

<https://doi.org/10.1038/s41612-024-00617-1>

Mapping the seamless hourly surface visibility in China: a real-time retrieval framework using a machine-learning-based stacked ensemble model

Check for updates

Xutao Zhang¹, Ke Gui¹✉, Zhaoliang Zeng², Ye Fei³, Lei Li¹, Yu Zheng¹, Yue Peng¹, Yurun Liu⁴, Nanxuan Shang¹, Hengheng Zhao¹, Wenrui Yao¹, Hong Wang¹, Zhili Wang¹, Yaqiang Wang², Huizheng Che¹✉ & Xiaoye Zhang¹

Surface visibility (SV), a key indicator of atmospheric transparency, is used widely in the fields of environmental monitoring, transportation, and aviation. However, the sparse distribution and limited number of SV monitoring sites make it difficult to fulfill the urgent need for spatiotemporally seamless fine-scale monitoring. Here, we developed the operational real-time SV retrieval (RT-SVR) framework for China that incorporates information from multiple data sources, including Chinese Land Data Assimilation System meteorological data, in situ observations, and other ancillary data. Seamless hourly SV data with 6.25-km spatial resolution are available in real time via the operational RT-SVR model, which was built using a two-layer stacked ensemble approach that combines multiple machine learning algorithms and a deep learning module. Sample-based cross-validation of the RT-SVR model on approximately 41.3 million data pairs revealed strong robustness and high accuracy, with a Pearson correlation coefficient (R) value of 0.95 and a root mean square error (RMSE) of 3.17 km. An additional hindcast-validation experiment, performed with continuous observations obtained over one year (approximately 20.8 million data pairs), demonstrated the powerful generalization capabilities of the RT-SVR model, albeit with slight degradation in performance ($R = 0.85$, RMSE = 5.28 km). The seamless hourly SV data with real-time update capability enable tracking of the generation, development, and dissipation of various low-SV events (e.g., fog, haze, and dust storms) in China. The developed framework might also prove useful for quantitative retrieval of aerosol-related parameters (e.g., PM_{2.5}, PM₁₀, and aerosol optical depth).

Surface visibility (SV), often used as a proxy for ambient air quality, exhibits strong association with national transportation and human health^{1–4}. In recent decades, severe and frequent low-SV episodes related to air quality during events such as haze and dust storms have received widespread attention^{5–7}. The main factors affecting SV are relative humidity (RH) and ambient particulate matter (PM), which can affect SV by directly or indirectly changing the transmission path of solar radiation^{8,9}. However, the

relationships between PM concentration and SV are complex and nonlinear owing to variations in the size distribution, mixing state, and chemical composition of aerosol particles^{10,11}. Moreover, this relationship is also affected by various meteorological factors (especially RH) that exhibit notable regional and seasonal differences¹².

Since 2015, the China Meteorological Administration (CMA) has conducted automated observations of SV, gradually replacing the previous

¹State Key Laboratory of Severe Weather & Key Laboratory of Atmospheric Chemistry of CMA, Chinese Academy of Meteorological Sciences, Beijing 100081, China. ²Institute of Artificial Intelligence for Meteorological, Chinese Academy of Meteorological Sciences, Beijing 100081, China. ³National Meteorological Information Center, Beijing 100081, China. ⁴Plateau Atmospheric and Environment Key Laboratory of Sichuan Province, College of Atmosphere Sciences, Chengdu University of Information Technology, Chengdu 610225, China. ✉e-mail: guike@cma.gov.cn; chezh@cma.gov.cn

manual method of conducting observations only 3–6 times daily¹³. The automated instruments for monitoring SV measure surface scattering coefficients and can provide high-quality SV observations on an hourly basis. Currently, the CMA has equipped more than 2400 sites with such instruments, thereby providing a good opportunity to track the diurnal cycle of SV variability. However, despite the considerable efforts made to date, the relatively sparse distribution and limited number of SV monitoring sites make it difficult to meet the current urgent need for spatiotemporally seamless fine-scale monitoring of SV. Thus, seamless hourly high-resolution SV retrievals would be useful for tracking various low-SV events caused by different weather conditions or background emissions. Moreover, such data would also strongly support quantitative retrieval of aerosol-related parameters, including fine PM (PM_{2.5})^{14–16}, coarse PM (PM₁₀)¹⁷, and aerosol optical depth (AOD)^{18,19}, for which SV is recognized as a key intermediate variable.

To address the dearth of SV observational data, studies in recent decades have investigated obtaining SV data indirectly. Initially, such studies used statistical models or radiation theory for indirect estimation of SV using synchronized observations of meteorological, aerosol optical, and chemical component properties as key predictors^{20,21}. Such a method is somewhat accurate but has strong geographical dependence, which leads to discrepancies in its spatial applicability. Additionally, it also relies on online observations of chemical components, which can hinder its application to large-scale estimation. In recent years, to resolve the difficulty in obtaining SV data on a large scale, some studies indirectly retrieved the spatial distribution of SV based on spectral signals from polar-orbiting or geostationary satellites and their derivatives (e.g., AOD)^{22–25}. However, satellite-derived SV products can have incomplete spatial coverage due to cloud interference, thereby limiting their application prospects^{26–28}. Furthermore, owing to the unavailability of nighttime aerosol products, satellite-based retrieval strategies are not yet able to achieve seamless SV monitoring on a 24-h cycle. Consequently, a real-time seamless hourly SV retrieval framework for China that can be operationalized and shared with the wider community remains lacking.

In this study, a spatiotemporally seamless SV retrieval framework for China with real-time operational capability was constructed based on multisource data fusion and machine learning (ML) methods. Here, we show the development of a real-time SV retrieval (RT-SVR) model in this framework, which is an end-to-end stacked ensemble model composed of bagging, boosting, and deep network algorithms. With the support of real-time availability of multisource input data, the RT-SVR model has the capability to provide real-time output of seamless hourly SV data for China with 6.25-km spatial resolution. Real-time retrieval of high-accuracy seamless hourly SV data is beneficial both for further improving the monitoring capability of national-scale low-SV events, and for advancing the application of spatiotemporally continuous multiple aerosol-related parameter retrievals.

Results

Advantages of the stacked ensemble model

In this study, we chose not to use an independent ML algorithm to establish the complex relationships between multiple predictors and observed SV. Instead, we developed a two-layer stacked ensemble model that combines five widely used ML algorithms (i.e., MLP, RF, CatBoost, XGBoost, and LightGBM; see Methods) with robust performance and a deep residual network (Fig. 1). In the following, we explain in detail the advantages of the stacked ensemble model developed in this study compared to an independent ML algorithm through consideration of a case that occurred at 02:00 CST (China Standard Time) on December 10, 2022 (Supplementary Fig. 1).

Supplementary Fig. 2a–f show the spatial distribution of SV at 02:00 CST on December 10, 2022, as estimated by the MLP, RF, CatBoost, XGBoost, LightGBM, and RT-SVR models, respectively. For the MLP model, although the estimated SV exhibits a continuous spatial pattern similar to that of the observed SV (see Supplementary Fig. 1), the model tends to be spatially over-smoothed in areas of high SV, e.g., northwestern China (Supplementary Fig. 2a). This is an inherent drawback of the MLP

approach, where the bias between estimated SV and observed SV tends to be large on the tabular data²⁹. Compared with the MLP model, the RF model provides better portrayal of the spatial details of SV in the Tibetan Plateau region, but overfitting occurs in the northwestern region (Supplementary Fig. 2b). Overall, in areas of high SV, the three boosting models (i.e., CatBoost, XGBoost, and LightGBM) successfully address the problems encountered by the MLP and RF models. However, for regions with extremely low SV, e.g., the North China Plain (NCP), the SV estimated by all three boosting models exhibits unrealistic anomalies with negative retrieval values (Supplementary Fig. 2c–e). The statistics indicate that the grid points with negative retrieval values encountered by CatBoost, XGBoost, and LightGBM over the NCP region comprise approximately 0.84%, 0.58%, and 0.63% of the total sample size, respectively. It should be noted that we revised the loss function of XGBoost to eliminate negative retrieval anomalies in the model, which explains the lower number of negative retrieval values for the XGBoost (Supplementary Fig. 2d)³⁰.

To address the various shortcomings encountered by the different ML models, we constructed a deep residual network in the second layer using the outputs of five independent ML models as inputs to comprehensively integrate the strengths and weaknesses of the base models in the first layer (Fig. 1). Supplementary Fig. 2f depicts the spatial map of SV retrieved by the RT-SVR model coupled with a two-layer stacked structure. It is evident that the deep residual network avoids the overfitting properties of the stacking methods and is spatiotemporally capable of combining the advantages of the first layer of the base model.

Relative importance of explanatory variables

We used the permutation method to interpret the relative importance of each predictor in the five base models in the first layer of the RT-SVR model. The same approach was also applied to account for the relative contribution of the retrieval from each base model in the first layer to the final retrieval from the second layer (a deep residual network). The permutation method calculates the relative importance of each feature by shuffling it randomly during training³¹. The relative importance of the predictors for retrieving SV in the RT-SVR model is illustrated in Supplementary Fig. 3. In the first layer of the RT-SVR model, each base model has different usage of the predictors. As expected, surface RH, which can directly or indirectly affect fog or cloud formation and modify light scattering and thereby lead to changes in SV, accounts for approximately 40% of the overall importance of predictors among the base models. For tree-based models, RH and PM_{2.5} are the dominant factors affecting SV retrievals because aerosol hygroscopic growth varies under different RH conditions¹². It is notable that in the MLP model, surface temperature (TEM) is the most dominant predictor. Statistically significant correlation between TEM and SV has also been identified in some previous studies. On the one hand, TEM can affect the rate of aerosol hygroscopic growth and indirectly affect SV^{32,33}; on the other hand, TEM affects the stability of the atmospheric boundary layer, which can influence pollutant diffusion²⁰. In addition to the predictors mentioned above, geographical factors (e.g., the normalized difference vegetation index (NDVI) and elevation) and temporal factors (i.e., day of year), which have notable spatial heterogeneity, are also critical in estimations of SV.

