Using real-time data and unsupervised machine learning techniques to study large-scale spatio-temporal characteristics of wastewater discharges and their influence on surface water quality in the Yangtze River Basin

Zhenzhen Di<sup>1</sup>, Miao Chang<sup>1\*</sup>, Peikun Guo<sup>1</sup>, Yang Li<sup>2</sup>, Yin Chang<sup>2</sup>

- 1 School of Environment, Tsinghua University, 100084, Beijing, China
- 2 Green Nest Smart Data Technologies (Beijing) Co. Ltd, 77 Shuangqing Rd, 100084, Beijing, China

## **Supplementary Material**

## **Figures**

## **Tables**

<sup>\*</sup> Corresponding author at: School of Environment, Tsinghua University, 100084, Beijing, China. E-mail addresses: changmiao@tsinghua.edu.cn (Chang, M.), <u>dzz17@tsinghua.org.cn</u> (Di, Z.).

methods based on yearly/weekly means of COD, NH <sub>3</sub> -N, pH and DO in wastewater
discharges from the sewage outlets and in surface water from the sites in the YRB in
2016 and 20179
Table S 4 Numbers and administrative regions of the 18 YRB sites and the 2213 YRB
wastewater-generating factories in different spatial PAM clusters in 2016 and 201710
Table S 5 Unpolluted week percentages and yearly means of pH, DO, COD and NH <sub>3</sub> -N in
surface water at the YRB sites in 2016 and 2017
Table S 6 Numbers and percentages of YRB sewage outlets in each EM cluster based on the
yearly means of COD, NH <sub>3</sub> -N, and pH in wastewater discharged in 2016 and 2017 12
Table S 7 YRB sewage outlet numbers in Clusters HPW (heavily polluted wastewater) and
UPW (unpolluted wastewater) in different industries by EM clustering in 2016 and 2017
Table S 8 Numbers of the YRB factories with heavily polluted wastewater discharges (HPW)
and their administrative regions and industrial classes in the heavily polluted zone (HPZ)
in 201614
Table S 9 Numbers of the YRB factories with heavily polluted wastewater discharges (HPW)
and their administrative regions and industrial classes in the heavily polluted zone (HPZ)
in 2017

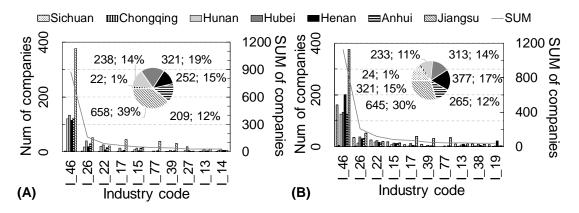


Figure S 1 Numbers of the wastewater-generating factories ranked in order of the top 10 industries in the YRB that published online monitoring data in 2016 (A) and in 2017 (B). (In the pie charts, numbers before the semicolon represent the sum of factories in a specific category and percentages after the semicolon represents the ratio of the YRB factories.)

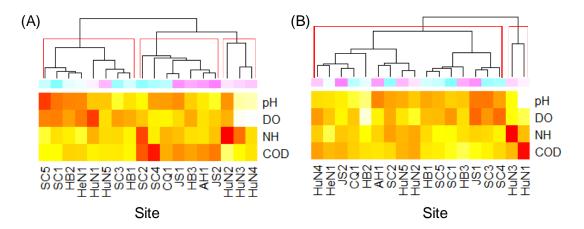


Figure S 2 Yearly concentration means of chemical oxygen demand (COD), ammonia nitrogen (NH<sub>3</sub>-N), dissolved oxygen (DO) and pH at the 18 surface water sections (sites) in the Yangtze River Basin (YRB) in 2016 (A) and 2017 (B) by hierarchical clustering with the Ward.D method (NH: NH<sub>3</sub>-N. The deeper the color is, the larger the value of some indicator is.).

