
<th>Region</th>
<th>Sample size</th>
<th>$R^2$ value</th>
<th>Slope</th>
<th>RMSE ($\mu$g N m$^{-3}$)</th>
<th>MAE ($\mu$g N m$^{-3}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast China</td>
<td>175</td>
<td>0.44</td>
<td>0.43</td>
<td>1.30</td>
<td>0.81</td>
</tr>
<tr>
<td>NCP</td>
<td>492</td>
<td>0.70</td>
<td>0.64</td>
<td>1.74</td>
<td>1.06</td>
</tr>
<tr>
<td>Southeast China</td>
<td>395</td>
<td>0.59</td>
<td>0.57</td>
<td>1.50</td>
<td>0.84</td>
</tr>
<tr>
<td>Southwest China</td>
<td>384</td>
<td>0.60</td>
<td>0.59</td>
<td>2.08</td>
<td>1.41</td>
</tr>
<tr>
<td>Northwest China</td>
<td>190</td>
<td>0.58</td>
<td>0.52</td>
<td>2.06</td>
<td>1.38</td>
</tr>
</tbody>
</table>

4.3 Spatial pattern of newly developed NO$_3^-$ dataset

The monthly NO$_3^-$ concentration displayed a similar distribution characteristic to PM$_{2.5}$ and PM$_1$ (Wei et al., 2019). Overall, the NO$_3^-$ concentration in east China was much higher than that in west China. The higher NO$_3^-$ concentration was concentrated on NCP ($3.55 \pm 1.25 \mu$g N m$^{-3}$), followed by the Yangtze River Delta (YRD) ($2.56 \pm 1.12 \mu$g N m$^{-3}$), Pearl River Delta (PRD) ($1.68 \pm 0.81 \mu$g N m$^{-3}$), and Sichuan Basin ($1.53 \pm 0.63 \mu$g N m$^{-3}$), and the lowest one was observed in the Tibetan Plateau ($0.42 \pm 0.25 \mu$g N m$^{-3}$) (Fig. 5). Most provinces over NCP such as Beijing, Hebei, Henan, and Shandong suffered from severe NO$_3^-$ pollution due to dense human activities and strong industry foundation (Li et al., 2017) (Fig. S7 in the Supplement), which released a large amount of N-bearing gaseous pollutants to the atmosphere, especially in winter. In Beijing–Tianjin–Hebei (BTH) ($2.97 \pm 1.97 \mu$g N m$^{-3}$), Wang et al. (2016) verified that this fresh NO$_x$ emitted from power plants or cement industries could be transformed into nitrate in the particulate phase by the aid of low air temperature. In the YRD and PRD, the combustion of fossil fuels and traffic emissions was considered to be the major source of NO$_x$ emission, which favoured the formation of nitrate events through the gas–particle conversion processes (Fu et al., 2017; Kong et al., 2020; Ming et al., 2017). Apart from the contributions of smelting industries, the poor topographical or meteorological conditions were also responsible for the severe NO$_3^-$ pollution in the Sichuan Basin (Tian et al., 2017; Wang et al., 2017). The Tibetan Plateau generally showed clean air quality due to the unique landscape and scarce industrial activity (Yang et al., 2018). In addition, it was interesting to note that...
the Altai region and Taklimakan desert in the Xinjiang autonomous region also showed some NO$_3^-$ hotspots, though these regions were often believed to be remote. It was assumed that the many petrochemical industries (e.g. Karamay oil field) were located in the Altai region (Liu et al., 2018). In addition, Qi et al. (2018) verified that the resuspension of soil dust might trigger the accumulation of NO$_3^-$ concentration in the aerosol.

