SCIENTIFIC REPORTS

OPEN

Received: 6 March 2018
Accepted: 22 October 2018
Published online: 02 November 2018

Forecasting extreme atmospheric events with a recurrence-interval-analysis-based autoregressive conditional duration model

Yue-Hua Dai[1], Zhi-Qiang Jiang[1,2] & Wei-Xing Zhou[1,2,3]

With most city dwellers in China subjected to air pollution, forecasting extreme air pollution spells is of paramount significance in both scheduling outdoor activities and ameliorating air pollution. In this paper, we integrate the autoregressive conditional duration model (ACD) with the recurrence interval analysis (RIA) and also extend the ACD model to a spatially autoregressive conditional duration (SACD) model by adding a spatially reviewed term to quantitatively explain and predict extreme air pollution recurrence intervals. Using the hourly data of six pollutants and the air quality index (AQI) during 2013–2016 collected from 12 national air quality monitoring stations in Beijing as our test samples, we attest that the spatially reviewed recurrence intervals have some general explanatory power over the recurrence intervals in the neighbouring air quality monitoring stations. We also conduct a one-step forecast using the RIA-ACD(1,1) and RIA-SACD(1,1,1) models and find that 90% of the predicted recurrence intervals are smaller than 72 hours, which justifies the predictive power of the proposed models. When applied to more time lags and neighbouring stations, the models are found to yield results that are consistent with reality, which evinces the feasibility of predicting extreme air pollution events through a recurrence-interval-analysis-based autoregressive conditional duration model. Moreover, the addition of a spatial term has proved effective in enhancing the predictive power.

Regardless of the air quality monitoring stations' commitment to present the latest air quality reports and exhaustive air quality information, Chinese residents are tending to suffer a longer stretch of stifling air pollution periods[1]. As such, a high premium has been put on explaining and forecasting the occurrence of extreme atmospheric events due to their influence on people's daily life and health[2,3]. Unlike other extreme events, the occurrence of extreme air pollution events is rarely studied for the following reasons[2]. First, the gaps between extreme value theory and its applications in atmospheric time series still exist. Second, the air monitoring stations have begun to offer high-frequency data only in recent years. In this paper, to model the recurrence intervals of extreme air quality events, we resort to the method that is widely used in modeling the financial market risk[4–6].

Research efforts to evaluate and predict the occurrence of extreme air pollution have been made from various perspectives[2,3,7]. Wang *et al.* established an early-warning system using a hybrid forecasting model based on some data processing methods (such as support vector machine and fuzzy set theory)[2,3]. Whereas their quest for optimal distributions to model the air pollution time series is partially similar to ours, differences do exist in the models to deal with the distributions. We gravitate towards the recurrence frequency of the extreme events and their spatiotemporal properties yet they followed with interest the distributions and model selections in evaluating the time series. Niu *et al.* proposed the ensemble empirical mode decomposition and least square support vector machine (EEMD-LSSVM) method based on phase space reconstruction (PSR) to analyze the pollution time series[7]. Their results boast a higher predictive accuracy when applied in Lanzhou and Guangzhou. Other relevant studies adopt multifractal analysis to check the long-term or short-term structure and self-organized properties of recurrence intervals. More importantly, with both numerical and model-based analytical approaches, prediction making and variants of Value-at-Risk estimates have been discussed intensively in recent literature based on the

[1]School of Business, East China University of Science and Technology, Shanghai, 200237, China. [2]Research Center for Econophysics, East China University of Science and Technology, Shanghai, 200237, China. [3]Department of Mathematics, East China University of Science and Technology, Shanghai, 200237, China. Correspondence and requests for materials should be addressed to W.-X.Z. (email: wxzhou@ecust.edu.cn)

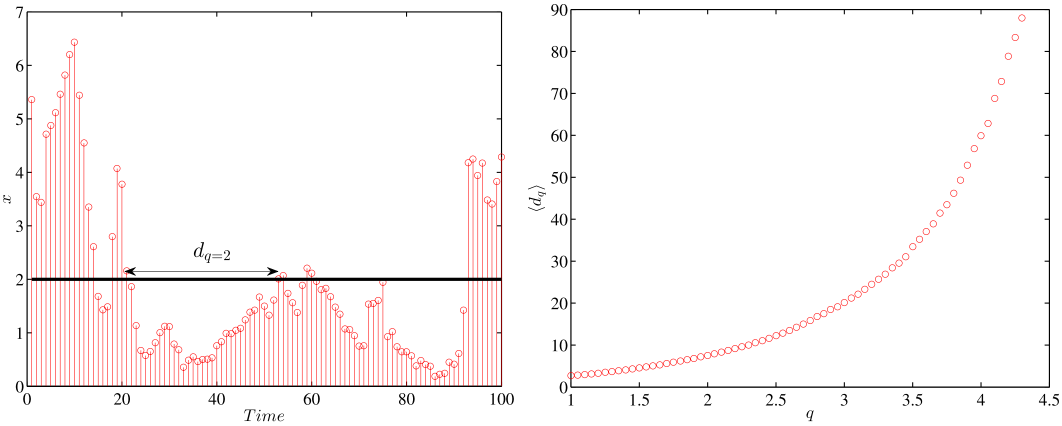

Figure 1. An illustrative example of the recurrence interval d_q (left panel) and the relationship between q and $\langle d_q\rangle$ (right panel). The selected x is diurnally adjusted $PM_{2.5}$ time series at station 1.

statistics of return intervals, conditional intervals and event expectation times[8–11]. However, we steer the wheel to another direction which focuses on the ARMA structure and spatial explanatory power. It's worth to mention that Beck and Cohen put forward a two-compound superstatistical model to model multiple fragments characterized by the exponential inter-session time distribution, which has been widely used in complex systems for risk estimation[11–15].

In this paper, we integrate the recurrence interval analysis with the autoregressive conditional duration model to measure the recurrence statistics of extreme air pollution events, which accords with Herrera and Schipp's methodology to predict the value at risk of stock market index[16]. In addition, peaks over threshold (POT) method is also employed to generate POT series and the autoregressive conditional duration (ACD) model is similarly applied to analyze the generated series. In this regard, our paper follows Herrera and Schipp's framework and tries to apply the ACD model into the recurrence intervals analysis. However, given the spatial characteristics of the air pollution time series, we incorporate the spatial term into the ACD model and attempt to unveil the spatial predictive power of neighbouring stations. In this study, we first analyze the basic statistics of recurrence interval series under different thresholds. We also propose the spatially reviewed recurrence intervals by adding another monitoring station as a benchmark to fully incorporate spatial reference information. Using the spatially reviewed recurrence interval as an exogenous variable in the conditional duration model, this paper extends the ACD model to a spatial ACD model. With the help of maximum likelihood estimation (MLE), we find that some coefficients of the spatial term are significant at 1% level, which indicates a strong predictive power of the spatially reviewed recurrence intervals. On top of that, this paper also goes from the simplest scenario to more temporal lags and spatially neighbouring stations to evaluate how the predictive power changes with time and distance. Finally, in this paper, we conduct a one-step forward forecast using both the classic ACD model and the spatial ACD model. Statistics show that 90% of the predicted recurrence intervals deviate by less than 72 hours from the actual value, which manifests the predictive power of the proposed model. Moreover, there is evidence that the spatial ACD model is slightly better than the ACD model due to the use of spatial information in predicting extreme pollution events.

Results

Recurrence intervals.

A recurrence interval analysis pertains to the time intervals between two consecutive extreme events[17,18]. It has been widely studied both in nature science[18–21] and social science[4,22,23]. For a given pollutant and threshold, long intervals indicate its inactivity which may then signify a period of weak industrial activity and gentle meteorological fluctuations. Conversely, increased industrial activities and unstable climatic conditions may give rise to excess concentration of the pollutant and result in shorter intervals. The dynamic behavior of the durations thus contains valuable information about both the human activities and meteorological changes.

The recurrence interval d_q is defined as the waiting time between two consecutive events in which the normalized concentration x exceeds the threshold q.

$$d_q = \min\{t - t'\colon x(t) > q,\ x(t') > q,\ t > t'\} \tag{1}$$

In this paper, the selected q values range from 2.0 to 4.0 with an increment of 0.5. Generally speaking, when $q = 2.0$, the air condition is categorized as mild pollution, whereas when $q = 4.0$, it is labelled as serious pollution. Specifically, judging from Eq. (2) and Fig. 1(b), it's apparent that qs in this bulk account for around 10% to 1% extreme air pollution conditions[5]

$$\frac{1}{\langle d_q\rangle} = \int_q^{+\infty} p(x)dx \tag{2}$$

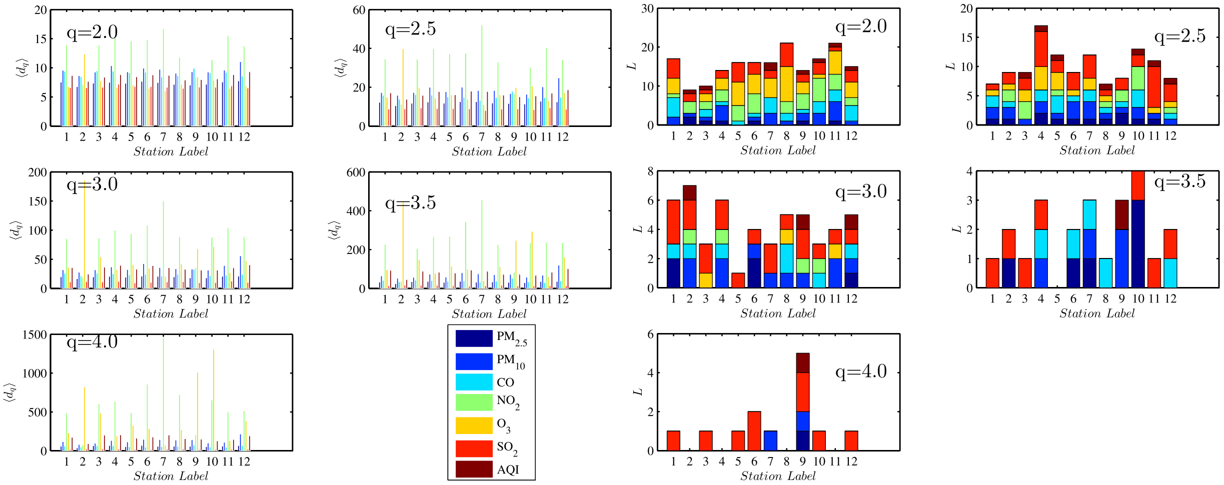

Figure 2. Average recurrence intervals (left panel) and AR(L) models (right panel) under five different *q*s. In the right panel, under each threshold *q*, we first obtain the recurrence interval series d_q and then model d_q using the autoregressive model. The maximal significant (95% confidence interval) lag is denoted as *L*. For one station, the *L*s of the six pollutants in question and AQI's d_qs are stacked in bars with distinct colors.