Evaluation of the RT-SVR model performance

The performance of the RT-SVR model was evaluated using a sample-based cross-validation (CV) approach and a hindcast-validation (HV) experiment. The k-fold CV is used most often to test model robustness. In this study, a 5-fold CV approach was chosen, i.e., all site-based samples were randomly divided into five subsets. Each time, the model was trained using data from four subsets and tested on the remaining subset. In contrast, the HV experiment evaluated the generalization ability of the model (i.e., the real prediction ability) on a completely independent dataset that was not involved in any training process of the model.

Figure 2 shows the overall fitting and the 5-fold CV results of the RT-SVR model for all 24-h periods during 2020–2021. Here, the performance of

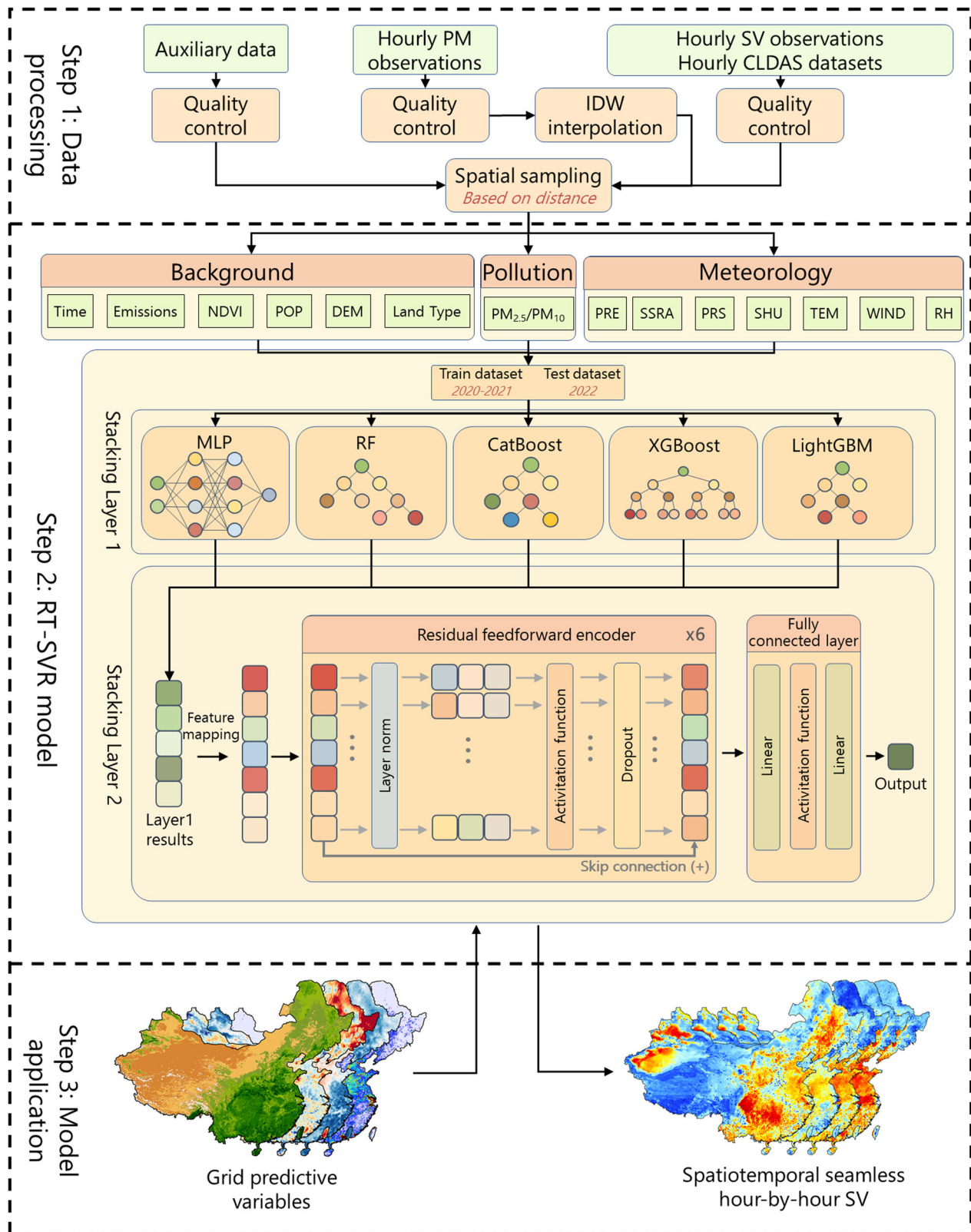


Fig. 1 | Operational procedure of the RT-SVR framework for generating seamless hour-by-hour SV data in China. The first and second steps are the data pre-processing and detailed model building process. The third step is the operational application of the model.

the RT-SVR model for SV retrieval is indicated using the Pearson correlation coefficient (R), root mean square error (RMSE), mean absolute error (MAE), and relative mean bias (RMB). Overall, the 5-fold CV results of the model are comparable to the fitting results, indicating that our model is

robust and does not experience overfitting problems. The RT-SVR SVs estimated from the 5-fold CV, tested on approximately 41.3 million data pairs, exhibit reasonable agreement with the in situ observations ($R = 0.95$, MAE = 2.16 km, and RMSE = 3.17 km).

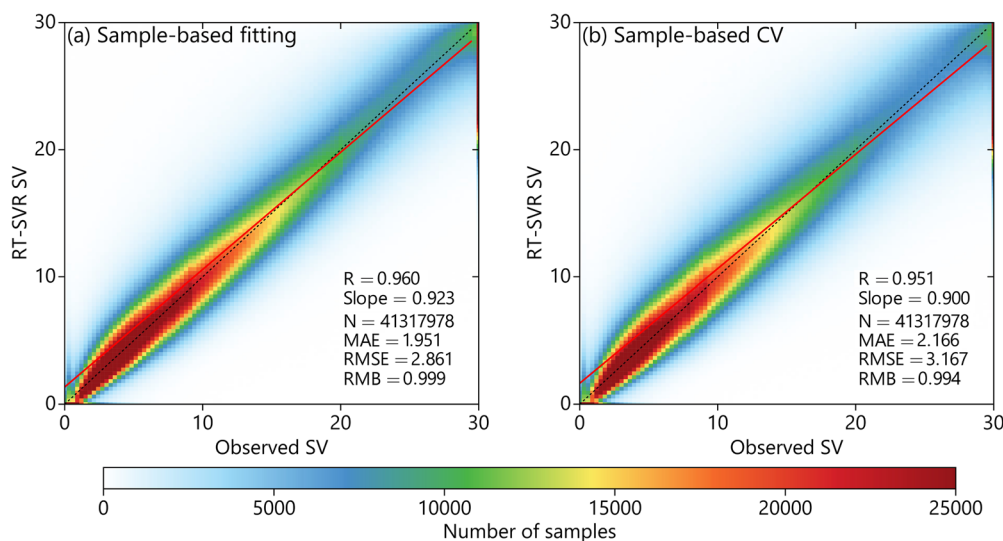


Fig. 2 | The performance of the RT-SVR model. Density scatterplots between observed SV and (a) fitted SV and (b) estimated SV across China for sample-based fivefold CV from 2020 to 2021. CV results are generated with hourly temporal resolution. The dashed black line is the 1-to-1 line and the solid red line is the linear

regression line. The number of sample matchups (N), Pearson correlation coefficient (R), slope, mean absolute error (MAE), root mean square error (RMSE), and relative mean bias (RMB) of the linear regression are shown in the lower-right corner of each panel.

To further check the performance of the model in real application scenarios, we performed an additional HV experiment to examine the model's ability to generalize when using unused data. For the HV experiment using one-year's (2022) continuous observations (approximately 20.8 million data pairs), the RT-SVR model still exhibits powerful generalization capabilities, albeit with slight degradation in performance ($R = 0.85$, MAE = 3.77 km, and RMSE = 5.28 km) (see Fig. 3i and Supplementary Fig. 4f). Similar HV experiments were also undertaken with five independent ML models in the first layer and the results are shown in Supplementary Fig. 4a–e. Overall, in addition to the improvement in performance spatially, the RT-SVR model also exhibits improved accuracy compared to all five ML models, with the lowest RMSE and the highest value of R (Supplementary Fig. 4f). Importantly, the RT-SVR model resolves the problems of anomalous negative retrievals encountered by the boosting models. Moreover, the RT-SVR model demonstrates outstanding predictive performance compared to that of other models that use satellite-retrieved AOD as the predictor²².