4.4 Long-term trend of ambient NO$_3^-$ across China

The temporal variation in NO$_3^-$ levels from 2005 to 2015 over China has been clarified in Figs. 6 and 7 and Table S3 in the Supplement. Overall, the ambient NO$_3^-$ concentration in China showed the significant increasing trend of 0.10 µg N m$^{-3}$ yr$^{-1}$ during 2005–2014, while it decreased sharply from 2014 to 2015 by $-0.40$ µg N m$^{-3}$ yr$^{-1}$. Overall, more than 90% of mainland China showed consistent temporal variation with a gradual increase from 2005 to 2013/14 and then rapid decrease from 2013/14 to 2015. However, the decreasing/increasing speed displayed significant spatial difference in some major regions of China. For instance, the ambient NO$_3^-$ level in BTH showed a remarkable increase during 2005–2013 by 0.20 µg N m$^{-3}$ yr$^{-1}$. Afterwards, the NO$_3^-$ level decreased rapidly from 2013 to 2015 at a rate of $-0.58$ µg N m$^{-3}$ yr$^{-1}$. The NO$_3^-$ concentrations in the YRD (0.11 µg N m$^{-3}$ yr$^{-1}$) and PRD (0.05 µg N m$^{-3}$ yr$^{-1}$) both showed slight increases during 2005–2013, though the statistical test revealed the increases were significant ($p<0.05$). However, the NO$_3^-$ concentrations in the YRD and PRD showed dramatic decreases with $-0.48$ and $-0.36$ µg N m$^{-3}$ yr$^{-1}$ during 2013–2015, respectively. As seen from 2005 to 2015, the NO$_3^-$ concentration in BTH displayed on slight increase during this period. Nevertheless, the NO$_3^-$ levels in the YRD and PRD both displayed slow decreases of $-0.01$ and $-0.03$ µg N m$^{-3}$ yr$^{-1}$, respectively.

Furthermore, the different provinces displayed disparate temporal variations, especially during 11th 5-year plan (2005–2010). A total of 31 provinces (municipalities/autonomous region) of China can be classified into three clusters based on the temporal trends of NO$_3^-$ concentrations during the 11th 5-year plan. The first cluster featured a gradual increase in NO$_3^-$ concentration during this period, which consisted of three provinces in northeast China (e.g. Heilongjiang) and central provinces in south China (e.g. Jiangxi, Anhui) (Table S3). The second cluster represented the provinces with stable increases of NO$_3^-$ during 2005–2007 and slight decreases during 2007–2010. Some provinces of the NCP (e.g. Beijing, Hebei, Henan) and northwest China (e.g. Gansu, Inner Mongolia, Ningxia) fell into the second cluster. The last cluster featured the opposite temporal trend to the second cluster during 2005–2010, which included many southern provinces such as Fujian, Guangdong, Zhejiang, and Guangxi. Although the central government proposed the emission reduction goal in 2006, the ambient NO$_3^-$ concentrations in most provinces did not display pronounced decreases, which was totally different from the decrease in PM$_{2.5}$ since 2007 (Xue et al., 2019). Especially in the provinces of northeast China (e.g. Liaoning), the ambient NO$_3^-$ concentrations in these provinces still showed rapid increases after the proposal of emission control measures. It was assumed that these provinces generally possessed a large amount of energy-intensive industries and coal-fired power plants (Zhang et al., 2018). Moreover, the result might be associated with the fact that the emission reduction measures focused on the reduction of SO$_2$ emission rather than NO$_x$ emission (Kanada et al., 2013). Schreifels et al. (2012) revealed that major control measures during this period included shutting down inefficient industries, increasing the pollution levy for excessive SO$_2$ emissions, and implementing energy conservation projects. Therefore, the total SO$_2$ emission in 2010 decreased by more than 14% compared with the emission in 1995. The ambient SO$_2$ concentrations in many provinces since 2005 have displayed significant decreases compared with those in the 1990s (Li et al., 2020b; Lu et al., 2013; Zhou et al., 2015). Nonetheless, the NO$_x$ emission in China did not display a significant decrease during this period (Duncan et al., 2016; Granier et al., 2017), and thus the ambient NO$_3^-$ in many provinces still maintained higher concentrations. It should be noted that the NO$_3^-$ concentrations in some provinces of the NCP exhibited slow decreases after 2007. It was supposed that the energy structure adjustment and elimination of backward production capacity promoted the small decrease in NO$_3^-$ concentrations (Ma et al., 2019). Unfortunately, the slight decreases were quickly offset by the rapid increase in energy consumption. Zhang et al. (2018) demonstrated that the industry-added value and private cars in BTH increased by 189.4% and 279.6% during 2005–2010, respectively. In addition, the decrease in SO$_2$ emission rather than NO$_x$ emission can further lead to a NO$_3^-$ increase because of decreased aerosol acidity, which was dictated by SO$_2^-$ in particulate matter (Xie et al., 2020; Vasilakos et al., 2018).
Figure 6. The annual mean predicted NO$_3^-$ concentrations (µg N m$^{-3}$) across the entire study area from (a–k) 2005–2015 based on the ensemble model.