When it comes to different pollutants under each threshold, Fig. 2 presents how the average recurrence interval $\langle d_q \rangle$ varies and how temporally persistent the recurrence interval series are. Generally, the recurrence intervals of NO_2 and O_3 are substantially higher than those of other pollutants, and this trend is especially true of higher *q*s. Thereby it can be concluded that fine particulate matters are the main pollutants of air pollution and this is congruent with the previous findings[24,25]. Apart from these two pollutants, pollution stemming from four other pollutants happens about every 10 hours on average and this tendency may be attributed to the diurnally cyclical patterns of air pollutants[26]. As *q* rises from 2 to 4, the recurrence intervals for these four pollutants are almost doubled. However, NO_2 and O_3 show a different picture, with the average recurrence intervals around 15 hours when $q = 2$ and rocketing to 1000 hours when $q = 4$, which suggests that these two pollutants are marginal in air pollution. The average recurrence intervals are also scrutinized by the monitoring stations. Figure 2(a) shows that all these stations share similar recurrence intervals on the aforementioned four pollutants and differ in the average recurrence intervals of NO_2 and O_3. In station 2, the average recurrence intervals for O_3 outweigh all other pollutants, and this is true for stations 9 and 10 when *q* rises. Finally, AQI, as a comprehensive air quality index, overtops $q = 2$ every 10 hours, and the interval is doubled when *q* increases by 0.5. This trend strong evidence that air pollution deserves to be put on the list of priorities[1].

On the temporal side, autoregressive model AR(L) is employed to fit d_qs for each adjusted time series to identify the autocorrelation structure of recurrence intervals. The first *L* significant (under 95% confidence interval) lags for d_qs are stacked in Fig. 2(b). The general trend shows that as the threshold increases, the autocorrelation is weakening across both monitoring stations and pollutant categories. When $q = 4$, SO_2 becomes the only pollutant whose recurrence intervals still autocorrelate. Two possible reasons may account for this. Firstly, the threshold is selected mainly based on normalized AQI and may be somewhat lower for normalized SO_2 series, yet this alternative is ruled out after a scrutiny of the original data. Secondly, it's more likely that the SO_2's autocorrelation structure is persistent compared with other time series[27,28]. As the threshold *q* climbs up to 3, the autocorrelations of most air pollutants' recurrence intervals are diminishing, which inspires our exploration of the correlation structure from the spatial side.

Scaling behaviour of the recurrence intervals. As many other natural phenomena[29–31], in this section, we try to check whether there exist scaling properties in the original time series and recurrence intervals series of the air pollutants. In Fig. 3, it's straightforward that most of the recurrence intervals occur between $0.01\langle d_q \rangle$ and $0.1\langle d_q \rangle$ and much smaller than $\langle d_q \rangle$. Longer inter-event intervals are less likely to occur. However, due to long-term correlation of the concentration time series of $PM_{2.5}$[26], the intervals series show only several unique numbers, making it hard to capture the term structure of the recurrence intervals series of the air pollutants. When it comes to different thresholds, the scaling behavior still shows no traceable patterns and almost 99% of the recurrence intervals are 1 hour, indicating that the air pollution is followed by one another and recur in clusters.

When checked the long-term correlation of the air pollutants at different locations, we do believe the finite sample size and data quantization effects influence the results[32]. Factually, sample size and data quantization affect the results through several channels. First, as *q* is lifted, less recurrence data will be generated, and the long-term correlation structure is influenced. Especially, as $q = 4$, the precision of the long-term correlation is certainly reduced, the choice between long-term correlation and short-term correlation is thus affected. Secondly, as *q* is rising, less data will lead more bias and convergence problems when calculating the Autoregressive Conditional Duration Model. To consolidate the influence of the sample size effect, we use the original data to check the long-term correlation, and as evinced in Fig. 4, when the threshold *q* is very high, the sample size is shortened,

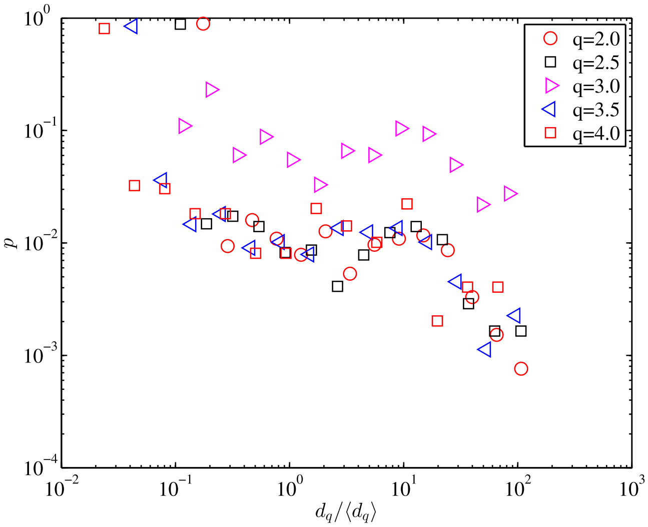

Figure 3. The probability of $d_q/\langle d_q\rangle$ of the $PM_{2.5}$ series in station one.

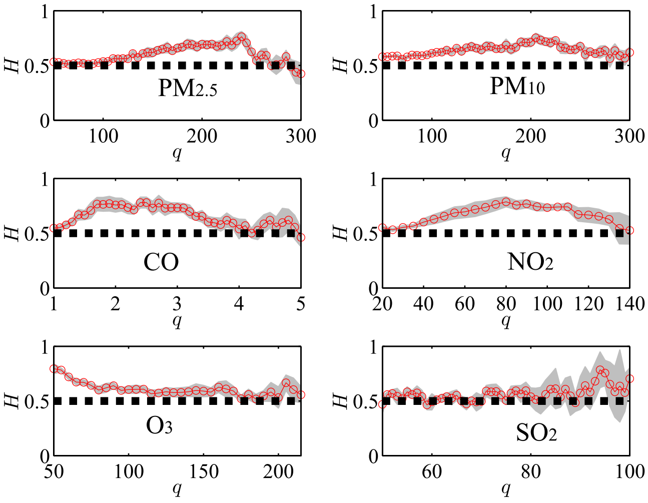

Figure 4. The Hurst exponent and the threshold for the six pollutants at Station 1.

and the Hurst exponent will be indistinguishable from 0.5 which signals that it is the very hard to identify the term structure as the sample size is limited.

Spatially reviewed recurrence intervals. In this part, we define the recurrence intervals from the spatial perspective. For two locations i and j, we define j's spatially reviewed recurrence intervals using i's recurrence intervals as the time breakpoint: Whenever location i's recurrence ends, the most recent recurrence interval at the same threshold q at location j is denoted as j's spatially reviewed recurrence interval $d_q^{(ij)}$. In this definition, we introduce i as a benchmark. Whenever the extreme pollution spell terminates, people at location i will receive a message from location j about how long the most recent recurrence lasts in j under the same threshold q. As Fig. 5(a) illustrates, these two locations will have observed their last recurrence intervals under the same threshold q. A problem may arise when t is too small, and location i has observed an occurrence while location j has not. In this case, we, ad hoc, adopt the average interval at location j as its spatially reviewed recurrence interval. In this way we can generate two equal lengths of recurrence interval series $d_q^{(i)}$ and $d_q^{(ij)}$. It should be noted that $d_q^{(ij)}$ is not exactly the same as $d_q^{(j)}$ at location j without any spatially reference introduced if generated with the above method. However, they will remain in high "correlation" with the single time series' recurrence intervals. Nonetheless this correlation is hard to measure because of the unequal lengths of the two series.

We simulate two independent standard distribution series x_i and x_j with correlation c_{x_i,x_j} varying from −1 to 1. Figure 5(b) shows i's recurrence interval series $d_q^{(i)}$ and j's spatially reviewed recurrence intervals $d_q^{(ij)}$ are of no significant correlation when c_{x_i,x_j} is lower than 0.5. However, when c_{x_i,x_j} is higher than 0.5, their recurrence intervals apparently becomes more correlated, which indicates that if the two original time series are of high correlations under the identical mean, variance and distribution conditions, the recurrence intervals of one series are

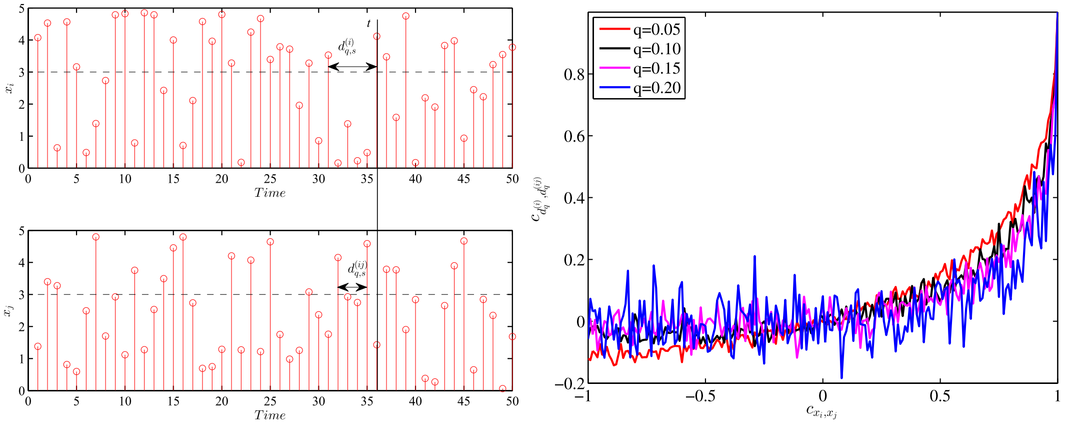

Figure 5. An illustrative example of $d_{q,s}^{(i)}$ and its spatially reviewed recurrence intervals $d_{q,s}^{(ij)}$, $q = 3.0$ in this example (left panel) and simulated relationship between the original time series' correlations and recurrence interval correlations under different thresholds (right panel). In the right panel, first we simulate standard distribution time series pairs and then use the spatially reviewed recurrence intervals scheme to obtain the $d_q^{(i)}$ and $d_q^{(ij)}$ and calculate the correlation.

	1	2	3	4	5	6	7	8	9	10	11	12
Panel A: Minimum correlation matrix between stations.												
1	1.00	0.60	0.91	0.89	0.89	0.92	0.85	0.73	0.64	0.68	0.87	0.83
2	0.60	1.00	0.59	0.59	0.59	0.61	0.61	0.63	0.74	0.80	0.58	0.63
3	0.91	0.59	1.00	0.85	0.93	0.92	0.82	0.71	0.60	0.67	0.90	0.79
4	0.89	0.59	0.85	1.00	0.86	0.84	0.79	0.68	0.59	0.68	0.81	0.78
5	0.89	0.59	0.93	0.86	1.00	0.90	0.82	0.74	0.61	0.68	0.88	0.79
6	0.92	0.61	0.92	0.84	0.90	1.00	0.86	0.72	0.63	0.70	0.88	0.81
7	0.85	0.61	0.82	0.79	0.82	0.86	1.00	0.73	0.66	0.67	0.81	0.83
8	0.73	0.63	0.71	0.68	0.74	0.72	0.73	1.00	0.72	0.69	0.72	0.72
9	0.64	0.74	0.60	0.59	0.61	0.63	0.66	0.72	1.00	0.73	0.58	0.66
10	0.68	0.80	0.67	0.68	0.68	0.70	0.67	0.69	0.73	1.00	0.65	0.68
11	0.87	0.58	0.90	0.81	0.88	0.88	0.81	0.72	0.58	0.65	1.00	0.77
12	0.83	0.63	0.79	0.78	0.79	0.81	0.83	0.72	0.66	0.68	0.77	1.00
Panel B: Distance matrix between stations. (*Km*)												
1	0.00	49.53	11.08	5.86	14.76	8.38	14.69	43.08	63.15	38.27	15.51	13.79
2	49.53	0.00	43.45	51.05	43.49	41.58	34.86	49.35	41.92	11.36	37.42	40.17
3	11.08	43.45	0.00	8.64	3.96	6.32	11.13	32.37	52.26	32.13	6.11	18.03
4	5.86	51.05	8.64	0.00	11.30	9.80	16.67	38.62	60.46	39.68	14.49	18.80
5	14.76	43.49	3.96	11.30	0.00	10.08	13.68	28.45	49.16	32.35	6.63	21.70
6	8.38	41.58	6.32	9.80	10.08	0.00	6.89	37.90	55.66	30.24	7.80	11.71
7	14.69	34.86	11.13	16.67	13.68	6.89	0.00	38.32	52.58	23.59	7.91	10.58
8	43.08	49.35	32.37	38.62	28.45	37.90	38.32	0.00	28.54	42.04	30.90	48.52
9	63.15	41.92	52.26	60.46	49.16	55.66	52.58	28.54	0.00	41.49	47.89	62.88
10	38.27	11.36	32.13	39.68	32.35	30.24	23.59	42.04	41.49	0.00	26.15	29.74
11	15.51	37.42	6.11	14.49	6.63	7.80	7.91	30.90	47.89	26.15	0.00	17.63
12	13.79	40.17	18.03	18.80	21.70	11.71	10.58	48.52	62.88	29.74	17.63	0.00