In addition to examining the overall performance of the model, the site-scale performance was also evaluated. Supplementary Figs. 5, 6 show the spatial distribution of the site-based statistical indicators associated with the 5-fold CV and HV experiment, respectively. For the 5-fold CV (Supplementary Fig. 5), the averaged R , RMSE, and MAE across China is 0.93, 3.10 km, and 2.16 km, respectively. Among them, the number of sites with $R > 0.94$ accounted for 45.4% (1091 sites) of the total number of sites, the majority of which are located on the NCP. Meanwhile, those sites with relatively low R -values are mainly located on the Tibetan Plateau and in northeastern China. This is attributed to the fact that the SV sites in those regions are sparsely distributed, making it difficult to acquire sufficient information for model development (Supplementary Fig. 1). For the HV experiment, the overall distribution pattern of the above metrics is similar to that of the 5-fold CV (Supplementary Fig. 6). Statistically, the averaged R , RMSE, and MAE for the HV experiment across China is 0.80, 5.08 km, and 3.77 km, respectively, and approximately 43.9% (1055 sites) of the sites have an R -value of >0.85 . Spatially, our model has greater bias in south-central and north-central China compared to that in the region of the NCP, mainly because of the more complex relationships between SV, RH, and $PM_{2.5}$ in those areas^{34,35}. The drier climatic background of the NCP allows SV variability in this region to be modulated mainly by atmospheric $PM_{2.5}$ pollution. The low bias in the region of the NCP is largely attributable to the relatively accurate pollutant and meteorological fields associated with the large number of monitoring sites distributed throughout the region. We also note that a

similarly high bias in the model also exists in northeastern and northwestern China, which might be related to the inadequate spatial representation of background PM_{10} information interpolated from the fewer observational sites in those regions (see Methods).

Consistent accuracy on the full 24-h diurnal cycle

Supplementary Figs. 7, 8 show the performance of the RT-SVR model on an hourly scale based on the sample-based 5-fold CV and HV experiments, respectively. On the CV dataset, the RT-SVR model shows superior temporal robustness and its statistical parameters do not exhibit significant fluctuations over time; the model has values of $R > 0.93$ and RMSE < 3.5 km at all 24-h periods (Supplementary Fig. 7). On the HV dataset, although the accuracy of the RT-SVR model is reduced compared to that on the CV dataset at all 24-h periods, the performance of the model is robust across hours (Supplementary Fig. 8).

We further selected six SV observation sites with different underlying surface types, i.e., desert (Urumqi site), basin (Wenjiang site), plateau (Nagqu site), and urban (Beijing, Nanjing, and Shenzhen sites), to further examine the performance of the RT-SVR model. Figure 3 and Supplementary Fig. 9 present the daily and hourly time series of SV at those sites retrieved by the RT-SVR model during the HV experiment in 2022, respectively. In Beijing, the estimated daily SV agrees well with the observed SV, with an R -value of 0.97 and RMSE of 2.47 km. Similarly good performance is observed in both Nanjing ($R = 0.96$, RMSE = 2.22 km) and Wenjiang ($R = 0.94$, RMSE = 2.28 km). At the other city site, i.e., Shenzhen, the model performance is slightly reduced. In contrast, there is notable reduction in model performance at the western sites of Urumqi ($R = 0.85$, RMSE = 2.68 km) and Nagqu ($R = 0.83$, RMSE = 2.02 km) compared to that of the eastern sites in China. However, it should be noted that there are also differences in model performance across seasons at different sites (Supplementary Fig. 9). For example, at the Urumqi site, larger error occurs in spring (Supplementary Fig. 9a), during which time frequent dust storm events tend to cause short-term abrupt changes in SV, leading to an under-response of the model^{36,37}. At the Nanjing and Wenjiang sites, the RT-SVR model has worst performance in summer (Supplementary Fig. 9c, e), which might be related to the seasonal increase in precipitation events caused by frequent strong convective activities^{38,39}. The combination of higher temperature and increased precipitation in summer highlights the difficulty the model has in accurately retrieving SV under conditions of high RH¹². The model performs poorly at the Nagqu site in all seasons except winter (Supplementary Fig. 9d). The

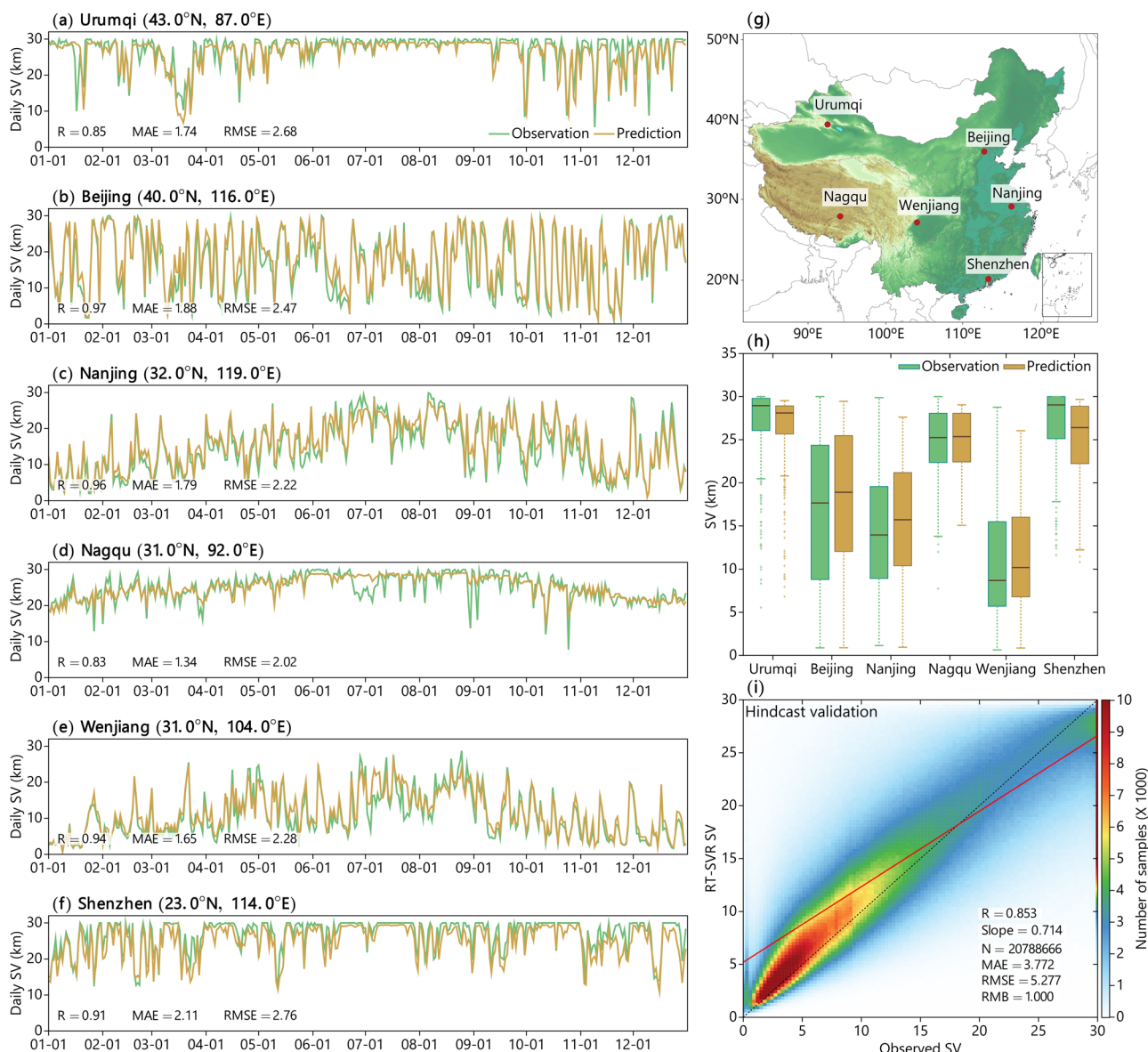


Fig. 3 | Performance of the RT-SVR model in the hindcast-validation (HV) experiment in 2022. Comparison of daily SV time series from observations and from the RT-SVR model at six different sites in China: (a) Urumqi, (b) Beijing, (c) Nanjing, (d) Nagqu, (e) Wenjiang, and (f) Shenzhen. **g** Geographic locations of the six independent sites selected for this study. **h** Boxplot of observed and estimated SV at the six sites (statistics performed on the daily time series for 2022). The boxplot illustrates the median, the interquartile range, the upper (lower) whisker extending

from the hinge to the largest (smallest) value no further than $1.5 \times$ interquartile range, and outliers as individual points. The box extends from lower to upper quartile values of the data, with a horizontal line at the median. The whiskers extend from the box to show the range of the data. **i** Density scatterplots between observed SV and estimated SV across China for the HV experiment in 2022 (results based on hourly temporal resolution).

uncertainty of the meteorological fields in the region of the Tibetan Plateau might contribute to this level of performance. Figure 4 displays the distribution of the retrieved hourly SV averaged over the period 2020–2022 for the RT-SVR model, trained using the 2020–2022 data. The site observations are shown in Supplementary Fig. 10. Overall, the RT-SVR model captures the features of the diurnal cycle of the seamless hourly SV in China. For example, in the early morning and evening, low surface temperatures (due to radiative cooling) and high RH are conducive to fog formation, leading to reduced SV⁴⁰. Additionally, during this period, such meteorological conditions coupled with low boundary layer heights favor accumulation of pollutants (especially surface PM), leading to enhanced extinction, which further reduces SV^{12,41}. Spatially, low SV values are mainly located in the regions of the NCP, central-eastern China, Sichuan Basin, and Taklimakan Desert. In contrast, as temperatures rise during the day, surface heating leads to

reduction in RH and enhancement of vertical motion, the latter of which contributes to greater dispersion of pollutants and thus improves SV. Supplementary Fig. 11 shows the multiyear averaged seasonal SV maps retrieved from the RT-SVR model during 2020–2022. During the summer, enhanced surface heating leads to vigorous convective activity in the lower atmosphere and significant precipitation, resulting in a notable improvement in SV. In contrast, during the winter, the prevalence of low-SV events escalates due to suboptimal atmospheric dispersion conditions that are compounded by the synergistic effects of adverse weather conditions and anthropogenic emissions, with this phenomenon being particularly pronounced in the central and eastern regions of China. Supplementary Figs. 12–15 show the multiyear averaged hour-by-hour SV maps retrieved from the RT-SVR model in the four seasons during 2020–2022, respectively. Consistent with the patterns observed in Fig. 4, the diurnal variations of SV in China exhibit a similar