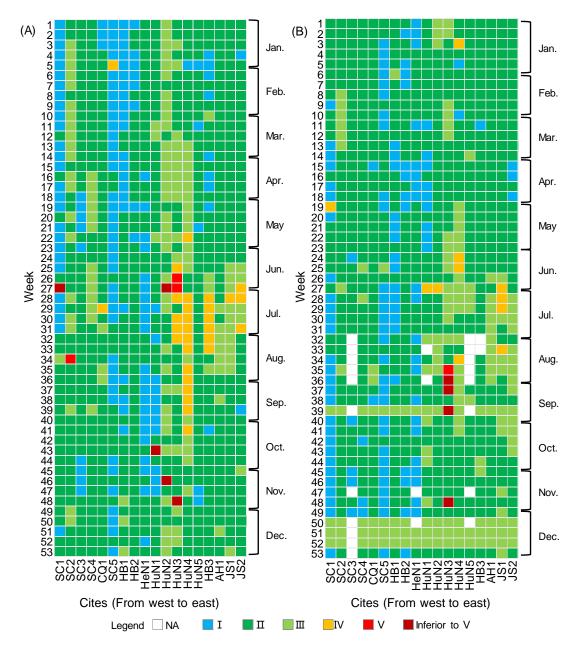


Figure S 3 Weekly water quality grades at the 18 YRB sites in 2016 (A) and 2017 (B).

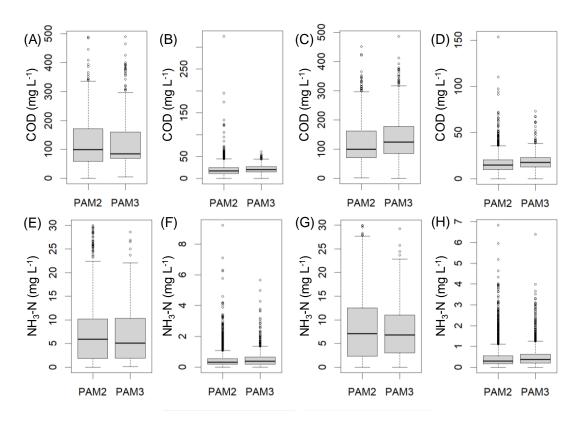


Figure S 4 COD and NH<sub>3</sub>-N weekly means in the YRB wastewater discharges in different spatial (PAM2 and PAM3) and water quality clusters (HPW clusters—clusters of heavily polluted wastewater: A, C, E and G; UPW clusters—clusters of unpolluted wastewater: B, D, F and H) in 2016 (A, B, C and D) and 2017 (E, F, G and H)

Table S 1 Basic information of the YRB sites, sewage outlets and their factories with monitoring data published online in 2016 and 2017 (from west to east).

			the YRB	Longitude	Latitude		City	Sewage	Factory	Wastewater	Data Availability
NO.	Site Code	River Name	Reaches	(East)	(North)	AR Name	Number	Outlet Number	Number	2016	2017
1	SC1	Yangtze River	Upper	101.66°E	26.59°N						
2	SC2	Minjiang River	Upper	103.76°E	29.51°N						
3	SC3	Minjiang River	Upper	104.43°E	28.78°N	Sichuan (SC)	21	335	321	No	Yes
4	SC4	Tuojiang River	Upper	105.45°E	28.90°N						
5	SC5	Jialing River	Upper	105.88°E	32.67°N						
6	CQ1	Yangtze River	Upper	105.85°E	29.02°N	Chongqing (CQ)	9	28	24	Yes	Yes
7	HuN1	Yuan River	Middle	112.13°E	28.92°N						
8	HuN2	Lishui River	Middle	112.13°E	29.47°N						
9	HuN3	Zishui River	Middle	112.63°E	28.80°N	Hunan (HuN)	15	272	241	Yes	Yes
10	HuN4	Xiangjiang River	Middle	112.84°E	28.34°N						
11	HuN5	Yangtze River	Middle	113.23°E	29.54°N						
12	HB1	Yangtze River	Middle	111.27°E	30.76°N						
13	HB2	Danjiangkou Reservoir	Middle	111.50°E	32.57°N	Hubei (HB)	14	356	322	Yes	Yes
14	HB3	Han River	Middle	114.22°E	30.58°N						
15	HeN1	Danjiangkou Reservoir	Middle	111.71°E	32.67°N	Henan (HeN)	18	416	378	Yes	Yes
16	AH1	Yangtze River	Lower	117.03°E	30.50°N	Anhui (AH)	16	298	269	Yes	Yes
17	JS1	Yangtze River	Lower	118.52°E	31.89°N	Lionagy (IC)	12	601	650	Vas	Vas
18	JS2	Jiajiang River	Lower	119.65°E	32.35°N	Jiangsu (JS)	13	681	658	Yes	Yes
SUM						7	106	2386	2213		