Table 1. Minimum correlation matrix and distance matrix between stations. The minimum correlation matrix is the minimum value of seven correlations between two stations.

more likely to be strongly and positively correlated with its neighbour's spatially reviewed recurrence intervals. In other words, if location *i*'s concentration time series are in high correlation with *j*'s, the message about recurrence intervals at location *j* is conducive to explaining and predicting the recurrence intervals at *i*.

Excitingly, Table 1 panel A shows that all the minimum correlations of the 7 selected time series between the 12 stations are over 0.50, which indicates a strong likelihood that one station's recurrence interval series over a threshold are also highly correlated with its neighbour's spatially reviewed recurrence intervals. This euphoria of high spatial correlation structure in reviewed recurrence intervals fuels our interest in extending traditional time series models to more general spatiotemporal models. The inter-station distances as tabulated in panel B are

	$q=2.0$		$q=2.5$		$q=3.0$		$q=3.5$		$q=4.0$	
	Exponential	**Weibull**	**Exponential**	**Weibull**	**Exponential**	**Weibull**	**Exponential**	**Weibull**	**Exponential**	**Weibull**
Panel A: Recurrence series of PM_{10} in Station 2										
ω_q	0.10**	0.11***	0.08	0.19***	0.31***	0.22***	0.10	0.51***	0.06*	0.04 ***
$\alpha_{q,1}$	0.08**	0.06***	0.10**	0.14***	0.26***	0.22***	0.12	0.00	0.17*	0.08
$\beta_{q,1}$	0.82***	0.75***	0.83***	0.49***	0.46***	0.35***	0.81***	0.00	0.83***	0.83 ***
k		0.57***		0.51***		0.46***		0.40***		0.38 ***
$Q_q(10)$	7.22	10.09	7.49	9.79	5.98	6.04	2.70	3.97	1.65	1.77
	(0.70)	(0.43)	(0.68)	(0.46)	(0.82)	(0.81)	(0.99)	(0.95)	(1.00)	(1.00)
Panel B: Recurrence series of AQI in Station 4										
ω_q	0.05***	0.04***	0.15**	0.16**	0.81***	0.33***	0.06	0.03***	0.60***	0.41 **
$\alpha_{q,1}$	0.05***	0.03***	0.27**	0.17*	0.40	0.06	0.10	0.03	0.00	0.01
$\beta_{q,1}$	0.90***	0.89***	0.66***	0.43*	0.00	0.00	0.88***	0.87***	0.40***	0.00
k		0.56***		0.50***		0.42***		0.37***		0.32 ***
$Q_q(10)$	11.63	13.82	5.38	10.43	11.69	13.23	4.34	3.50	4.61	4.59
	(0.31)	(0.18)	(0.86)	(0.40)	(0.31)	(0.21)	(0.93)	(0.97)	(0.92)	(0.92)
Panel C: Recurrence series of NO_2 in Station 7										
ω_q	0.09**	0.05***	0.18***	0.16**	0.23	0.12***	0.54***	0.18***	0.05***	0.64
$\alpha_{q,1}$	0.19***	0.08***	0.50***	0.44***	0.62	0.32	0.29	0.22	0.00***	0.00
$\beta_{q,1}$	0.75***	0.79***	0.50***	0.28**	0.38	0.43***	0.10***	0.15**	1.00***	0.00
k		0.52***		0.43***		0.36***		0.31***		0.26 ***
$Q_q(10)$	0.98	1.00	3.68	4.23	3.01	3.02	1.14	0.73	3.00	3.22
	(1.00)	(1.00)	(0.96)	(0.94)	(0.98)	(0.98)	(1.00)	(1.00)	(0.98)	(0.98)

Table 2. Partial estimated results of the RIA-ACD(1,1) model. We estimate the above model under five thresholds from $q=2.0$ to $q=4.0$ with 0.5 increments each time. The table reports the estimated results using both exponential and Weibull distributions. We also estimate the Newey-West corrected standard errors and label the significance of estimations using asterisks, with *, **, ***representing the statistical significance at 10%, 5%, and 1% level. The Ljung-Box Q-test for residuals autocorrelation (lags = 10) is reported in the last row of each panel and p-value of the Ljung-Box statistic is in the parentheses.

all below 100 *Km*, which has twofold implications. First, once the inter-station distance is longer than 100 *Km*, it's likely to introduce the variations of geographical and meteorological conditions to this model and reduce the explanatory and prediction powers of the neighbouring stations. Second, when observing the air conditions and the frequency of extreme pollution events, people are more apt to refer to the situation of an adjacent area instead of remote ones. To recapitulate, the aforementioned reasons are the main inspirations to define the recurrence intervals from the spatially reviewed angle.

Primary results of RIA-ACD(1,1) and RIA-SACD(1,1,1). Starting from the simplest scenario, we evaluate the RIA-ACD(1,1) and the RIA-SACD(1,1,1) models on the 84 selected time series respectively. To fully incorporate the recurrence intervals' distribution information, the results are displayed from both exponential distribution and Weibull distribution of ε. Some scholars opted for power-law distribution with an exponential cutoff and q-exponential distribution to shelter the recurrence intervals' heavy tails in stock market[4,5]. Simiu and Heckert showed that the recurrence intervals of wind data are better fitted by General Pareto distribution in the tails[33]. When checking the distributions of recurrence intervals of air pollutants, we find General Pareto distribution, with more parameters, to be more adaptive. However, involving more parameters in the distributions of standardized innovations surely means more complexity and instability of the estimation process though the results may be more accurate. To strike a balance between generality and complexity, we select the above two distributions. As noted before, in order to make the distributions meet the trend of recurrence intervals, we confine the shape parameter k in the Weibull distribution within [0, 1]. As the Weibull distribution possesses one more parameter than the exponential distribution, without loss of the precision and generality, this work reports the results for the Weibull distribution after presenting the primary results.

Tables 2 and 3 tabulate partial results of all the time series. While Table 2 mainly focuses on whether RIA-ACD model is valid in evaluating the expected recurrence intervals[16], Table 3 probes whether incorporating spatial information in the above models improves the predictive power. In Table 2, the two distributions produce similar results for the estimated parameters ω_q, $\alpha_{q,1}$, $\beta_{q,1}$ and Q(10). Although some estimations may vary in significance, the estimated magnitudes of results are quite close under both exponential distribution and Weibull distribution. The diagnostic statistics $Q(10)$ s are never significant in the first 10 lags for the normalized innovations, which indicates that the RIA-ACD model is successful in capturing the persistent features of the recurrence intervals in the time dimension[34]. Both exponential and Weibull distribution are successful in identifying the recurrence intervals' probability distributions. In Table 2, as the value of q varies from 2.0 to 4.0, $\beta_{q,1}$ is more likely to stay significant than $\alpha_{q,1}$, indicating that the conditional expected recurrence intervals' autocorrelation structure is more stable. In other words, in predicting the recurrence intervals of more severe atmospheric pollution, the

	$q = 2.0$		$q = 2.5$		$q = 3.0$		$q = 3.5$		$q = 4.0$	
	Exponential	**Weibull**	**Exponential**	**Weibull**	**Exponential**	**Weibull**	**Exponential**	**Weibull**	**Exponential**	**Weibull**
Panel A: Recurrence series of PM_{10} in Station 2										
$\omega_q^{(i)}$	0.04	0.12***	0.53***	0.34***	0.06**	0.22***	0.06***	0.03**	0.59***	0.27***
$\alpha_{q,1}^{(i)}$	0.04***	0.06***	0.26***	0.21***	0.15*	0.23***	0.04	0.01	0.02	0.02
$\beta_{q,1}^{(i)}$	0.89***	0.69***	0.12	0.11	0.82***	0.34***	0.80***	0.85***	0.00	0.00
$\gamma_{q,1,1}^{(i)}$	0.02	0.04***	0.13***	0.08***	0.02	0.00	0.14***	0.06***	0.42**	0.53
k		0.58***		0.51***		0.46***		0.41***		0.39 ***
$Q_q^{(i)}(10)$	5.01	5.83	11.90	11.31	2.46	1.20	16.64	18.82	0.43	0.71
	(0.89)	(0.83)	(0.29)	(0.33)	(0.99)	(1.00)	(0.08)	(0.04)	(1.00)	(1.00)
Panel B: Recurrence series of AQI in Station 4										
$\omega_q^{(i)}$	0.87***	0.04***	0.51***	0.27***	0.70***	0.30***	0.08	0.36***	0.02**	0.41**
$\alpha_{q,1}^{(i)}$	0.08***	0.03***	0.62***	0.27**	0.55	0.09	0.05	0.03	0.00	0.01
$\beta_{q,1}^{(i)}$	0.00	0.88***	0.05	0.04	0.00	0.00	0.88***	0.00	0.94***	0.00
$\gamma_{q,1,1}^{(i)}$	0.09	0.00	0.17***	0.10***	0.10**	0.04	0.01	0.01	0.06	0.01
k		0.56***		0.50***		0.42***		0.36***		0.32***
$Q_q^{(i)}(10)$	55.92	11.91	15.18	15.26	8.52	8.26	2.57	4.13	1.62	3.55
	(0.00)	(0.29)	(0.13)	(0.12)	(0.58)	(0.60)	(0.99)	(0.94)	(1.00)	(0.97)
Panel C: Recurrence series of NO_2 in Station 7										
$\omega_q^{(i)}$	0.12*	0.07**	0.51***	0.15***	0.19	0.11***	0.41	0.09	0.02	0.29***
$\alpha_{q,1}^{(i)}$	0.17***	0.07***	0.12***	0.40***	0.65**	0.38	0.28	0.19	0.00	0.00
$\beta_{q,1}^{(i)}$	0.70***	0.75***	0.00	0.32***	0.35*	0.33***	0.29***	0.41*	1.00	0.00
$\gamma_{q,1,1}^{(i)}$	0.02	0.01	0.59***	0.00	0.15	0.12	0.05	0.03	0.02	0.29***
k		0.52***		0.43***		0.37***		0.31***		0.27***
$Q_q^{(i)}(10)$	0.97	0.95	2.90	4.08	2.80	2.77	2.25	0.53	2.99	3.34
	(1.00)	(1.00)	(0.98)	(0.94)	(0.99)	(0.99)	(0.99)	(1.00)	(0.98)	(0.97)