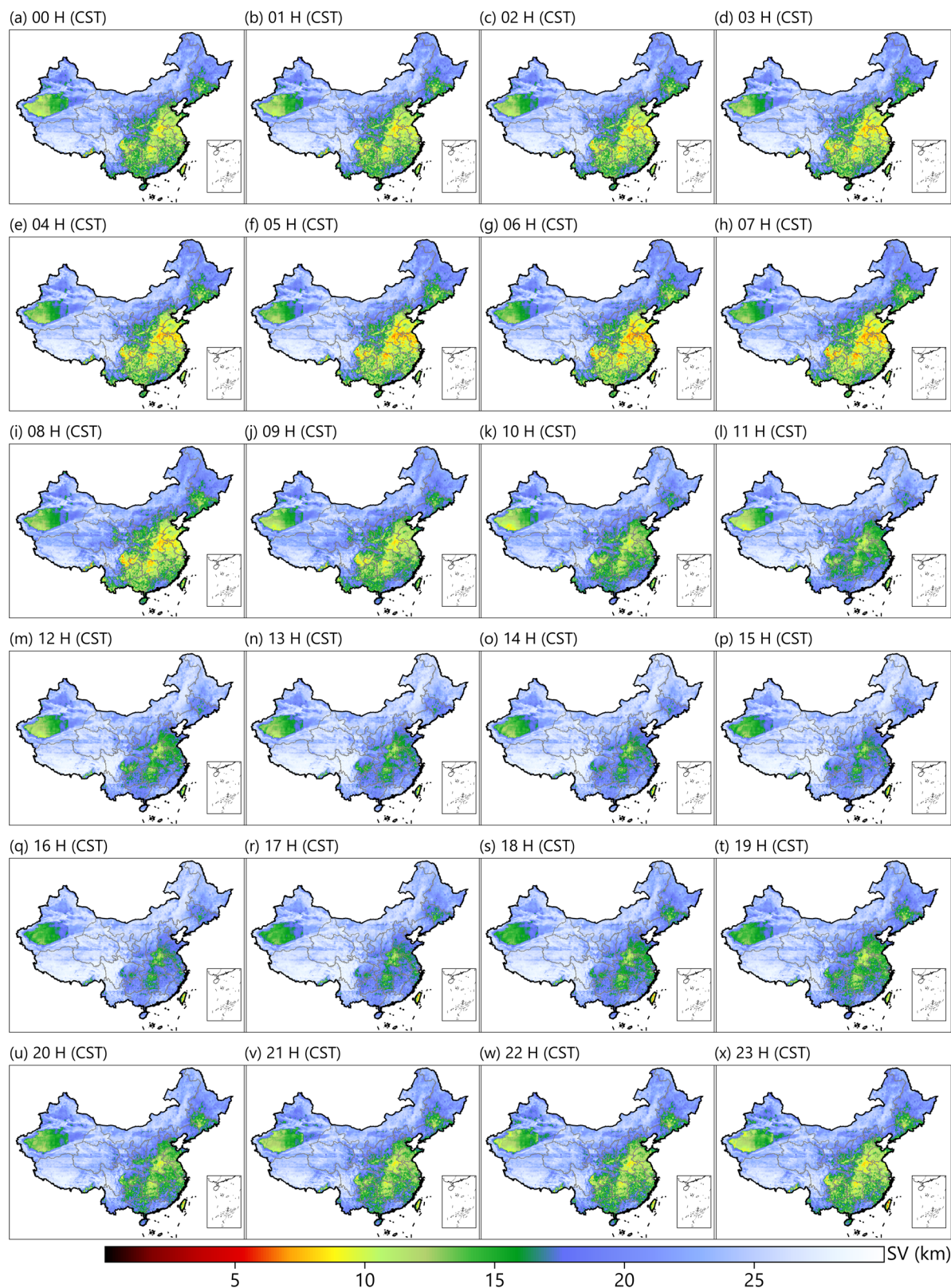


Fig. 4 | Hourly SV distributions in China. Multiyear averaged hour-by-hour SV maps retrieved from the RT-SVR model during 2020–2022.

pattern across different seasons, characterized by the lowest SV in the early morning or evening, and the highest SV in the afternoon. Overall, the diurnal and seasonal variations of SV in China are driven by a complex process that is influenced by a combination of factors such as topography, season, meteorological conditions, and ambient air pollution.

Model applications: tracking low-SV events

To fully demonstrate the model’s ability to track the entire process of occurrence, development, and dissipation of various extreme low-SV weather episodes, we considered three typical events that can have major impact on SV: fog, haze, and dust storms.

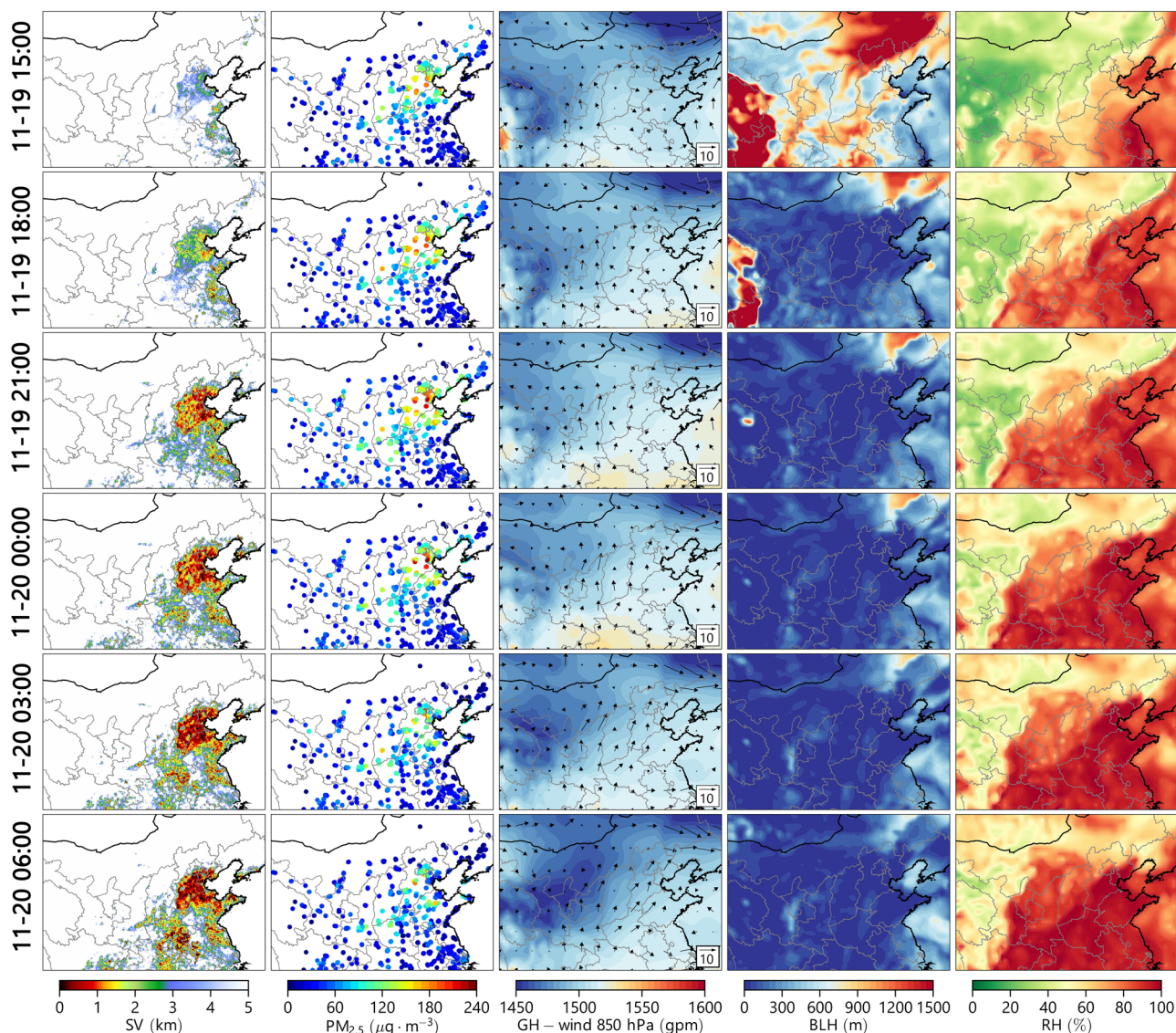


Fig. 5 | Model application to track typical low-SV events. Evolutionary maps of 3-h SV (first column), surface $PM_{2.5}$ observations (second column), and meteorological fields (third column: geopotential height (GH) and wind field at 850 hPa; fourth

column: boundary layer height (BLH); fifth column: surface relative humidity (RH)) during a fog event that occurred on November 19–20, 2022. Meteorological fields are from the fifth-generation ECMWF atmospheric reanalysis (ERA5).