Table S 2 Industrial class list and numbers of the YRB wastewater-generating factories in 2016 and 2017

Industry Code	Industry Name Acronym	Industry Full Name	SUM in 2016	SUM in 2017	ISIC.Rev.4 Division ID	ISIC.Rev.4 Name
I_46	WPS	Water production and supply	872	1126	36 & 37	Water collection, treatment and supply & Sewerage
I_26	CCPM	Manufacture of chemicals and chemical products	160	205	20	Manufacture of chemicals and chemical products
I_22	PMPA	Manufacture of paper articles of paper	91	121	17	Manufacture of paper and paper products
I_15	BM	Manufacture of beverages	55	80	10 & 11	Manufacture of food products & Manufacture of beverages
I_17	TI	Textile industry	69	71	13	Manufacture of textiles
I_39	CCOEEM	Manufacture of computers, communication and other electrical equipment	44	53	26	Manufacture of computer, electronic and optical products
I_77	EPER	Ecological protection and environmental restoration	45	45	38 & 39 & 91	Waste collection, treatment and disposal activities; materials recovery & Remediation activities and other waste management services & Libraries, archives, museums and other cultural activities
I_13	ASFP	Processing of agricultural and sideline foods	30	44	10	Manufacture of food products
I_38	<b>EMEM</b>	Manufacture of electrical machinery and equipment	27	40	27	Manufacture of electrical equipment
I_19	LFFF	Leather, fur, feather and related products and manufacture of footwear	19	39	14 & 15	Manufacture of wearing apparel & Manufacture of leather and related products
I_27	PMCM	Manufacture of pharmaceuticals and medicinal chemical	32	38	21	Manufacture of pharmaceuticals, medicinal chemical and botanical products
I_32	SCPNFM	Smelting, calendering and processing of non-ferrous metals	26	35	24	Manufacture of basic metals

I_06	CMW	Coal mining and washing	22	33	05 & 08	Mining of coal and lignite & Other mining and quarrying
I_14	FPM	Manufacture of food products	29	31	10	Manufacture of food products
I_33	MP	Metal products	21	27	25	Manufacture of fabricated metal products, except machinery and equipment
I_09	MBNFMO	Mining and beneficiation of non-ferrous metal ores	22	26	07	Mining of metal ores
I_31	SCPFM	Smelting, calendering and processing of ferrous metals	23	25	24	Manufacture of basic metals
I_44	EPHGS	Electric power and heat generation and supply	20	21	35	Electricity, gas, steam and air conditioning supply
I_25	PPCNFP	Petroleum processing, coking and nuclear fuels processing	15	16	19	Manufacture of coke and refined petroleum products
I_28	CFM	Manufacture of chemical fibres	9	15	20	Manufacture of chemicals and chemical products
I_30	ONMP	Other non-metallic mineral products	9	10	23	Manufacture of other non-metallic mineral products
I_16	TP	Tobacco products	5	6	12	Manufacture of tobacco products
I_35	SPEM	Manufacture of special-purpose equipment	3	6	28 & 32	Manufacture of machinery and equipment n.e.c. & Other manufacturing
I_41	OM	Other manufacturing	5	5	32	Other manufacturing
I_81	OSA	Other service activities	4	4	81 & 96	Services to buildings and landscape activities & Other
_						personal service activities
I_20	WBRPS	Wood processing and products of wood, bamboo, rattan, palm and straw	2	2	16	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
I_21	FM	Manufacture of furniture	1	1	31	Manufacture of furniture
SUM			1700	2178		

Table S 3 Partitioning around medoids (PAM) and expectation—maximization (EM) clustering methods based on yearly/weekly means of COD, NH<sub>3</sub>-N, pH and DO in wastewater discharges from the sewage outlets and in surface water from the sites in the YRB in 2016 and 2017