Table 3. Partial estimated results of the RIA-SACD(1,1,1) model under the Weibull distribution. We estimate the above model under five thresholds from $q = 2.0$ to $q = 4.0$ with 0.5 increments each time. The table reports the estimated results using both exponential and Weibull distributions. We also estimate the Newey-West corrected standard errors and label the significance of estimations using asterisks, with *, **, ***representing the statistical significance at 10%, 5%, and 1% level. The Ljung-Box Q-test for residuals autocorrelation (lags = 10) is reported in the last row of each panel and *p*-value of the Ljung-Box statistic is in the parentheses.

impact from last conditional expected interval $\psi_{q,s-1}$ is more helpful than last realized interval $d_{q,s-1}$. Another notable finding is that as q increases, the shape parameter k of Weibull distribution decreases, the implication of which is quite straightforward: The most extreme polluted days are rare and moderately polluted days are found within longer periods and the length of recurrence interval series is thus shortened, making the distribution flatter. In this sense, the declining k is quite consistent with the intuition[35]. But caution should be used to interpret the empirical meanings of the coefficients $\alpha_{q,1}$ and $\beta_{q,1}$. As noted before, $\alpha_{q,1}$ mainly measures how significantly last realized recurrence interval influences this expected recurrence interval, and $\beta_{q,1}$ evaluates how significantly last conditional expected recurrence interval influences this expected recurrence interval. For example, when $q = 3$, with the recurrence interval series of PM_{10} in station 2 under the exponential distribution, $\alpha_{q,1} = 0.26$, which means that once last recurrence interval increases by 1 hour, it's more likely that the expected interval increases by 0.26 hours on average. $\beta_{q,1} = 0.46$ shows that once last conditional expected interval increases by 1 hour, the expected interval increases by 0.46 hours on average. Under the stationary conditions, the future expected recurrence interval is $\omega_q/(1-\alpha_{q,1}-\beta_{q,1}) = 1.11\langle d_q \rangle$.

To fully present the RIA-ACD(1,1) results for all the time series, we present $\alpha_{q,1}$ s and $\beta_{q,1}$ s in Fig. 6. The above findings from Table 2 are consistent with the trend shown in Fig. 6. Generally, when $q = 2$, most estimations are significantly greater than 0 whereas the percentage of this significance slumps when $q = 4$, which is either an indirect evidence that the recurrence interval analysis with the RIA-ACD model somewhat fails to capture the very extreme air pollution occurrences or an indicator that the recurrence intervals of very extreme air pollution occurrence seem untraceable. Technically speaking, the paucity of the recurrence interval series at very extreme level makes it difficult to evaluate and predict. Another notable finding is that $\beta_{q,1}$ outweighs $\alpha_{q,1}$ on the whole. As mentioned above, this immediately shows that by integrating the past temporal information, the autoregressive property is efficiently archived in the conditional expected recurrence interval ψ. Finally, when we focus on the AQI series specifically, it's easy to find that as the threshold improves to $q = 4$, only station 4, 9 and 11's α s are still significant and the β values of stations 1, 2, 6, 8 and 12 are insignificant. For the same City of Beijing, the discordancy in the temporal-spatial correlation structure logged by different monitoring stations provides a

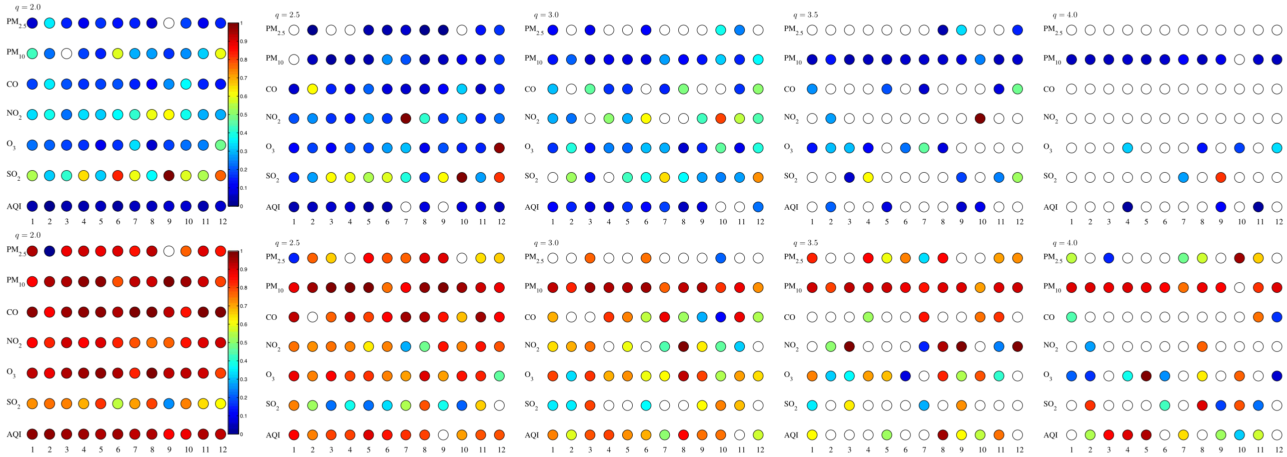

Figure 6. *The coefficients* $\alpha_{q,1}$ and $\beta_{q,1}$ of each time series are estimated from the RIA-ACD(1,1) model under Weibull distribution. The magnitudes of the parameters are represented by different colors filled in each circle, and the circle with no color filled in signifies that the estimated parameter is not significant at 95% level. The top row is the $\alpha_{q,1}$ and the bottom row is the $\beta_{q,1}$. Presented from the left column to the right are $q = 2.0$ to $q = 4.0$ respectively. The standard error is adjusted using Newey-West method.

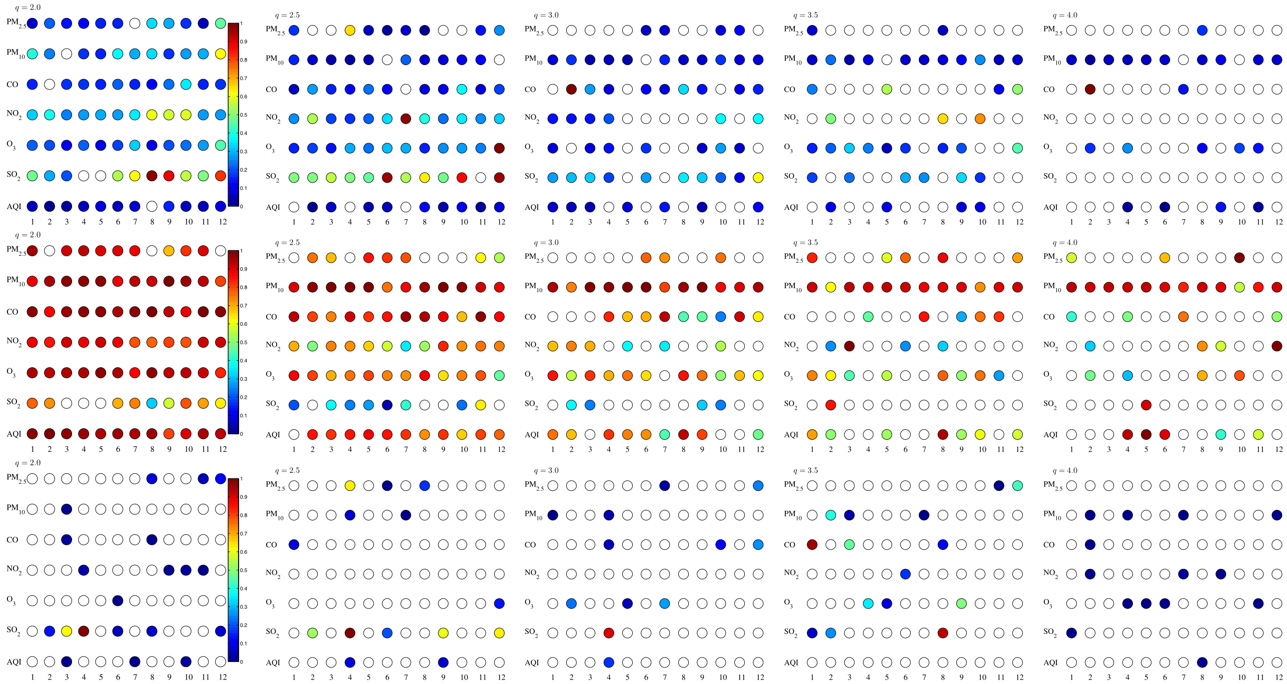

Figure 7. $\alpha_{q,1}^{(i)}$, $\beta_{q,1}^{(i)}$ and $\gamma_{q,1,1}^{(i)}$ of each time series estimated from the RIA-SACD(1,1,1) model. The magnitudes of the parameters are represented by different colors filled in each circle, and the circle with no color filled in signifies that the estimated parameter is not significant at 95% level. The top row is the $\alpha_{q,1}^{(i)}$, the middle row is the $\beta_{q,1}^{(i)}$ and the bottom row is $\gamma_{q,1,1}^{(i)}$. Presented from the left column to the right are $q = 2.0$ to $q = 4.0$ respectively. The standard error is adjusted using Newey-West method.

conformational evidence that the occurrence of extreme air pollution in the same place may be driven by distinct mechanisms in different small areas[36].

Table 3 incorporates the spatial term into the RIA-ACD model on the basis of Table 2. As mentioned before, the chief reason to adopt this model is to check whether integrating spatial information is helpful in evaluating the extreme air pollution periods. Compared with the results in Table 2, the values of α and β are generally close to the model without spatial term due to the orthogonal relationship between temporal dimension and spatial dimension. But when it comes to the γ's in Fig. 7, it's quite obvious that the spatial correlation only exists in 20% of all the selected time series. Moreover, this spatial significance varies with the threshold q while the overall percentage of significant spatial terms is close. The difficulty in capturing high extreme pollution level from

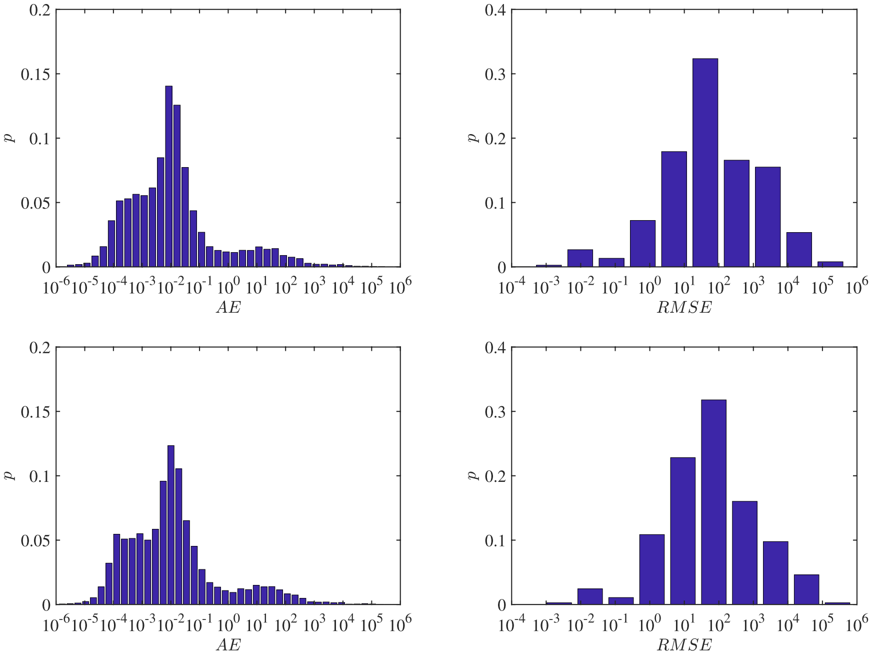

Figure 8. Distributions of *AE* and *RMSE* from the RIA-ACD(1,1) model and the RIA-SACD(1,1,1) model. The top row is the RIA-ACD(1,1) model and the bottom row is the RIA-SACD(1,1,1) model.