Fog event

Figure 5 shows the 3-h variability of SV retrieved from the RT-SVR model, together with the observed $PM_{2.5}$ and the geopotential height, wind vector, boundary layer height, and RH from the fifth-generation ECMWF atmospheric reanalysis (ERA-5) data, during the formation of a fog event that occurred on the NCP during November 19–20, 2022. The full hour-by-hour evolution of this process is depicted in Supplementary Fig. 16. This event began after 12:00 CST on November 19, 2022. As shown in Fig. 5, this event was primarily driven by meteorological conditions because the regional $PM_{2.5}$ concentrations were at low levels. Specifically, from the night of November 19 to the early morning of November 20, 2022, with the rapid increase in RH (>90%) and the reduction in the boundary layer height, a persistent foggy event affected the NCP and surrounding areas, which caused rapid reduction in the regional SV to the 100-m level. Our retrieval products well captured this process, demonstrating the ability of the RT-SVR model to reproduce extreme low-SV events.

Haze event

Apart from fog, increase in atmospheric aerosols can also cause rapid decline in SV. We further examined the performance of the RT-SVR model

in tracking extreme low-SV events caused by anthropogenic aerosols (described as a haze event) and dust aerosols (described as a dust storm event). Supplementary Fig. 17 shows the 6-hourly variability of SV during the formation of a haze event that occurred in eastern China during January 20–22, 2021. Despite the spatial extent of the impact of this haze event being comparable to that of the fog event (see Fig. 5), its duration was longer, persisting for nearly three days.

Overall, this haze event was mainly caused by gradual accumulation of local anthropogenically derived $PM_{2.5}$ under unfavorable meteorological conditions. From the early morning of January 21, 2021, gradual increase in $PM_{2.5}$ (up to $240 \mu\text{g m}^{-3}$) was observed in eastern China, especially in southern Hebei, Shandong, Shaanxi, and Henan. Correspondingly, our retrieval product also revealed gradual decline in regional SV, reaching a minimum (approximately 500 m) at approximately 10:00 CST on January 22, 2021 (Supplementary Figs. 17 and 18). This process was mainly regulated by changes in local meteorological conditions caused by the large-scale circulation. During this period, the reduction in the planetary boundary layer height and the diminished wind speed were not conducive to the dispersion of pollutants. Meanwhile, the increase in RH further promoted the hygroscopic growth of aerosols and contributed to the increase in $PM_{2.5}$ ^{42,43}. The

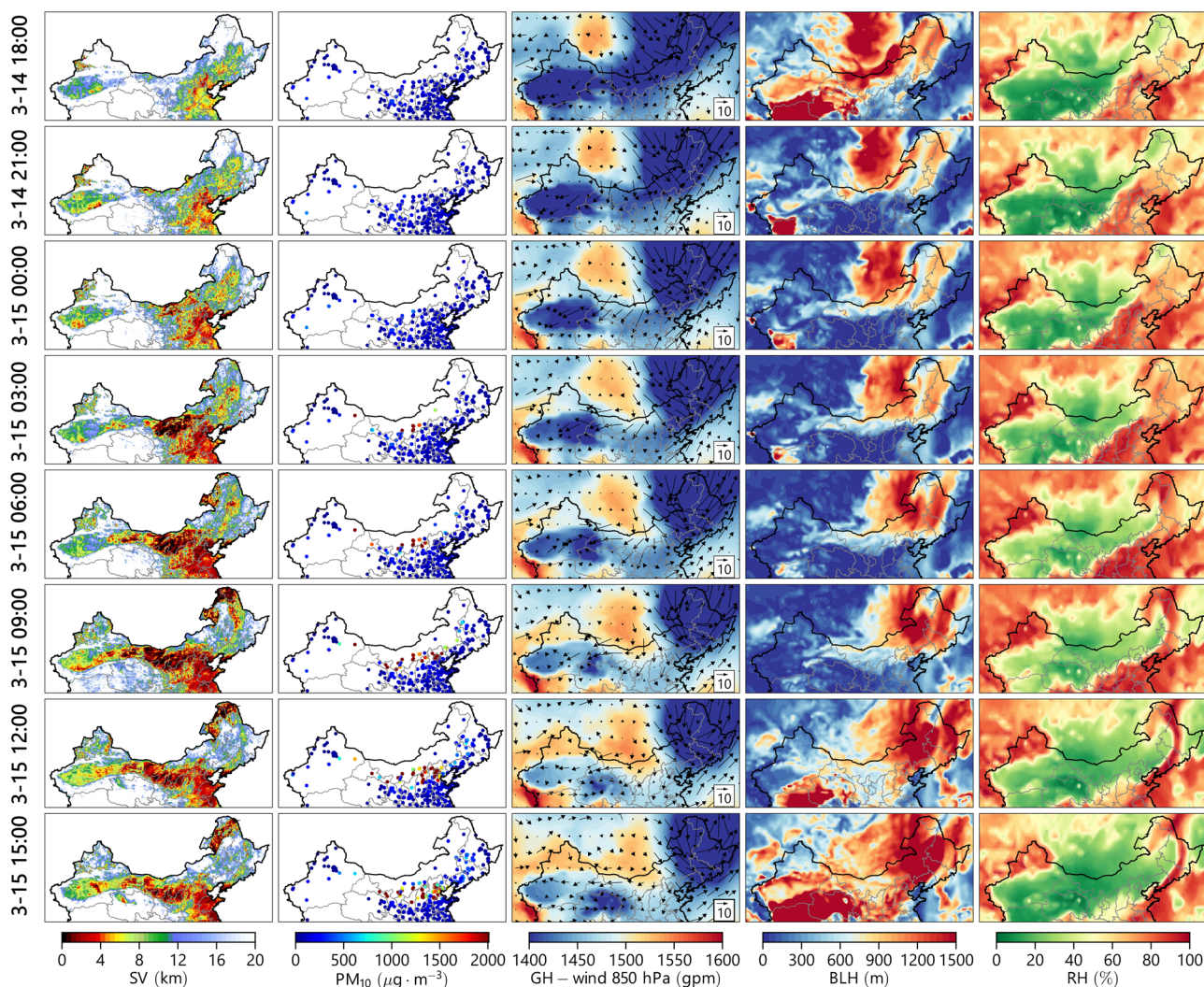


Fig. 6 | Model application to track typical low-SV events. Evolutionary maps of 3-h SV (first column), surface PM₁₀ observations (second column), and meteorological fields (third column: GH and wind field at 850 hPa; fourth column: BLH; fifth column: RH) in a mega dust storm event that occurred during March 14–15, 2021.

formation of these unfavorable meteorological conditions collectively contributed to the rapid accumulation of PM_{2.5} and the subsequent rapid decline in SV. After 10:00 CST on January 22, 2021, as atmospheric conditions led to improved dispersion of pollutants and reduced RH, the regional PM_{2.5} concentration began to decrease, leading to improvement in SV.

Dust event

Figure 6 illustrates the evolution of a mega dust storm event that occurred in northern China during March 14–15, 2021. A recent study, based on long-term satellite observations, registered the intensity of the dust loading during this event as the strongest episode over the past two decades⁵. Thus, examining the retrieved SV products during such an event is beneficial for deepening our understanding of the model’s reproducibility under large-scale, dynamic, and extremely low-SV conditions.

ERA5 data showed that this dust storm event was triggered by a strong Mongolian cyclone on March 14, 2021, in conjunction with a surface cold anticyclonic system. Under the control of such powerful weather systems, a sharp difference in pressure gradient between Mongolia and northeastern China formed, inducing exceptionally strong air movement (wind speed at 850 hPa of up to 20 m s⁻¹), which dragged dust aerosols from the dust source into the atmosphere. Subsequently, these dust aerosols moved rapidly from the west toward the east in association with the trough. In terms of timing, the dust storm entered China from northern-central Inner Mongolia at approximately 21:00 CST on March 14, 2021, and it moved rapidly toward

the NCP region (Fig. 6 and Supplementary Fig. 19). Synergistic analysis showed that our SV products correspond well with ground-based PM₁₀ observations in terms of spatial and temporal variations. Specifically, at 03:00 CST on March 15, 2021, surface PM₁₀ concentrations increased rapidly at several sites in northern China, corresponding to sharp decline in SV in southern Inner Mongolia. At 09:00 CST on March 15, 2021, the dust belt was transported to the NCP and mixed with local PM_{2.5} pollution, further reducing SV. Until 12:00 CST on March 15, 2021, as dust aerosols continued to be deposited, SV in northern China began to improve, but it remained below 5 km in some areas, e.g., southern Inner Mongolia. Overall, the RT-SVR model proved itself capable of accurately tracking the formation, transport, and dissipation processes of large-scale mega dust events both spatially and temporally.

Real-time retrieval capability

We designed the RT-SVR framework with the ability to retrieve hourly SV in real time. Several advantageous factors support this technical realization. First, the two dynamic datasets involved in the model are available in real time, i.e., the PM observations and the CMA Land Data Assimilation System (CLDAS) data. In terms of the PM data, the National Meteorological Information Center releases to the public the most recent hourly national-scale observations in real time (with delay of approximately one-half hour). The assimilation system of the CMA can assimilate quality-controlled meteorological data from tens of thousands of ground-based sites to

generate hourly CLDAS data in real time (with delay of approximately 40 min). Second, although the RT-SVR model is a combination of several base models, each model is lightweight and efficient, and can be deployed on a GPU. As tested on a workstation with a GeForce RTX 3090 graphics card, for each hourly SV retrieval, it was found that the entire RT-SVR framework took approximately 35–40 min (including time waiting for the PM and CLDAS data to become available) in total from the processing of the model input variables to the retrieval, of which model extrapolation took only approximately 12 s. The entire framework is expected to run even faster when running on a higher performance server. Therefore, the above conditions powerfully support the RT-SVR framework to be operationalized to provide real-time seamless hourly SV output in the future.