Figure 7. The annual mean NO$_3^-$ concentrations in major regions across China during 2005–2015. The solid lines denote the mean NO$_3^-$ concentrations, and the shadow represents the range of NO$_3^-$ concentrations.
Since 2010, the central government has begun to implement severe limitations in PM$_{2.5}$, NO$_x$, and soot emissions, and thus the total NO$_x$ emission during the 11th 5-year plan (2011–2015) showed a slow decrease (10%) across China (Ma et al., 2019). However, the NO$_3^-$ concentrations across China did not show a rapid response to the emission control measures. For instance, the NO$_3^-$ concentrations in most provinces of China still showed rapid increases during 2010–2013 (2014) (Figs. 7 and 8). The result suggested that the control measures about the NO$_x$ emissions from vehicles and ships might not be very effective. Until 2013, the central government had issued the Action Plan for Air Pollution Prevention and Control (APPC-AP) in order to enhance air pollution prevention measures (Li et al., 2017, 2019). Many powerful economic and policy means including pricing (tax) policy and optimisation of industrial layout caused rapid decreases in NO$_3^-$ concentrations after 2013 in many provinces (e.g. Beijing, Hebei, Zhejiang). Y. Wang et al. (2019a) also verified that the NO$_3^-$ level in PM$_{2.5}$ over BTH decreased by 20% during 2013–2015, which was in accordance with the findings of our study. In addition to the impact of emission reduction, the rapid decrease in NO$_3^-$ concentration over China after 2013 might be linked with beneficial meteorological factors because Z. Chen et al. (2019) demonstrated that favourable meteorological conditions led to about 20% of the PM$_{2.5}$ decrease in BTH during 2013–2015. However, the decreasing trend of NO$_3^-$ concentration during 2014/15 in the PRD (−0.36 µg N m$^{-3}$ yr$^{-1}$) was significantly slower than that in BTH (−0.58 µg N m$^{-3}$ yr$^{-1}$) and the YRD (−0.48 µg N m$^{-3}$ yr$^{-1}$) (Table 4). Y. Wang et al. (2019a) found that the ambient NO$_3^-$ concentration in a background site of the PRD even showed an upward trend during 2014–2016. Thus, it was necessary to strengthen the control of nitrogen oxide emissions.

In general, the ambient NO$_3^-$ concentration varied greatly at the seasonal scale (Fig. 9). China underwent the most serious NO$_3^-$ pollution in winter (1.57 ± 0.63 µg N m$^{-3}$), followed by autumn (1.09 ± 0.52 µg N m$^{-3}$) and spring (0.78 ± 0.50 µg N m$^{-3}$), and the lowest one in summer (0.63 ± 0.40 µg N m$^{-3}$) (Table S4 in the Supplement). The higher NO$_3^-$ concentration observed in winter might be contributed by the dense coal combustion in north China and unfavourable meteorological conditions (Ishii et al., 2017; Quan et al., 2014; Y. L. Wang et al., 2019). The lightest NO$_3^-$ pollution in summer was attributable to the abundant precipitation, which promoted the diffusion and removal of pollutants and reduced ambient NO$_3^-$ level (Hu et al., 2005). The ratio of NO$_3^-$ concentration in winter (NO$_3^-$ winter) and that in summer (NO$_3^-$ summer) varied greatly at the spatial scale. The NO$_3^-$ winter / NO$_3^-$ summer in some provinces (municipalities) including Tianjin (2.11), Hebei (2.25), and Henan (2.84) displayed higher values compared with other provinces. The higher NO$_3^-$ winter / NO$_3^-$ summer in the NCP might be affected by the fossil fuel combustion for domestic heating, while some southern provinces did not need domestic heating in winter. In contrast, the ratio of NO$_3^-$ winter / NO$_3^-$ summer exhibited lower values in some western provinces such as Tibet and Qinghai. This might be associated with less aerosol emission from anthropogenic sources and higher wind speed (Wei et al., 2019).