	$p\,(AE < 1)$	$p\,(AE < 6)$	$p\,(AE < 12)$	$p\,(AE < 24)$	$p\,(AE < 48)$	$p\,(AE < 96)$	$Avg\,(AE)$
ACD (1,1)	0.84	0.88	0.89	0.90	0.91	0.93	137
SACD (1,1,1)	0.83	0.88	0.89	0.90	0.91	0.93	123
	$p\,(RMSE < 1)$	$p\,(RMSE < 6)$	$p\,(RMSE < 12)$	$p\,(RMSE < 24)$	$p\,(RMSE < 48)$	$p\,(RMSE < 96)$	$Avg\,(RMSE)$
ACD (1,1)	0.04	0.13	0.18	0.25	0.31	0.44	717
SACD (1,1,1)	0.04	0.13	0.17	0.23	0.30	0.43	642

Table 4. p-values of some AE and RMSE breakpoints and average AEs and RMSEs.

temporal side has been construed before and the same is also true with *q* from the spatial side. Anyway, though the RIA-SACD model has merely made limited improvements, it still can explain part of the spatial correlations.

Out-of-sample test.

In this section, we conduct a sliding window scheme of out-of-sample test to explore whether the RIA-ACD(1,1,1) model is valid in predicting the recurrence intervals and to what extent the predictive power can be improved using the RIA-SACD(1,1,1) model.

The expected next recurrence duration at threshold *q* using the RIA-SACD(1,1,1) model is given by

$$E\left[d_{q,s+1}^{(i)}|I_s\right] = \hat{\psi}_{q,s+1}^{(i)} = \hat{w}_q^{(i)} + \hat{\alpha}_{q,1}\,\psi_{q,s}^{(i)} + \hat{\beta}_{q,1}\,d_{q,s}^{(i)} + \hat{\gamma}_{q,j,1}^{(i)}\,d_{q,s}^{(ij)} \tag{3}$$

For each recurrence interval series $d_{q,s}^{(i)}$ with $s = 1, 2, \ldots, N$, we choose $s = 1, 2, \ldots, N - 30$ as the in-sample estimation, and recursively estimate the next recurrence duration $d_{q,s+1}^{(i)}$. By comparing $\hat{\psi}_{q,s+1}^{(i)}$ and $d_{q,s+1}^{(i)}$, we report the absolute error (AE)

$$AE_{q,s+1}^{(i)} = \left|\hat{\psi}_{q,s+1}^{(i)} - d_{q,s+1}^{(i)}\right| \tag{4}$$

and the root-mean-square error (RMSE)

$$RMSE_q^{(i)} = \sqrt{\frac{1}{30}\sum_{s=1}^{30}\left(d_{q,s+1}^{(i)}\hat{\psi}_{q,s+1}^{(i)}\right)^2}. \tag{5}$$

Figure 8 displays the distributions of *AE* and *RMSE* of the out-of-sample test using both the RIA-ACD(1,1) model and the RIA-SACD(1,1,1) model. Together with Table 4, it shows that more than 90% of the predictions deviate from the true recurrence intervals within 3 days under both models, which demonstrates the good predictive power of these models. In terms of the *p* values of *AE* and *RMSE*, incorporating the spatial term does no better than the model without spatial term. However, the average absolute error and root-mean-squared error

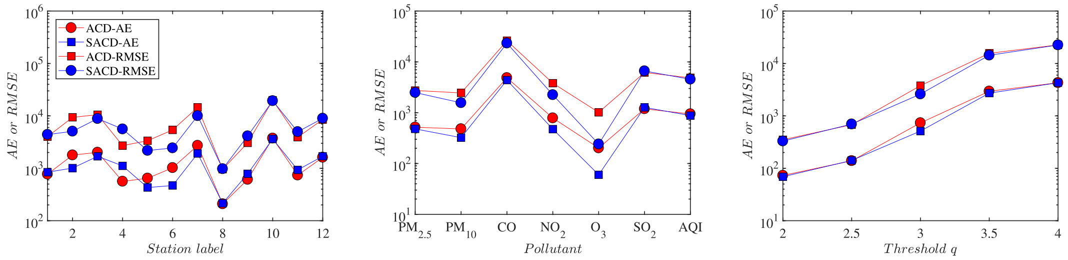

Figure 9. Average *AE* or *RMSE* from the ACD(1,1) model and the SACD(1,1,1) model from three dimensions.

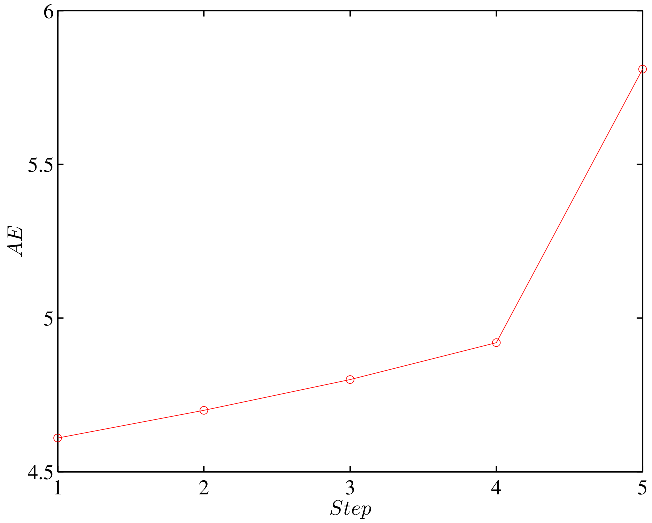

Figure 10. Absolute errors of multi-step out-of-sample test of $PM_{2.5}$ series and station one under the threshold $q = 2.0$. The results are obtained from RIA-SACD model.

decrease by 10% if the spatial term is included in the ACD structure. In this sense, the spatial autoregressive conditional model generally performs slightly better than the model considering no spatial interaction. Figure 9 shows the average *AE* or *RMSE* from the ACD(1,1) model and the SACD(1,1,1) model from three dimensions.

To further consolidate the predictive power of the integrated model, our research goes from the one-step forward to multi-step forward. However, as illustrated in Eq. (3) and noted above, a moving window scheme is employed when the out-of-sample prediction is examined. Here, we only present results from the recurrence intervals of $PM_{2.5}$ in station one and under the threshold $q = 2.0$ in Fig. 10 due to limitation of computational consumption. It's quite straightforward that as we go further steps, the average absolute errors increase, which means that the predictive power will decrease when the most recent information is used to predict the further extreme air pollution event. However, this declining trend of the predictive power is still within our expectation. Specifically, the absolute error that forecasts the most recent approaching extreme air pollution is 30% lower than that forecasts the 5th approaching extreme air pollution from now on. In fact, due to the presence of persistence, one can easily get highly reliable one-step forecasts using a simple persistence model. On the other hand, when making prediction, we intuitively make full use of the past information and make predictions on the nearest future event. Practically speaking, one-step forward is still of great significance in the air pollution prediction.

More general settings. The above preliminary results have shown that both time dimension and spatial dimension have the ability to capture the structures of recurrence interval series. But the predictability from the temporal dimension is much stronger than the spatial dimension and the predictive power varies with q. In this section, the model has been extended to more time lags and more neighbouring stations in order to check the temporal and spatial persistence as well as how the explanatory power is debilitated with lags and distances. On the temporal side, finding out the temporal persistence is vital to identify how extreme air pollution events happen and build up the knowledge of statistics about air pollution. On the spatial side, detecting the relationship between explanatory power and inter-station distances is salutary to understand the spatial interactions of air pollutants.

Figure 11(a–d) displays significant βs in each lag. These percentages, in other words, measure the Markovian properties of the recurrence intervals in question. Generally speaking, about half of the series show strong autocorrelation in the first lag, and the percentage goes down to only 20% when it comes to the sixth lag. For one thing, the declining persistence accounts for short-term correlations of the recurrence intervals, which supports the exponential distributions of the recurrence intervals. For another, it shows the specifications included in the

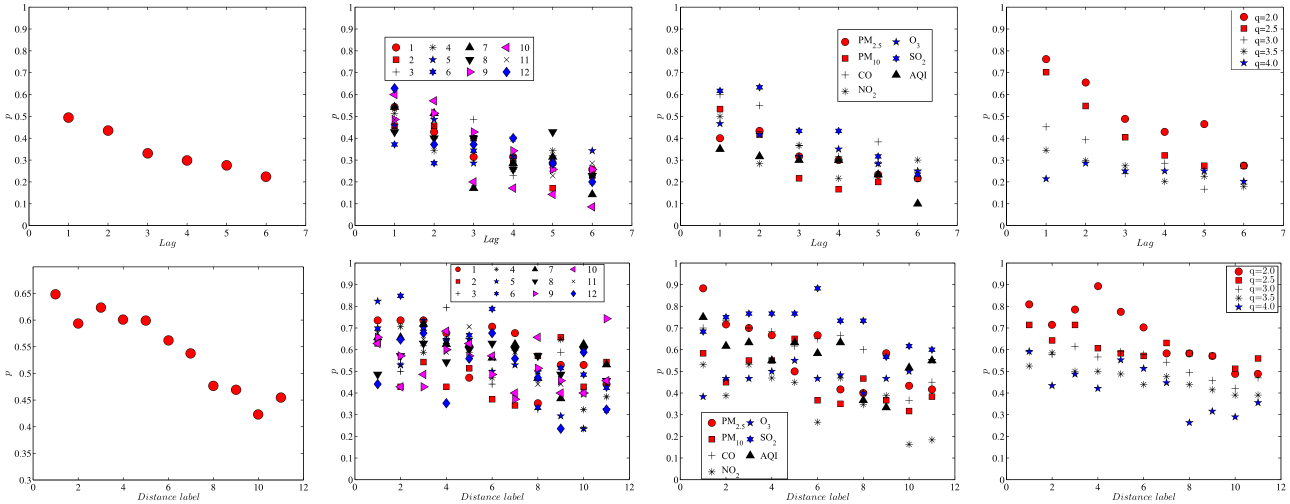

Figure 11. Percentages of significant $\beta_{q,v}^{(i)}$ s across lags and significant $\gamma_{q,j,1}^{(i)}$ s across distances based on Eq. (6) and Eq. (7). We first estimate each time series using the above extended models and then calculate the percentage of $\beta_{q,u}^{(i)}$ s and $\gamma_{q,j,1}^{(i)}$ s above the 95% significant level in each dimension.

first several lags are sufficient in the estimation. When classified into three dimensions, the trend in each dimension is consistent with the overall trend in Fig. 11(a). However, some outliers such as those for SO$_2$ with $q = 4.0$ at station 5 are found to deviate from this trend. In these outliers, more significant β's are observed in the second lag instead of the first lag,

$$\psi_{q,s}^{(i)} = \omega_q^{(i)} + \sum_{u=1}^{L_d} \alpha_{q,u}^{(i)} d_{q,s-u}^{(i)} + \sum_{v=1}^{L_\psi} \beta_{q,v}^{(i)} \psi_{q,s-v}^{(i)}, \tag{6}$$

which suggests that we cannot exclude the possibility that some recurrence interval series possess strong long-term correlations. The factor immediately relevant to the explanatory power is the distance between stations that may be mutually beneficial in explaining and predicting. In the above nearest neighbours' results, it has been found that the nearest neighbour's recurrence intervals have predictive power in the station's recurrence intervals. Here in a more general setting, we go a step further to accommodate all the other stations in the spatial terms of Eq. (7) in recognizing the spatial effects:

$$\psi_{q,s}^{(i)} = \omega_q^{(i)} + \alpha_{q,1}^{(i)} d_{q,s-1}^{(i)} + \beta_{q,1}^{(i)} \psi_{q,s-1}^{(i)} + \sum_{j=1}^{11} \gamma_{q,j,1}^{(ij)} d_{q,s-1}^{(ij)}. \tag{7}$$

Before performing the following calculation, we first sort out i's spatial neighbours in an ascending order, with $j = 1$ as the nearest station away from i while $j = 11$ as the farthest.