Discussion

In this study, an operational SV retrieval framework (i.e., RT-SVR) was developed, which takes advantage of the real-time availability of ground-based observations and CLDAS meteorological data to achieve real-time retrieval of spatiotemporally seamless SV data. The retrieval accuracy and reliability of the RT-SVR model were proven satisfactory through HV experiments on more than 20 million data pairs.

The RT-SVR model adopts a double-stacked structure combining multiple ML algorithms and deep residual neural networks, which resolves several problems encountered by individual ML algorithms, including low retrieval accuracy, negative anomalies of retrieval values, and spatial discontinuities. However, the RT-SVR model still has several limitations. First, we noted underestimation of SV levels, as suggested by the slopes of 0.9 and 0.71 between the retrievals and the observations for the CV and HV experiment (Figs. 2b and 3i, respectively). Such underestimation might be related to inaccurate information on dynamic PM pollution, insufficient spatial prediction capabilities, and uncertainties regarding the input predictors. In this study, the seamless $PM_{2.5}$ and PM_{10} background fields were generated through the spatial inverse difference weighting interpolation technique, which caused large spatial bias in some areas with a limited distribution of air quality monitoring sites, e.g., western parts of China. The limited number of SV monitoring sites also resulted in insufficient learning of the spatial information by the model. Moreover, some uncertainties exist regarding the CLDAS data used in the model, and regarding all the other predictors (e.g., anthropogenic emission inventories); these uncertainties, in conjunction with modeling errors, contribute to the uncertainty in the final SV retrievals. It is thus difficult to quantify the uncertainties of the model and its specific sources. Here, we expect that well-estimated air pollutant concentration fields could be developed in the future to further improve model performance. As demonstrated in a previous study^{44–47}, instead of using traditional interpolation methods, more accurate pollutant reanalysis fields can be estimated from site observations using deep learning models to learn multivariable spatial correlations from chemical transport models. Second, compared to other observational instruments, SV observation instruments are special because most have an upper limit (i.e., up to 30 km). This leads to large biases in SV under similar meteorological and PM conditions. Additionally, although rigorous quality control was implemented, automatic SV sensors inherently have different uncertainties at different threshold ranges (e.g., 10% for $SV < 10$ km and 20% for $SV > 10$ km)¹³. This calls for the need to develop distinct models to address the uncertainties that exist within the different ranges of SV observations in future studies.

Despite these limitations, the RT-SVR framework developed in this study has potential for application in several research areas. First, the seamless hourly SV product enables us to provide refined real-time monitoring of various low-SV events (e.g., fog, haze, and dust storms) at horizontal resolution of 6.25 km, which addresses the shortcomings of incomplete spatial coverage of observation sites. Second, once the complex decoupling relationships between SV and the $PM_{2.5}/PM_{10}$ and prevailing meteorological conditions have been resolved, such a product has potential to serve as a key parameter in real-time retrieval of the seamless diurnal cycle of surface $PM_{2.5}/PM_{10}$ pollution in China. This will contribute to resolving the limitations in spatial coverage and temporal resolution of current

retrieval strategies for surface $PM_{2.5}/PM_{10}$ using satellite-based AOD products^{17,23–25,48–50}. Meanwhile, it might also be used for retrieval of key aerosol property parameters such as AOD and single-scattering albedo, as implied by previous studies^{18,19,51,52}. Moreover, with the availability of long-term SV datasets, a short-term forecast model for SV could be built in the future through ML-based spatiotemporal extrapolation techniques.

Methods

Figure 1 illustrates the entire modeling architecture for the real-time surface visibility retrieval (RT-SVR) framework, which is divided into three main steps: multisource data processing and fusing, development of the stacked ensemble model, and model application.

Multisource input data

This data uses multi-source datasets, including in situ observations, meteorological fields, emission inventories, and other ancillary data such as total population, elevation, normalized difference vegetation index (NDVI), and land cover type. The in situ observations comprised surface visibility (SV), fine particulate matter ($PM_{2.5}$), and coarse particulate matter (PM_{10}). The hourly SV observations from 2020 to 2022, recorded at approximately 2400 stations distributed throughout China, were collected from the National Meteorological Information Center (<http://data.cma.cn/>) of the China Meteorological Administration (CMA). To ensure the quality of the SV data, we used only those SV records that were labeled “good” after the quality control procedure. The quality control algorithm for the SV records was proposed by the National Meteorological Information Center to eliminate a few random and systematic errors in the raw data. We collected hourly surface $PM_{2.5}$ and PM_{10} concentrations for the same period from the China National Environmental Monitoring Center network. For each site, we removed $PM_{2.5}/PM_{10}$ outliers that exceeded three standard deviations from the 1-month moving average⁵⁰. The meteorological fields were taken from the CMA Land Data Assimilation System version 2 (CLDAS-V2.0) at resolution of $0.0625^\circ \times 0.0625^\circ$ ⁵³. The CLDAS products are generated by fusing high-density automated surface meteorological observations, multi-satellite retrieval products, and numerical model analysis and forecast fields using multiple grid variational data assimilation techniques, and they have lower uncertainties and higher spatiotemporal resolution (spatial resolution of 6.25 km and temporal resolution of 1 h) in China compared with other similar products (e.g., Global Land Data Assimilation System) used in the international arena^{54,55}. Currently, this product is updated in real time with a delay of <1 h (approximately 40 min). In this study, we utilized the following parameters extracted from the CLDAS dataset: temperature at 2-m height (TEM), specific humidity at 2-m height (SHU), wind speed at 10-m height (WIN), U wind component at 10-m height, V wind component at 10-m height, surface pressure (PRS), and downward surface shortwave radiation (SSRA).

To further improve the retrieval capability of the model, multi-source ancillary data were used. We utilized annual land cover type and monthly NDVI data for 2020 from the Moderate Resolution Imaging Spectroradiometer (MODIS) with spatial resolution of 500 and 250 m, respectively^{56,57}. Population datasets for 2020 with resolution of 30 arcseconds were taken from the Gridded Population of the World (GPW) version 4 dataset and calibrated using the total population reported in China City Yearbooks⁵⁸. Monthly anthropogenic emission inventories for 2020 for multiple species, including primary $PM_{2.5}$ and PM_{10} , sulfur dioxide (SO_2), organic carbon (OC), black carbon (BC), nitrogen oxides (NO_x), ammonia (NH_3), and volatile organic compounds (VOCs), were taken from the Multi-resolution Emission Inventory for China (MEIC) with spatial resolution of $0.25^\circ \times 0.25^\circ$. We also download elevation data at spatial resolution of 300 m from the 2-min Gridded Global Relief Data (ETOPO2). Apart from the features mentioned above, we also considered temporal features (day of year, month, and hour) and spatial features (longitude and latitude) in our model to reduce the effect of spatiotemporal heterogeneity.

The data fusion step in the RT-SVR framework consists of three procedures. First, we removed samples containing missing values from all datasets. Second, PM_{2.5}/PM₁₀ observations and other ancillary data (except for population, elevation, and land cover data; the first two are averaged within the 6.25 km grid and the last are summed within the 6.25 km grid) were interpolated to the same spatial resolution (i.e., 6.25 km) as that of the CLDAS data using the inverse distance weighting method. Finally, we sampled the predictors for every SV site based on geographic location and time using nearest neighbor interpolation. In this way, a dataset containing predictors and observed SV was constructed. The full datasets were divided into two parts: the data from 2020 to 2021 were used to train the RT-SVR model to perform sample-based 5-fold cross-validation, and the data from 2022 were used to conduct a hindcast-validation experiment to evaluate the generalization capability of the model.

Stacked ensemble model description

A stacked ensemble model (RT-SVR), incorporating multiple machine learning (ML) algorithms and a deep learning module, was developed in this study to generate seamless hourly SV in China, as illustrated in Fig. 1. Overall, the RT-SVR model consists of two structural layers. In the first layer, five base ML models are used to establish the relationships between the predictors and SV. These five base models comprise the Multilayer Perceptron (MLP)⁵⁹, Random Forest (RF)⁶⁰, Categorical Boosting (CatBoost)⁶¹, eXtreme Gradient Boosting (XGBoost)⁶², and Light Gradient Boosting Machine (LightGBM)⁶³. We chose five base models with different principles because the purpose of stacked ensemble ML is that the diversity of base models needs to be ensured⁶⁴. The MLP model is a deep network that fits the target through the backpropagation method⁵⁹. The RF model is a type of bagging algorithm that reduces fitting errors through use of the bootstrap aggregating method^{60,65}. The boosting models (CatBoost, XGBoost, and LightGBM) are currently the tree methods used most commonly for retrieving surface pollutants (e.g., PM_{2.5}, PM₁₀, and ozone) because the advantage of the gradient descent method can achieve better results⁶⁶. In addition, in our experiments, we found that due to the complexity of the relationship between SV, PM, and meteorological elements, traditional statistical-based methods for constructing models are less effective in retrieving SV (e.g., linear regression models, generalized additive model, etc.). To achieve better performance, we performed hyperparameter tuning for each base model. Supplementary Table 1 summarizes the key parameter information of the five base models finally selected for this study.