4.5 Uncertainty analysis of NO$_3^-$ estimation

The ensemble model of three machine-learning algorithms captured better accuracy in predicting the NO$_3^-$ level from OMI data. Nonetheless, the ensemble model still showed some improvement space in terms of the $R^2$ value. At first, meteorological data collected from reanalysis on the ECMWF website generally showed high uncertainty, which inevitably increased the error of NO$_3^-$ estimation. In our study, we validated the gridded $T_{2m}$ and $T_p$ datasets against the ground-observed datasets and found that the $R^2$ values of $T_{2m}$ and $T_p$ reached 0.98 and 0.83 (Table S5 in the Supplement), respectively. The result suggested that $T_{2m}$ showed a lower uncertainty, while $T_p$ displayed relatively higher uncertainty. Except $T_{2m}$ and $T_p$, the ground-level datasets for other meteorological factors were not open access, and thus we cannot assess their uncertainties. Thus, we only reviewed some references and evaluated their uncertainties. For instance, Guo et al. (2019) found that the reanalysis BLH data also exhibited large uncertainties because few sounding data were assimilated. These uncertainties derived from predictors could be passed to the ensemble model and thus increase the uncertainties of ambient NO$_3^-$ estimates.

The second reason was closely linked to the missing NO$_2$ column amount across China. The NO$_2$ column amount retrieval showed many nonrandom biases, especially for the arid or semi-arid area with high surface reflectance. The missing NO$_2$ column amounts over China were not filled in our study due to the increased uncertainty of filling the NO$_2$ column. Moreover, it should be noted that the monthly NO$_2$ column amounts were averaged based on the daily one, and the missing ratio of daily NO$_2$ columns during 2005–2015 reached 57.64%. The higher missing ratio might increase the uncertainty of NO$_3^-$ simulation.

Lastly, the developed ensemble model did not integrate the direct spatiotemporal weight indicators (e.g. the distance of observed sites and contiguous grids) though many predictors (e.g. month of year) reflecting spatiotemporal autocorrelation were input into the original model as the key predictors. Furthermore, the developed model was the ensemble one of three original models, which ignored the spatiotemporal autocorrelation of estimation residue from first-stage model. In future work, the ensemble model could be combined with a spacetime model to further enhance the modelling performance.
The newly developed NO$_3^-$ dataset showed remarkable predictive accuracy compared with previous products developed by CTMs and linear regression models. The result might be linked to two key reasons. First of all, the new product assimilated high-resolution NO$_2$ column amount instead of the NO$_x$ emission inventory used by CTMs. The imperfect knowledge about the chemical modules with regard to Earth Syst. Sci. Data, 13, 2147–2163, 2021 https://doi.org/10.5194/essd-13-2147-2021

Figure 8. The long-term trends of NO$_3^-$ concentrations (µg N m$^{-3}$) and significance levels in China (a, b, and c denote the annual variation in ambient NO$_3^-$ concentration during 2005–2015, 2005–2014, and 2014/15, respectively. Panels (d), (e), and (f) represent the significance level of the NO$_3^-$ trend during these periods). The pale green colour denotes the regions with significant variation in ambient NO$_3^-$ concentrations ($p < 0.05$), while the grey colour represents the regions with insignificant variation in NO$_3^-$ concentrations.