After testing all the recorded time series, we display the percentages of $\gamma_{q,j,1}^{(i)}$ s above the 5% significance level in Fig. 11. Generally speaking, it's cogent that the nearest station possesses the strongest power to explain the recurrence intervals, which has a decreasing trend along with the increasing distance. According to Fig. 11(a), of all the series, around 65% of the nearest neighbours' spatially reviewed recurrence intervals show a strong relationship with the conditional recurrence intervals and this percentage dwindles to about 40% when it comes to the farthest station. Table 1 reveals that even with the farthest station, they still maintain a high correlation in the original concentration series, which supports 40% of significant spatial interactions. In Fig. 11(b–d), the percentages are presented specifically from three dimensions. Although the overall trend in each dimension is congruent with that in (a), there are still some outliers that deviate from the track. For instance, in the station dimension the most powerful neighbouring stations are mainly the two closest stations, and station 9′s most powerful neighbour is the farthest one instead of the nearest one. In the pollutant dimension, the most powerful station to explain the SO$_2$'s trend seems to be neither the nearest nor the farthest one. In the threshold dimension, when $q = 2$, the most powerful station to explain the recurrence intervals may be the fourth closest station. In spite of the data noise and model misspecification, these outliers are partly ascribed to the unexplained spatial interactions. In our setting, one primary assumption is that the strength of spatial interaction is inversely related to the inter-station distance, which is true of the sample population but not necessarily specific individuals. Notwithstanding, it's quite instrumental for us to get clues on heterogeneous spatial interactions from the above discussions, which in turn profits the choice of a proper j in Eq. (12), not simply the nearest neighbour.

Label	Code	Lat.	Long.	Label	Code	Lat.	Long.
1	10001A	116.37	39.87	7	10007A	116.32	39.99
2	10002A	116.17	40.29	8	10008A	116.72	40.14
3	10003A	116.43	39.95	9	10009A	116.64	40.39
4	10004A	116.43	39.87	10	10010A	116.23	40.20
5	10005A	116.47	39.97	11	10011A	116.41	40.00
6	10006A	116.36	39.94	12	10012A	116.22	39.93

Table 5. List of national air pollutants monitoring stations in Beijing.

Discussion

Motivated by the methods that model irregular spaced transaction data in the stock market[22,34], we apply the autoregressive conditional duration model into the recurrence interval analysis of extreme air pollution events. Meanwhile, the spatial interaction and similarity between stations are taken into account and the spatial reviewed recurrence intervals are proposed to check whether the recurrence intervals of neighbouring stations are of high correlations. A simple simulation shows that when the original time series are of high correlations, the generated recurrence interval series are more likely to maintain high correlations, whereby we attempt to add the spatial reviewed term into the autoregressive conditional duration model to fully incorporate the spatial effects. Therefore, this paper copes with two crucial issues. One is to make an attempt to integrate two models which are commonly used in risk assessment and the other is to explore to what extent the spatial interaction can be utilized to predict the value at risk. This paper embarks on the exploration of these two issues from the simplest setting: the RIA-ACD(1,1) model and the RIA-SACD(1,1,1) model. Both the partial results and the general framework show that RIA-ACD is valid in capturing the characteristics of recurrence intervals of different pollutants and the AQI series. However, the predictive power of the proposed models varies in different threshold *q*s. We conclude that as the threshold rises, which means fewer extreme cases, the memory property of the recurrence intervals would become hard to measure due to the lack of enough data. The RIA-ACD and RIA-SACD model will lose some power in capturing the properties. From the spatial side, adding spatial information only increases some recurrence intervals. No matter how the threshold varies, the percentage of the series that have spatial connections is around 20%, which means the use of spatial information to predict the recurrence interval is not pervasively efficient in all pollutants at all stations.

By extending the model to more time lags and farther stations, we find a downward trend of the coefficients' significance. This temporal and spatial contracting of recurrence intervals is quite consistent with intuition and can partly be explained by some empirical results[37]. This declining significance is also displayed from the station level, the threshold level and the pollutant level. Despite the general trend, some outliers do exist and unveil the spatial transmission and interaction of different pollutants[26], probably caused by some meteorological conditions[38].

Although this method provides a statistical direction to identify extreme air pollution occurrence intervals and the results are carefully presented, there are several challenges that require further efforts. The first one is how to identify the best distributions to fit the model. In this paper, we choose exponential and Weibull distribution as a simple case. Although other distributions such as generalized Parato distribution probably provide better fits[33], the critical issue is how to build in a more adaptive distribution while minimizing the likelihood function. The second challenge as shown in the general setting part is that we don't integrate temporal lags and distant stations in the model at the same time due to the difficulty of estimation: once more items are involved in this model, the maximization of the likelihood is not convergent any longer, which is also true of the number of parameters. Thirdly, this paper presents the spatial information of recurrence intervals by proposing the spatially reviewed recurrence intervals, which offers a direct means to generate two recurrence interval series of equal length. To completely present the spatial interaction, a more accurate and informative method should be developed. Be that as it may, this paper still provides a promising direction to explore and predict the occurrence of extreme air pollution. In addition, a more general framework that aggregates all the displaced stations and all the effective time lags would be of great significance in improving the predictive power for extreme air pollution events. Since more stations and time lags mean more complexity which might add some fixed or random effects of the model and reduce the predictive power, a Bayesian updating scheme similar to looking one step back with spatial aggregation over multiple stations might help[15].

Methods

Data sets.

The hourly pollutant data are collected from Shanghai Qingyue Open Environmental Protection Data Center (QOEPDC). We use Beijing as our research object for several reasons. Firstly, the air quality in Beijing, the capital of People's Republic of China, has always been a focal point of public attention and criticism in recent years. Secondly, air conditions there are more volatile than those in other cities due to a bunch of political reasons. The 12 national air pollutants monitoring stations in Beijing are listed in Table 5, and the geographic distributions of these stations can be found in Fig. 12. We choose the hourly records of $PM_{2.5}$, PM_{10}, CO, NO_2, O_3, SO_2 as well as the AQI series in each station as our research samples, spanning from 1 January 2013 to 31 December 2016.

In total, $12 \times 7 = 84$ time series (AQI included) are collected. The original concentration time series is denoted as $c(t(d, h))$, where d is the d-th day and h is the h-th hour of a particular day. In view of the influence of diurnal effect, we remove the intraday effect by dividing the average level at the same hour $h(h = 1, 2, \ldots, 24)$ in each day

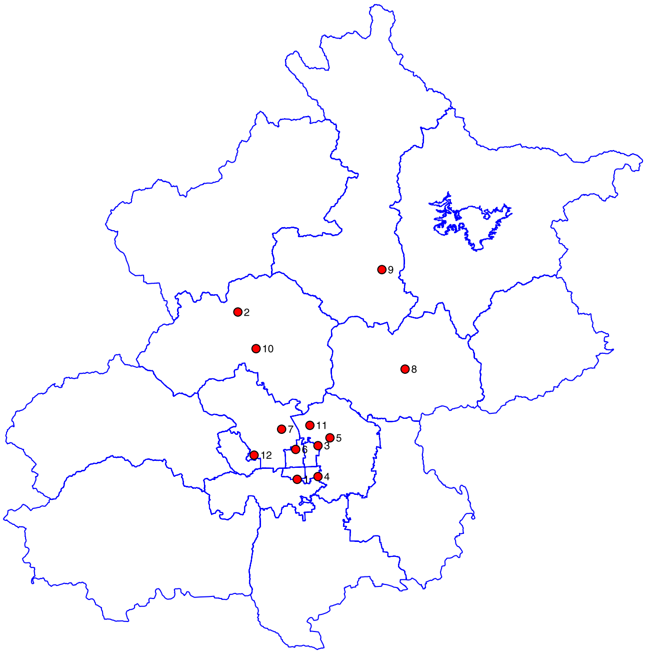

Figure 12. Distributions of the twelve national observation stations in Beijing.

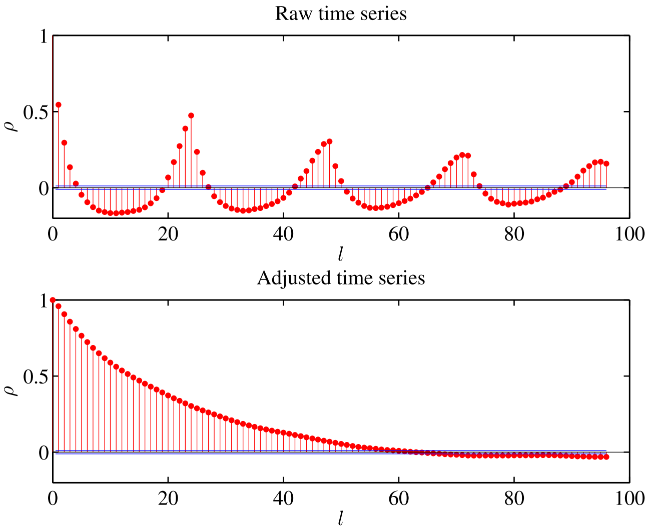

Figure 13. Autocorrelations of two $PM_{2.5}$ time series in station one. (**a**) is the c and (**b**) is x.

$$x(t(d,\ h))=\frac{c(t(d,\ h))}{\frac{1}{N_d}\sum_{d=1}^{N_d}c(t(d,\ h))},\tag{8}$$

where N_d is the number of days. Through this method, most intraday patterns are removed, and as shown in Fig. 13, the cyclic autocorrelation pattern is transformed into a gradual downward trend. Moreover, this method makes the mean of the series generally close to 1.

RIA-ACD model and RIA-SACD model. In this section, we first introduce the autoregressive conditional durations (ACD) model which is widely used in analyzing the irregular event durations and then extend the ACD model to a spatiotemporal autoregressive conditional (SACD) model by adding spatial terms in the conditional duration structure.