We found that different base models have distinct advantages and disadvantages in retrieving the SV in this study, depending on their accuracy and spatial performance (see Results). Therefore, a deep residual network was developed in the second layer to fully utilize (eliminate) the different advantages (disadvantages) of the base models in the first layer. This residual network comes from the transformer model, and its fixed feature size and residual connection method can effectively resolve the problem of overfitting in the stacked ensemble model^{67,68}.

In the residual feedforward network, the encoder block is repeated five times to fully learn the features from the different base models. We employ feature embedding and subsequent mapping in each block using the following equation:

$$G(x) = \max(0, xw_1 + b_1)w_2 + b_2 \quad (1)$$

where max(x) represents the Rectified Linear Unit (ReLU) function. w represents the weight, and b represents the bias in the feedforward network.

Compared to the models presented in previous studies, our model has two major advantages. In the first layer, conventional gradient boosting models typically employ the square error as the loss function, which results in the same punishment for underestimation and overestimation for the model. In experimental tests, we found that the estimated SV from the boosting model suffered from unrealistically frequent negative retrievals. To address this issue, we modified the gradient and hessian of the loss function

in the boosting model:

$$\frac{\partial[L(Y|f(x))]}{\partial[f(x)]} = -2 \bullet (Y_i - f(x_i)) \quad (2)$$

$$\frac{\partial[L(Y|f(x))]}{\partial[f(x)]} = \begin{cases} -2\alpha \bullet (Y_i - f(x_i)), & f(x_i) < Y_i \\ -2\beta \bullet (Y_i - f(x_i)), & f(x_i) \geq Y_i \end{cases} \quad (3)$$

$$\frac{\partial^2[L(Y|f(x))]}{\partial[f(x)]^2} = \begin{cases} 2\alpha, & f(x_i) < Y_i \\ 2\beta, & f(x_i) \geq Y_i \end{cases} \quad (4)$$

Equation 2 represents the gradient of the conventional loss function used in boosting models. We add two penalty coefficients α and β in the gradient (Eq. 3) and hessian function (Eq. 4) to constraint the extreme values obtained from the boosting model. $f(x)$ and Y_i represent the retrieved and observed SV respectively. In order to make the loss function have a penalizing effect on the low values of the retrieval SV, α was always kept greater than β during the experiments. In the field of artificial intelligence, the loss function serves as a crucial metric for quantifying the disparity between the predicted outputs of a model and the target values. The optimization of ML models often involves the minimization of this loss function to enhance the model's ability to accurately fit the training data. The gradient informs about the rate of change of the loss function at the current parameter values, indicating the direction in which the loss function is ascending or descending most rapidly. The hessian derivative provides insights into the curvature of the loss function⁶⁹. During the model training process, we modified the gradient method of XGBoost, whilst keeping CatBoost and LightGBM unchanged.

Additionally, compared to the traditional stacked ensemble model constructed with a simple statistical model (e.g., generalized additive model, linear regression model) in the second layer, our meta model in the second layer has the following advantages. First, compared with simple statistical models, residual feedforward networks are better able to capture the nonlinear relationships between predictors and the target by using nonlinear activation functions and residual connections between layers⁷⁰. Second, deep residual networks have been theoretically demonstrated to possess superior expressive capacity, enabling them to approximate more complex functions with greater accuracy⁷⁰. This endows them with an advantage in handling large-scale and highly nonlinear datasets. Moreover, in the first layer, there are boosting models with negative outputs, which can have a large impact on the linear models. Thus, the use of this method improves the accuracy of the stacked ensemble model and maintains the same retrieval time.

Operational process of the RT-SVR framework

In the model application stage (step 3 in Fig. 1), the well-trained RT-SVR model in step 2 can be directly deployed to a cloud server, such as the CMA meteorological big data cloud platform “Tianqing”⁷¹. Benefiting from the real-time access to hourly PM_{2.5} and PM₁₀ observations and CLDAS meteorological fields on the “Tianqing” platform, the RT-SVR model can generate seamless hourly SV in real time once these data are made available and automatically processed spatiotemporally to a uniform grid resolution (6.25 km) to form an input predictor matrix.

Data availability

The PM and SV observations are accessible at <https://air.cnemc.cn:18007/> and <http://data.cma.cn/data/detail/dataCode/A.0012.0001.html>, respectively. The CLDAS meteorological fields are accessible at http://data.cma.cn/data/cdcdetail/dataCode/NAFP_CLDAS2.0_RT.html. The land cover type and NDVI data are available at <https://lpdaac.usgs.gov/products/mcd12c1v006/> and <https://lpdaac.usgs.gov/products/myd13q1v006/> respectively. The population data is available at <https://beta.sedac.ciesin.columbia.edu>. The elevation data is accessible at <https://www.ncei.noaa.gov/products/etopo-global-relief-model>. The emission data across China is accessible at <http://meicmodel.org>.

Code availability

All codes needed to perform the analyses are available upon reasonable request from the corresponding author (guik@cma.gov.cn).

Received: 25 December 2023; Accepted: 4 March 2024;

Published online: 13 March 2024

References

- Gultepe, I. et al. Fog research: a review of past achievements and future perspectives. *Pure Appl. Geophys.* **164**, 1121–1159 (2007).
- Han, Y. Q. & Zhu, T. Health effects of fine particles (PM_{2.5}) in ambient air. *Sci. China Life Sci.* **58**, 624–626 (2015).
- Kan, H., Chen, B., Chen, C., Fu, Q. & Chen, M. An evaluation of public health impact of ambient air pollution under various energy scenarios in Shanghai, China. *Atmos. Environ.* **38**, 95–102 (2004).
- Silva, R. A. et al. Future global mortality from changes in air pollution attributable to climate change. *Nat. Clim. Chang.* **7**, 647–651 (2017).
- Gui, K. et al. Record-breaking dust loading during two mega dust storm events over northern China in March 2021: aerosol optical and radiative properties and meteorological drivers. *Atmos. Chem. Phys.* **22**, 7905–7932 (2022).
- Shen, H. et al. Urbanization-induced population migration has reduced ambient PM_{2.5} concentrations in China. *Sci. Adv.* **3**, 1–13 (2017).
- Tao, J., Zhang, L., Cao, J. & Zhang, R. A review of current knowledge concerning PM_{2.5} chemical composition, aerosol optical properties and their relationships across China. *Atmos. Chem. Phys.* **17**, 9485–9518 (2017).
- Chen, S. L., Chang, S. W., Chen, Y. J. & Chen, H. L. Possible warming effect of fine particulate matter in the atmosphere. *Commun. Earth Environ.* **2**, 1–9 (2021).
- Song, Z., Wang, M. & Yang, H. Quantification of the impact of fine particulate matter on solar energy resources and energy performance of different photovoltaic technologies. *ACS Environ. Au.* **2**, 275–286 (2021).
- Chen, Z. et al. Influence of meteorological conditions on PM_{2.5} concentrations across China: A review of methodology and mechanism. *Environ. Int.* **139**, 105558 (2020).
- Yang, H., Peng, Q., Zhou, J., Song, G. & Gong, X. The unidirectional causality influence of factors on PM_{2.5} in Shenyang city of China. *Sci. Rep.* **10**, 1–12 (2020).
- Wang, X., Zhang, R. & Yu, W. The effects of PM_{2.5} concentrations and relative humidity on atmospheric visibility in Beijing. *J. Geophys. Res. Atmos.* **124**, 2235–2259 (2019).
- Fei, Y. et al. Spatiotemporal variability of surface extinction coefficient based on two-year hourly visibility data in mainland China. *Atmos. Pollut. Res.* **10**, 1944–1952 (2019).
- Gui, K. et al. Construction of a virtual PM_{2.5} observation network in China based on high-density surface meteorological observations using the extreme gradient boosting model. *Environ. Int.* **141**, 105801 (2020).
- Zhong, J. et al. Robust prediction of hourly PM_{2.5} from meteorological data using LightGBM. *Natl. Sci. Rev.* **8**, nwaa307 (2021).
- Zeng, Z. et al. Estimating hourly surface PM_{2.5} concentrations across China from high-density meteorological observations by machine learning. *Atmos. Res.* **254**, 105516 (2021).
- Chen, B. et al. Estimation of atmospheric PM₁₀ concentration in China using an interpretable deep learning model and top-of-the-atmosphere reflectance data from china's new generation geostationary meteorological satellite, FY-4A. *J. Geophys. Res. Atmos.* **127**, 1–20 (2022).
- Wu, J. et al. Improvement of aerosol optical depth retrieval using visibility data in China during the past 50 years. *J. Geophys. Res. Atmos.* **119**, 13,370–13,387 (2014).
- Lin, J. & Li, J. Spatio-temporal variability of aerosols over East China inferred by merged visibility-GEOS-Chem aerosol optical depth. *Atmos. Environ.* **132**, 111–122 (2016).
- Qu, W., Zhang, X., Wang, Y. & Fu, G. Atmospheric visibility variation over global land surface during 1973–2012: Influence of meteorological factors and effect of aerosol, cloud on ABL evolution. *Atmos. Pollut. Res.* **11**, 730–743 (2020).
- Pitchford, M. et al. Revised algorithm for estimating light extinction from IMPROVE particle speciation data. *J. Air Waste Manag. Assoc.* **57**, 1326–1336 (2007).
- Hu, B., Zhang, X., Sun, R. & Zhu, X. Retrieval of horizontal visibility using MODIS data: a deep learning approach. *Atmosphere* **10**, 1–15 (2019).
- Bai, K. et al. LGHAP: the long-term gap-free high-resolution air pollutant concentration dataset, derived via tensor-flow-based multimodal data fusion. *Earth Syst. Sci. Data* **14**, 907–927 (2022).
- Geng, G. et al. Tracking air pollution in China: near real-time PM_{2.5} retrievals from multisource data fusion. *Environ. Sci. Technol.* **55**, 12106–12115 (2021).
- Wei, J. et al. Reconstructing 1-km-resolution high-quality PM_{2.5} data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. *Remote Sens. Environ.* **252**, 112136 (2021).
- Chen, Z. Y. et al. Extreme gradient boosting model to estimate PM_{2.5} concentrations with missing-filled satellite data in China. *Atmos. Environ.* **202**, 180–189 (2019).
- Zheng, C. et al. Analysis of influential factors for the relationship between PM_{2.5} and AOD in Beijing. *Atmos. Chem. Phys.* **17**, 13473–13489 (2017).
- Guo, J. et al. Impact of diurnal variability and meteorological factors on the PM_{2.5} - AOD relationship: implications for PM_{2.5} remote sensing. *Environ. Pollut.* **221**, 94–104 (2017).
- Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **23**, 18932–18943 (2021).
- Ciampiconi, L., Elwood, A., Leonardi, M., Mohamed, A. & Rozza, A. A survey and taxonomy of loss functions in machine learning. Preprint at <https://arxiv.org/abs/2301.05579> (2023).
- Pereira, J. P. B., Stroes, E. S. G., Zwinderman, A. H. & Levin, E. Covered information disentanglement: model transparency via unbiased permutation importance. *Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022* **36**, 7984–7992 (2022).
- Kroll, J. H. & Seinfeld, J. H. Chemistry of secondary organic aerosol: formation and evolution of low-volatility organics in the atmosphere. *Atmos. Environ.* **42**, 3593–3624 (2008).
- Duan, J. et al. Summertime and wintertime atmospheric processes of secondary aerosol in Beijing. *Atmos. Chem. Phys.* **20**, 3793–3807 (2020).
- Fei, Y., Liao, J. & Zhang, Z. Consistency and discrepancy between visibility and PM_{2.5} measurements: potential application of visibility observation to air quality study. *Sensors* **23**, 898 (2023).
- Xu, W. et al. Current challenges in visibility improvement in southern China. *Environ. Sci. Technol. Lett.* **7**, 395–401 (2020).
- Yang, B., Bräuning, A., Zhang, Z., Dong, Z. & Esper, J. Dust storm frequency and its relation to climate changes in Northern China during the past 1000 years. *Atmos. Environ.* **41**, 9288–9299 (2007).
- Zhang, J. et al. Prolonged drought enhances northwest China dust storm activity. *J. Geophys. Res. Atmos.* **127**, e2022JD037088 (2022).
- Bai, L., Chen, G. & Huang, L. Convection initiation in monsoon coastal areas (South China). *Geophys. Res. Lett.* **47**, 1–11 (2020).
- Chen, G., Wang, B. & Liu, J. Study on the sensitivity of initial perturbations to the development of a vortex observed in Southwest China. *J. Geophys. Res. Atmos.* **126**, e2021JD034715 (2021).
- Lakra, K. & Avishek, K. *A Review on Factors Influencing Fog Formation, Classification, Forecasting, Detection and Impacts. Rendiconti Lincei* vol. 33 (Springer International Publishing, 2022).
- He, Q., Wang, M. & Yim, S. H. L. The spatiotemporal relationship between PM_{2.5} and aerosol optical depth in China: Influencing factors