Table 4. The trend analysis of NO$_3^-$ concentrations in China, BTH, YRD, and PRD regions during 2005–2015.

<table>
<thead>
<tr>
<th>Period</th>
<th>Trend (µg N m$^{-3}$ yr$^{-1}$)</th>
<th>China</th>
<th>BTH</th>
<th>YRD</th>
<th>PRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005–2014</td>
<td>0.08</td>
<td>0.13</td>
<td>0.08</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Significance</td>
<td>$p &lt; 0.05$</td>
<td>$p &lt; 0.05$</td>
<td>$p &lt; 0.05$</td>
<td>$p &lt; 0.05$</td>
</tr>
<tr>
<td>2014/15</td>
<td>$-0.40$</td>
<td>$-0.76$</td>
<td>$-0.79$</td>
<td>$-0.59$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Significance</td>
<td>$p &lt; 0.05$</td>
<td>$p &lt; 0.05$</td>
<td>$p &lt; 0.05$</td>
<td>$p &lt; 0.05$</td>
</tr>
<tr>
<td>2005–2015</td>
<td>$0.04$</td>
<td>$0.04$</td>
<td>$-0.01$</td>
<td>$-0.03$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Significance</td>
<td>$p &lt; 0.05$</td>
<td>$p &gt; 0.05$</td>
<td>$p &gt; 0.05$</td>
<td>$p &lt; 0.05$</td>
</tr>
</tbody>
</table>

5 Code availability
The code used for this article is not publicly available, but can be obtained on request from the corresponding authors.

6 Data availability
The monthly NO$_3^-$ datasets at 0.25° resolution across China during 2005–2015 are available at https://doi.org/10.5281/zenodo.3988307 (Li et al., 2020c), which can be downloaded in xlsx format. The missing values are shown as NaN.

7 Conclusions and implications
In this study, RF, GBDT, and XGBoost algorithms were combined to establish a high-resolution (0.25°) NO$_3^-$ dataset over China during 2005–2015 on the basis of multi-source predictors. The NO$_3^-$ product showed a high cross-validation $R^2$ value (0.78) but low RMSE (1.19 µg N m$^{-3}$) and MAE (0.81 µg N m$^{-3}$). The NO$_3^-$ dataset showed the marked spatial-temporal discrepancy. The $R^2$ value was in the order of summer (0.85) > spring (0.80) = autumn (0.80) > winter (0.75) across China, and the $R^2$ showed the highest value in the NCP. In addition, the dataset exhibited excellent transferability ($R^2 = 0.85$, RMSE = 0.74 µg N m$^{-3}$, and MAE = 0.55 µg N m$^{-3}$) on the basis of the non-learning-observed data in 10 sites.

the NO$_3^-$ formation and the inaccurate emission inventory decreased the predictive performance of CTMs. In contrast, the new product was obtained using an ensemble machine-learning model, which did not need to consider the photochemical or aqueous process from gaseous NO$_2$ to particulate NO$_3^-$. Compared with the NO$_3^-$ product estimated by linear regression model ($R^2 = 0.21$), the new product significantly elevated the modelling performance of NO$_3^-$ concentration. It was supposed that the ensemble model for the development of the new NO$_3^-$ dataset did not predefine the potential relationships between explanatory variables and NO$_3^-$ level like the multiple regression model, which must assume the linear linkage between dependent variables and predictors before model establishment.