The ACD model was first proposed from the generalized autoregressive conditional heteroskedasticity (GARCH) model[34]. In the ACD model, the sth recurrence interval or duration $d_{q,s}$ is postulated to follow

$$d_{q,s} = \psi_{q,s}\varepsilon_{q,s}, \tag{9}$$

where $\{\varepsilon_{q,s}\}$ is a sequence of independent and identically distributed random variables and $\psi_{q,s}$ is the expected recurrence interval when all the information set $I_{q,s-1}$ is known. Similar to the GARCH model, the expected recurrence interval is modelled as an ARMA process[34],

$$\psi_{q,s} = \omega_q + \sum_{u=1}^{L_d} \alpha_{q,u} d_{q,s-u} + \sum_{v=1}^{L_\psi} \beta_{q,v}\psi_{q,s-v}. \tag{10}$$

Given that all recurrence intervals are positive, all the coefficients in Eq. (10) must be positive. The logarithmic ACD model has been proposed to avoid this restriction[39–41]. Hautsch also proposed the Box-Cox transformation of Eq. (10) to make the model more adaptive[42]. Since $E[\varepsilon_{q,s}] = 1$ such that $E[d_{q,s}|I_{q,s-1}] = E[\psi_{q,s}\varepsilon_{q,s}|I_{q,s-1}] = \psi_{q,s}$, $\{\varepsilon_{q,s}\}$ has a positive support and unity expected value. Different distribution assumptions of $\{\varepsilon_{q,s}\}$ result in different ACD models[34].

Based on the ACD model, we propose the spatial ACD model by adding spatial terms in Eq. (10) as follows

$$d_{q,s}^{(i)} = \psi_{q,s}^{(i)}\varepsilon_{q,s}^{(i)} \tag{11}$$

with

$$\psi_{q,s}^{(i)} = \omega_q^{(i)} + \sum_{u=1}^{L_d} \alpha_{q,u}^{(i)} d_{q,s-u}^{(i)} + \sum_{v=1}^{L_\psi} \beta_{q,v}^{(i)}\psi_{q,s-v}^{(i)} + \sum_{k=1}^{L_j}\sum_{j=1}^{N(i)} \gamma_{q,j,k}^{(i)} w_{ij} d_{q,s-k}^{(ij)}, \tag{12}$$

where the added terms include spatially reviewed recurrence intervals $d_{q,s-k}^{(ij)}$ with k lags and w_{ij} is the adjacent matrix or deterrence function which is inversely related to the distance[43].

For simplicity's sake, we start with the nearest neighbour of location i, $NN(i)$, with one lag behind. So Eq. (12) reduces to

$$\psi_{q,s}^{(i)} = \omega_q^{(i)} + \alpha_{q,1}^{(i)} d_{q,s-1}^{(i)} + \beta_{q,1}^{(i)}\psi_{q,s-1}^{(i)} + \gamma_{q,1,1}^{(i)} d_{q,s-1}^{(NN(i))} \tag{13}$$

In this setting, the adjacent matrix $w_{ij} = 1$ if j is the nearest neighbour of i and $w_{ij} = 0$ otherwise. In this paper, the classic ACD model is integrated with the recurrence interval analysis, referred to as the RIA-ACD model to highlight different thresholds. Moreover, Eq. (13) is the case we mainly focus on since this specification is numerically feasible to solve and the properties can be easy to extend. This simplified spatial autoregressive conditional duration model is denoted as the RIA-SACD(1,1,1) model and the model without the spatial terms is hence categorized as the RIA-ACD(1,1) model. The RIA-SACD(1,1,1) model will nest the RIA-ACD(1,1) model if no spatial interaction occurs. However, if two stations are not sufficiently close, the story will be very different. The duration series of the two stations will be "correlated". The specification also considers the mutual impact on the extreme event occurrence intervals between two locations.

From Eq. (13), we have

$$\begin{cases} \psi_{q,s}^{(i)} &= \omega_q^{(i)} + \alpha_{q,1}^{(i)} d_{q,s-1}^{(i)} + \beta_{q,1}^{(i)}\psi_{q,s-1}^{(i)} + \gamma_{q,1,1}^{(i)} d_{q,s-1}^{(ij)} \\ \psi_{q,s}^{(j)} &= \omega_q^{(j)} + \alpha_{q,1}^{(j)} d_{q,s-1}^{(j)} + \beta_{q,1}^{(j)}\psi_{q,s-1}^{(j)} + \gamma_{q,1,1}^{(j)} d_{q,s-1}^{(ji)} \end{cases} \tag{14}$$

Moreover, $E[d_{q,s}^{(i)}] = \psi_{q,s}^{(i)} = E[\psi_{q,s}^{(i)}]$ and $E[d_{q,s}^{(j)}] = \psi_{q,s}^{(j)} = E[\psi_{q,s}^{(j)}]$. We obtain that, if two observations i and j are mutually nearest neighbours (they are close to each other and far away from other stations) and both recurrence interval series are under weak stationary conditions, the expected recurrence interval for station i is

$$E\left[d_q^{(i)}|(j)\right] = \frac{\omega_q^{(j)}\gamma_{q,1,1}^{(i)} + \omega_q^{(i)}\left(1 - \alpha_{q,1}^{(j)} - \beta_{q,1}^{(j)}\right)}{\left(1 - \alpha_{q,1}^{(i)} - \beta_{q,1}^{(i)}\right)\left(1 - \alpha_{q,1}^{(j)} - \beta_{q,1}^{(j)}\right) - \gamma_{q,1,1}^{(i)}\gamma_{q,1,1}^{(j)}} \tag{15}$$

It follows that, if $\gamma_{q,1,1}^{(i)} = 0$,

$$E\left[d_q^{(i)}|(j)\right] = \frac{\omega_q^{(i)}}{1 - \alpha_{q,1}^{(i)} - \beta_{q,1}^{(i)}} \triangleq \mu_i, \tag{16}$$

and if $\gamma_{q,1,1}^{(j)} = 0$,

$$E\left[d_q^{(i)}|(j)\right] = \frac{\omega_q^{(i)}}{1 - \alpha_{q,1}^{(i)} - \beta_{q,1}^{(i)}} + \frac{\omega_q^{(j)}}{1 - \alpha_{q,1}^{(j)} - \beta_{q,1}^{(j)}}\frac{\gamma_{q,1,1}^{(i)}}{1 - \alpha_{q,1}^{(i)} - \beta_{q,1}^{(i)}} = \mu_i + \mu_j\frac{\gamma_{q,1,1}^{(i)}}{1 - \alpha_{q,1}^{(i)} - \beta_{q,1}^{(i)}}. \tag{17}$$

In the first case, location j's recurrence intervals of extreme events have no impact on i's conditional intervals. Under weak stationary conditions, the expected intervals will reduce to classic ACD expectations. However, in the second case, i doesn't influence j, but the other way around (in fact, it can be assumed that directed wind from j to i always exists). Hence, the first term is a non-spatial term and the second is the spatial interaction triggered intervals.

The next model specification is the distribution of the innovation sequence $\{\varepsilon_{q,s}\}$. With a decreasing trend of the recurrence intervals' histogram, we resort to the exponential distribution

$$f(\varepsilon) = e^{-\varepsilon} \tag{18}$$

and the Weibull distribution

$$f(\varepsilon) = \frac{k}{\lambda}\left(\frac{\varepsilon}{\lambda}\right)^{k-1} e^{-(\varepsilon/\lambda)^k}. \tag{19}$$

In other words, small intervals are more likely than large intervals whatever the threshold q is. It should be pointed out that in order to make the Weibull distribution take on this trend, we restrict Weibull's shape parameter k in (0, 1).

Since $E[\varepsilon_{q,s}^{(i)}] = 1$, we obtain that

$$\lambda = \frac{1}{\Gamma(1 + 1/k)}. \tag{20}$$

Therefore, the Weibull distribution becomes

$$f(\varepsilon) = k\Gamma(1 + 1/k)[\varepsilon\Gamma(1 + 1/k)]^{k-1}\exp[-(\varepsilon\Gamma(1 + 1/k))^k]. \tag{21}$$

Maximum likelihood estimation. With the simplest SACD model as an illustrative example, suppose $D_q^{(i)} = \left\{d_{q,1}^{(i)}, d_{q,2}^{(i)}, \cdots, d_{q,N}^{(i)}\right\}$ are the realizations of an SACD model. For the parameter set $\theta_q^{(i)} = \left(\omega_q^{(i)}, \alpha_{q,1}^{(i)}, \beta_{q,1}^{(i)}, \gamma_{q,1,1}^{(i)}\right)$, the likelihood function of these recurrence interval realizations is

$$f\left(D_q^{(i)}|\theta_q^{(i)}\right) = f\left(d_{q,1}^{(i)}|\theta_q^{(i)}\right)\prod_{s=2}^{N} f\left(d_{q,s}^{(i)}|d_{q,s-1}^{(i)}, \theta_q^{(i)}\right). \tag{22}$$

For the exponential distribution, we have

$$f\left(d_{q,s}^{(i)}|d_{q,s-1}^{(i)}, \theta_q^{(i)}\right) = \frac{1}{\psi_{q,s}^{(i)}}\exp\left[-\frac{d_{q,s}^{(i)}}{\psi_{q,s}^{(i)}}\right]. \tag{23}$$

For the Weibull distribution, similarly, we have

$$f\left(d_{q,s}^{(i)}\middle| d_{q,s-1}^{(i)},\ \theta_q^{(i)}\right) = \frac{1}{\psi_{q,s}^{(i)}}k\Gamma(1 + 1/k)\left[-\frac{d_{q,s-1}^{(i)}}{\psi_{q,s}^{(i)}}\Gamma(1 + 1/k)\right]^{k-1}\exp\left[-\left(-\frac{d_{q,s-1}^{(i)}}{\psi_{q,s}^{(i)}}\Gamma(1 + 1/k)\right)^k\right] \tag{24}$$

The conditional log likelihood function of the data becomes

$$L\left(\theta_q^{(i)}|D_q^{(i)}\right) = -\sum_{s=2}^{N}\log\left[f\left(d_{q,s}^{(i)}|d_{q,s-1}^{(i)},\ \theta_q^{(i)}\right)\right]. \tag{25}$$

The estimate of $\hat{\theta}_q^{(i)}$ is obtained by maximizing $L\left(\theta_q^{(i)}|D_q^{(i)}\right)$ analytically or numerically.

The standard errors of the estimates are obtained through the Fisher information matrix. In view of the effect of serial correlation and heteroskadacity, we adopt the Newey-West robust standard errors by correcting the residuals with the weight of $1 - s/N$[44].

Model diagnosis. Similar to the classic ACD model, we use the Ljung-Box Q as a test statistic for the model[45]:

$$Q_q^{(i)}(L) = N(N + 2)\sum_{k=1}^{L}\frac{\rho_{\varepsilon_q^{(i)}}^2(k)}{N - k} \tag{26}$$

where $\rho_{\varepsilon_q^{(i)}}(k)$ is the k-th lag autocorrelation of $\varepsilon_q^{(i)}$. If the fitted model is adequate, the standardized innovation $\varepsilon_q^{(i)}$ should take on an i.i.d sequence of random variables with the assumed distribution. Therefore, there will be no serial correlations detected from the innovations and $Q_q^{(i)}(L)$ is insignificant. Although some drawbacks have been pinpointed and other statistics are proposed[46], the Q-test is still the most popular one to check the autocorrelations of the residuals.