- and implications for satellite PM_{2.5} estimations using MAIAC aerosol optical depth. *Atmos. Chem. Phys.* **21**, 18375–18391 (2021).
42. Lou, C. et al. Relationships of relative humidity with PM_{2.5} and PM₁₀ in the Yangtze River Delta, China. *Environ. Monit. Assess.* **189**, 582 (2017).
 43. Cheng, Y. et al. Humidity plays an important role in the PM_{2.5} pollution in Beijing. *Environ. Pollut.* **197**, 68–75 (2015).
 44. Lyu, B., Huang, R., Wang, X., Wang, W. & Hu, Y. Deep-learning spatial principles from deterministic chemical transport models for chemical reanalysis: An application in China for PM_{2.5}. *Geosci. Model Dev.* **15**, 1583–1594 (2022).
 45. Ding, Y. et al. Retrieving hourly seamless PM_{2.5} concentration across China with physically informed spatiotemporal connection. *Remote Sens. Environ.* **301**, 113901 (2024).
 46. Song, G., Li, S. & Xing, J. Lightning nowcasting with aerosol-informed machine learning and satellite-enriched dataset. *npj Clim. Atmos. Sci.* **6**, 1–10 (2023).
 47. Song, G. et al. Surface UV-assisted retrieval of spatially continuous surface ozone with high spatial transferability. *Remote Sens. Environ.* **274**, 112996 (2022).
 48. Chen, G. et al. Spatiotemporal patterns of PM₁₀ concentrations over China during 2005–2016: a satellite-based estimation using the random forests approach. *Environ. Pollut.* **242**, 605–613 (2018).
 49. Yang, Q., Yuan, Q., Li, T. & Yue, L. Mapping PM_{2.5} concentration at high resolution using a cascade random forest based downscaling model: evaluation and application. *J. Clean. Prod.* **277**, 123887 (2020).
 50. Yan, X. et al. Cooperative simultaneous inversion of satellite-based real-time PM_{2.5} and ozone levels using an improved deep learning model with attention mechanism. *Environ. Pollut.* **327**, 121509 (2023).
 51. Zhang, Z. et al. Aerosol optical depth retrieval from visibility in China during 1973–2014. *Atmos. Environ.* **171**, 38–48 (2017).
 52. Dong, Y. et al. Retrieval of aerosol single scattering albedo using joint satellite and surface visibility measurements. *Remote Sens. Environ.* **294**, 113654 (2023).
 53. Shi, C., Jiang, L., Zhang, T., Xu, B. & Han, S. Status and Plans of CMA Land Data Assimilation System (CLDAS) Project. *EGU Gen. Assem. Conf. Abstr.* **16**, 5671 (2014).
 54. Han, S. et al. Evaluation of CLDAS and GLDAS datasets for near-surface air temperature over major land areas of China. *Sustainability* **12**, 4311 (2020).
 55. Shi, C. et al. A review of multi-source meteorological data fusion products. *Acta Meteorol. Sin.* **77**, 774–783 (2019).
 56. Li, Z. et al. Accuracy assessment of land cover products in China from 2000 to 2020. *Sci. Rep.* **13**, 1–11 (2023).
 57. Tucker, C. J. & Sellers, P. J. Satellite remote sensing of primary production. *Int. J. Remote Sens.* **7**, 1395–1416 (1986).
 58. Lloyd, C. T. High resolution global gridded data for use in population studies. *Sci. Data* **42**, 117–120 (2017).
 59. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
 60. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
 61. Doroogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. Preprint at <https://arxiv.org/abs/1810.11363v1> (2018).
 62. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **13**–17(August-2016), 785–794 (2016).
 63. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, 3147–3155 (2017).
 64. Mohammed, A. & Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ. - Comput. Inf. Sci.* **35**, 757–774 (2023).
 65. Islam, P., Khosla, S., Lok, A. & Saxena, M. Analyzing Bagging Methods for Language Models. Preprint at <https://arxiv.org/abs/2207.09099v1> (2022).
 66. Ferreira, A. J. & Figueiredo, M. A. T. boosting algorithms: a review of methods, theory, and applications BT - ensemble machine learning: methods and applications. in (eds. Zhang, C. & Ma, Y.) 35–85 (Springer New York, New York, NY, 2012).
 67. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**(December), 5999–6009 (2017).
 68. Touvron, H. et al. ResMLP: feedforward networks for image classification with data-efficient training. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 5314–5321 (2023).
 69. Wang, Q., Ma, Y., Zhao, K. & Tian, Y. A comprehensive survey of loss functions in machine learning. *Ann. Data Sci.* **9**, 187–212 (2022).
 70. Muchlinski, D. Machine learning and deep learning. Preprint at <https://arxiv.org/abs/2104.05314v2> (2022).
 71. Wang, P. et al. Design and implementation of multi-element integration platform based on tianqing data. *Meteorol. Environ. Res.* **13**, 41–42 (2022).

Acknowledgements

This research was supported by grants from the National Natural Science Foundation of China project (42175153 and 42030608), the Science and Technology Plan Project of CMA (CMAJBGS202325), Third Xinjiang Scientific Expedition Program (2022xjkk0903), National Science Fund for Distinguished Young Scholars (41825011), Young Elite Scientists Sponsorship Program by BAST, and Basic Research Fund of CAMS (2021Y001 and 2023Z021).

Author contributions

H.C., K.G. and X.Z. designed this study. X.Z. developed the retrieval framework and drafted the paper with the help of H.C. and K.G., K.G. performed the data analysis work; N.S., Y.F., Y.L., H.Z., W.Y., Y.L., Y.Z., L.L., H.W., Z.W. and X.Z. provided constructive suggestions regarding this study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41612-024-00617-1>.

Correspondence and requests for materials should be addressed to Ke Gui or Huizheng Che.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024