On the basis of the dataset, the spatiotemporal variation in NO$_3^-$ concentration over China during 2005–2015 was clarified. The annual mean NO$_3^-$ concentration followed the order of NCP (3.55 ± 1.25 µg N m$^{-3}$) > YRD (2.56 ± 1.12 µg N m$^{-3}$) > PRD (1.68 ± 0.81 µg N m$^{-3}$) > Sichuan Basin (1.53 ± 0.63 µg N m$^{-3}$) > Tibetan Plateau (0.42 ± 0.25 µg N m$^{-3}$). The higher NO$_3^-$ concentrations in the NCP, YRD, and PRD were mainly contributed by the intensive industrial and traffic emissions. The Sichuan Basin suffered serious NO$_3^-$ pollution due to the high loadings of aerosols and unfavourable terrain conditions. The Tibetan Plateau showed the lightest NO$_3^-$ pollution because of the scarce anthropogenic emissions and favourable meteorological factors. Additionally, we also found that the ambient NO$_3^-$ concentration showed significant increasing trend of 0.10 µg N m$^{-3}$ yr$^{-1}$ during 2005–2014, while it decreased sharply from 2014 to 2015 at a rate of −0.40 µg N m$^{-3}$ yr$^{-1}$. The ambient NO$_3^-$ levels in BTH, YRD, and PRD displayed slight increases at the rate of 0.20, 0.11, and 0.05 µg N m$^{-3}$ yr$^{-1}$ during 2013–2015, respectively. Afterwards, the NO$_3^-$ concentrations decreased sharply at the speed of −0.58, −0.48, and −0.36 µg N m$^{-3}$ yr$^{-1}$. Although National Economic and Social Development of China issued an emission reduction goal in 2006, the NO$_3^-$ concentrations in most provinces did not show significant decreases during 2005–2010. This might be contributed by the increase in energy consumption and non-targeted emission control measures. Since 2010, the government began to decrease the NO$_x$ emission over China, whereas the NO$_3^-$ concentrations in many provinces still showed slight increases during 2010–2014 because the benefits of control measures for NO$_x$ emission could be neutralised by elevated energy consumption along with rapid economic development. Since 2014, the Chinese government has issued APPC-AP and further enhanced the emission control measures, which triggered the dramatic decrease in NO$_3^-$ concentration over China. Apart from the effect of emission reduction, the favourable meteorological conditions might have led to the rapid decrease in NO$_3^-$ levels over China during 2014/15. Compared with the powerful emission control measures, meteorological factors only contributed a small portion of NO$_3^-$ reduction in China. In addition, the decrease speed of NO$_3^-$ level in China also displayed pronounced spatial heterogeneity, and some background regions have even featured an increase in air pollution in

Figure 9. The mean concentrations of ambient NO$_3^-$ in spring (a), summer (b), autumn (c), and winter (d) during 2005–2015 over the study area.
It must be acknowledged that our study still suffers from some limitations. First of all, the NO$_3^-$ dataset was developed by machine-learning models, which lacked the chemical module concerning the transformation pathway from NO$_2$ to NO$_3^-$, and might underestimate the ambient NO$_3^-$ concentration across China. In future work, the output results of CTMs including the conversion ratio from NO$_2$ to NO$_3^-$ and dry–wet deposition flux of NO$_2$ and NO$_3^-$ in the atmosphere should be incorporated into the machine-learning model to develop a next-generation NO$_3^-$ product. Second, the low-time-resolution (monthly) observation data hindered the daily estimation of NO$_3^-$ concentration. The daily NO$_3^-$ datasets are warranted in the future because they can be used to assess the potential impact on human health. In addition, the ultrahigh-resolution satellite (TROPOMI) can allow continuation and enhancement of the spatiotemporal NO$_3^-$ estimation, and the OMI product could capture enough spatial variations across China.

**Supplement.** The supplement related to this article is available online at: https://doi.org/10.5194/essd-13-2147-2021-supplement.

**Author contributions.** RL, LC, and HF conceived and designed the study. RL, LC, YZ, and WZ collected and processed the data. RL wrote this paper with contributions from all of the coauthors.

**Competing interests.** The authors declare that they have no conflict of interest.

**Acknowledgements.** This work was funded by Chinese Postdoctoral Science Foundation (2020M680589) and National Natural Science Foundation of China (no. 21777025).

**Financial support.** This research has been supported by the National Natural Science Foundation of China (grant no. 91744205).

**Review statement.** This paper was edited by David Carlson and reviewed by two anonymous referees.

**References**