Data Availability

The data are owned by the Shanghai Qingyue Open Environmental Protection Data Center (https://data.epmap.org/). The center provides two options for accessing the data. Interested readers can browse the web site or send an email to support@epmap.org.cn for detailed information.

References

1. Chan, C. K. & Yao, X. H. Air pollution in mega cities in China. *Atmos. Environ.* **42**, 1–42, https://doi.org/10.1016/j.atmosenv.2007.09.003 (2008).
2. Wang, J., Zhang, X., Guo, Z. & Lu, H. Developing an early-warning system for air quality prediction and assessment of cities in China. Expert. *Sys. Appl.* **84**, 102–116, https://doi.org/10.1016/j.eswa.2017.04.059 (2017).
3. Xu, Y., Yang, W. & Wang, J. Air quality early-warning system for cities in China. *Atmos. Environ.* **148**, 239–257, https://doi.org/10.1016/j.atmosenv.2016.10.046 (2017).
4. Xie, W.-J., Jiang, Z.-Q. & Zhou, W.-X. Extreme value statistics and recurrence intervals of NYMEX energy futures volatility. *Econ. Model.* **36**, 8–17, https://doi.org/10.1016/j.econmod.2013.09.011 (2014).
5. Jiang, Z.-Q., Canabarro, A. A., Podobnik, B., Stanely, H. E. & Zhou, W.-X. Early warning of large volatilities based on recurrence interval analysis in Chinese stock markets. *Quant. Financ.* **16**, 1713–1724, https://doi.org/10.1080/14697688.2016.1175656 (2016).
6. Jiang, Z.-Q. *et al.* Short term prediction of extreme returns based on the recurrence interval analysis. *Quant. Financ.* **18**, 353–370, https://doi.org/10.1080/14697688.2017.1373843 (2018).
7. Niu, M., Gan, K., Sun, S. & Li, F. Application of decomposition-ensemble learning paradigm with phase space reconstruction for day-ahead $PM_{2.5}$ concentration forecasting. *J. Econom.* **196**, 110–118, https://doi.org/10.1016/j.jenvman.2017.02.071 (2017).
8. Bogachev, M. I. & Bunde, A. Improved risk estimation in multifractal records: Application to the value at risk in finance. *Phys. Rev. E* **80**, 026131, https://doi.org/10.1103/PhysRevE.80.026131 (2009).
9. Bogachev, M. I. & Bunde, A. On the predictability of extreme events in records with linear and nonlinear long-range memory: Efficiency and noise robustness. *Phys. A* **390**, 2240–2250, https://doi.org/10.1016/j.physa.2011.02.024 (2011).
10. Deluca, A., Moloney, N. R. & Corral, A. Data-driven prediction of thresholded time series of rainfall and self-organized criticality models. *Phys. Rev. E* **91**, 052808, https://doi.org/10.1103/PhysRevE.91.052808 (2015).
11. Denys, M., Gubiec, T., Kutner, R., Jagielski, M. & Stanley, H. E. Universality of market superstatistics. *Phys. Rev. E* **94**, 042305, https://doi.org/10.1103/PhysRevE.94.042305 (2016).
12. Beck, C. & Cohen, E. G. D. Superstatistics. *Phys. A* **322**, 267–275, 10.1016/S0378-4371(03)00019-0 (2003).
13. Beck, C., Cohen, E. G. D. & Rizzo, S. Atmospheric turbulence and superstatistics. *Europhys. News* **36**, 189–191, https://doi.org/10.1051/epn:2005603 (2005).
14. Tamazian, A., Nguyen, V. D., Markelov, O. A. & Bogachev, M. I. Universal model for collective access patterns in the Internet traffic dynamics: A superstatistical approach. *EPL (Europhys. Lett.)* **115**, 10008, https://doi.org/10.1209/0295-5075/115/10008 (2015).
15. Mark, C. *et al.* Bayesian model selection for complex dynamic systems. *Nat. Commun.* **9**, 1803, https://doi.org/10.1038/s41467-018-04241-5 (2018).
16. Herrera, R. & Schipp, B. Value at risk forecasts by extreme value models in a conditional duration framework. *J. Empir. Financ.* **23**, 33–47, https://doi.org/10.1016/j.jempfin.2013.05.00 (2013).
17. Bunde, A., Eichner, J. F., Havlin, S. & Kantelhardt, J. W. Return intervals of rare events in records with long-term persistence. *Phys. A* **342**, 308–314 (2004).
18. Bunde, A., Eichner, J. F., Kantelhardt, J. W. & Havlin, S. Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys. Rev. Lett.* **94**, 048701, https://doi.org/10.1103/PhysRevLett.94.048701 (2005).
19. Mazzarella, A. & Rapetti, F. Scale-invariance laws in the recurrence interval of extreme floods: An application to the upper Po river valley (northern Italy). *J. Hydrol.* **288**, 264–271 (2004).
20. Liu, C., Jiang, Z.-Q., Ren, F. & Zhou, W.-X. Scaling and memory in the return intervals of energy dissipation rate in three-dimensional fully developed turbulence. *Phys. Rev. E* **80**, 046304 (2009).
21. Cai, S.-M., Fu, Z.-Q., Zhou, T., Gu, J. & Zhou, P.-L. Scaling and memory in recurrence intervals of Internet traffic. *EPL (Europhys. Lett.)* **87**, 68001, https://doi.org/10.1209/0295-5075/87/68001 (2009).
22. Meng, H. *et al.* Effects of long memory in the order submission process on the properties of recurrence intervals of large price fluctuations. *EPL (Europhys. Lett.)* **98**, 38003, https://doi.org/10.1209/0295-5075/98/38003 (2012).
23. Ren, F. & Zhou, W.-X. Recurrence interval analysis of high-frequency financial returns and its application to risk estimation. *New J. Phys.* **12**, 075030, https://doi.org/10.1088/1367-2630/12/7/075030 (2010).
24. Du, P., Du, R., Ren, W., Lu, Z. & Fu, P. Seasonal variation characteristic of inhalable microbial communities in $PM_{2.5}$ in Beijing city. China. *Sci. Tot. Environ.* **610–611**, 308–315, https://doi.org/10.1016/j.scitotenv.2017.07.097 (2018).
25. Li, X. *et al.* Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environ. Pollut.* **231**, 997–1004, https://doi.org/10.1016/j.envpol.2017.08.114 (2017).
26. Dai, Y.-H. & Zhou, W.-X. Temporal and spatial correlation patterns of air pollutants in Chinese cities. *PLos One* **12**, e0182724, https://doi.org/10.1371/journal.pone.0182724 (2017).
27. Shi, K. & Liu, C. Q. Self-organized criticality of air pollution. *Atmos. Environ.* **43**, 3301–3304, https://doi.org/10.1016/j.atmosenv.2009.04.013 (2009).
28. Chelani, A. Long-memory property in air pollutant concentrations. *Atmos. Res.* **171**, 1–4, https://doi.org/10.1016/j.atmosres.2015.12.007 (2016).
29. Santhanam, M. S. & Kantz, H. Long-range correlations and rare events in boundary layer wind fields. *Phys. A* **345**, 713–721, https://doi.org/10.1016/j.physa.2004.07.012 (2005).
30. Bogachev, M. I. & Bunde, A. Universality in the precipitation and river runoff. *EPL (Europhys. Lett.)* **97**, 48011, https://doi.org/10.1209/0295-5075/97/48011 (2012).
31. Eichner, J. F., Koscielny-Bunde, E., Bunde, A., Havlin, S. & Schellnhuber, H. J. Power-law persistence and trends in the atmosphere: A detailed study of long temperature records. *Phys. Rev. E* **68**, 046133, https://doi.org/10.1103/Phys-RevE.68.046133 (2012).
32. Markelov, O., Duc, V. M. & Bogachev, M. Statistical modeling of the Internet traffic dynamics: To which extent do we need long-term correlations? *Phys. A* **485**, 48–60, https://doi.org/10.1016/j.physa.2017.05.023 (2017).
33. Simiu, E. & Heckert, N. A. Extreme wind distribution tails: A 'peaks over threshold' approach. *J. Struct. Eng.* **122**, 539–547, https://doi.org/10.1061/(ASCE)0733-9445(1996)122:5(539) (1996).
34. Engle, R. F. & Russell, J. R. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econom.* **66**, 1127–1162, https://doi.org/10.2307/2999632 (1998).
35. Ji, Z. & Kang, S. Evaluation of extreme climate events using a regional climate model for China. Int. *J. Clim.* **35**, 888–902, https://doi.org/10.1002/joc.4024 (2014).
36. Ning, G. *et al.* Characteristics of air pollution in different zones of Sichuan Basin, China. *Sci. Tot. Environ.* **612**, 975–984, https://doi.org/10.1016/j.scitotenv.2017.08.205 (2018).
37. Patton, A. P. *et al.* Spatial and temporal differences in traffic-related air pollution in three urban neighborhoods near an interstate highway. *Atmos. Environ.* **99**, 309–321, https://doi.org/10.1016/j.atmosenv.2014.09.072 (2014).
38. Li, X., Ma, Y., Wang, Y., Liu, N. & Hong, Y. Temporal and spatial analyses of particulate matter (PM_{10} and $PM_{2.5}$) and its relationship with meteorological parameters over an urban city in northeast China. *Atmos. Res.* **198**, 185–193, https://doi.org/10.1016/j.atmosres.2017.08.023 (2017).
39. Lunde, A. A generalized gamma autoregressive conditional duration model. Working paper (1999).
40. Luc, B. & Pierre, G. The logarithmic ACD model: An application to the bid-ask quote process of three NYSE stocks. *Annales d'Économie et de Stat.* **60**, 117–149, https://doi.org/10.2307/20076257 (2000).

41. Ng, K., Peiris, S. & Gerlach, R. Estimation and forecasting with logarithmic autoregressive conditional duration models: A comparative study with an application. *Expert. Sys. Appl.* **41**, 3323–3332, https://doi.org/10.1016/j.eswa.2013.11.024 (2014).
42. Hautsch, N. Assessing the risk of liquidity suppliers on the basis of excess demand intensities. *J. Financ. Econ.* **1**, 189–215, https://doi.org/10.1093/jjfinec/nbg010 (2003).
43. Openshaw, S. & Connolly, C. J. Empirically derived deterrence functions for maximum performance spatial interaction models. *Environ. Plan. A* **9**, 1068–1079, https://doi.org/10.1068/a091067 (1977).
44. Newey, W. K. & West, K. D. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. *Econom.* **55**, 703–708, https://doi.org/10.2307/1913610 (1987).
45. Ljung, G. M. & Box, G. E. P. On a measure of lack of fit in time series models. *Biom.* **65**, 297–303, https://doi.org/10.1093/biomet/65.2.297 (1978).
46. Bagnato, L., De Capitani, L. & Punzo, A. A diagram to detect serial dependencies: an application to transport time series. *Qual. Quant.* **51**, 581–594, https://doi.org/10.1007/s11135-016-0426-y (2017).

Acknowledgements

This work was partially supported by the China Scholarship Council (http://www.csc.edu.cn/) under Grant Number 201506740040 and the Fundamental Research Funds for the Central Universities (http://www.moe.gov.cn) under Grant Number 222201818006. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Yue-Hua Dai: Data collection, methodology, data analysis, draft; Zhi-Qiang Jiang: Methodology, supervision, validation, review; Wei-Xing Zhou: Methodology, supervision, validation, review and editing.

Additional Information

Competing Interests: The authors declare no competing interests.

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