<u>-</u>					Input D	ata											
Year of data	je			Yearly mea	ns		Weekly means						Number (k) identification of o	clusters			
r of	longitude	latitude	Sewage outle	ets	Sites	Se	ewage out	lets		Sites							
Yea	long	latit	COD NH3-N	pH COD	NH <sub>3</sub> -N DO	pH COD	NH <sub>3</sub> -N	pН	COD	NH <sub>3</sub> -N	DO		Model	k			
2016 &	DA.	M										Optimum av	erage silhouette width maximum	4(0(154)			
2017	PA	IVI											(asw)	4 (asw: 0.6154)			
2016			EM_A											6 (BIC: -6412.77)			
2016			EM_B	/									Mclust VVI (diagonal, varying	8 (BIC: -17467.57)			
2017			EM_C									ш	volume and shape) model	8 (BIC: -6997.97)			
2017			EM_D	/								iteri		7 (BIC: -21390.04)			
2016					EM_SA							n C	Mclust EEV (ellipsoidal, equal	5 (BIC: -39.27)			
2017					EM_SB							Bayesian Information Criterion (BIC)	volume and shape) model	2 (BIC: -88.06)			
2016							EM E					nforr (B)	Mclust VVI (diagonal, varying	2 (DIC: 900 56)			
2016							EM_E					an Ir	volume and shape) model	3 (BIC: -800.56)			
2017							EM_F					ayesi	Mclust VEI (diagonal, equal	2 (BIC: -611.20)			
2016										EM_SC		ŭ	shape) model	2 (BIC: -285.2)			
2017										EM_SD			Mclust EVI (diagonal, equal volume, varying shape) mode				

Table S 4 Numbers and administrative regions of the 18 YRB sites and the 2213 YRB wastewater-generating factories in different spatial PAM clusters in 2016 and 2017

PAM Cluster	Sample SUM	Number Factori		of	Nu	mber of Sites	Province
PAM1	351	345	321		6	5	SC
PAWII	331	343	24		O	1	CQ
			241			5	HuN
PAM2	551	544	302		7	2 (HB1, HB3)	НВ
			1				AH
			20			1 (HB2)	HB
PAM3	453	451	378		2	1 (HeN1)	HeN
PAIVIS	433	431	48		2		AH
			5				JS
DAM4	876	873	220		3	1	AH
PAM4	0/0	8/3	653		3	2	JS
SUM	2231	2213			18		

Note: The names in parentheses are the site codes.

Table S 5 Unpolluted week percentages and yearly means of pH, DO, COD and NH<sub>3</sub>-N in surface water at the YRB sites in 2016 and 2017

		2	016		2	017				
Site Code	Unpolluted Week		Year	ly Means	S	Unpolluted Week		Yearly	y Means	
	Percentage	pН	DO	COD	NH <sub>3</sub> -N	Percentage	pН	DO	COD	NH <sub>3</sub> -N
SC1	98.1%	8.04	9.07	1.85	0.17	98.1%	7.61	8.98	1.67	0.11
		(0.03)	(0.11)	(0.48)	(1.81)		(0.04)	(0.12)	(0.42)	(0.21)
SC2	98.1%	7.4	8.08	2.88	0.46	100.0%	7.85	8.72	2.18	0.34
		(0.04)	(0.14)	(0.33)	(0.55)		(0.06)	(0.11)	(0.27)	(0.32)
SC3	100.0%	731	8.63	1.98	0.18	100.0%	8.10	8.99	1.88	0.20
		(0.04)	(0.11)	(0.25)	(0.23)		(0.04)	(0.11)	(0.21)	(0.49)
SC4	100.0%	7.83	7.49	3.30	0.15	100.0%	7.86	9.61	2.19	0.13
		(0.02)	(0.21)	(0.26)	(0.37)		(0.04)	(0.17)	(0.23)	(0.28)
CQ1	98.1%	7.84	7.45	2.36	0.25	100.0%	7.23	8.33	2.30	0.22
		(0.05)	(0.16)	(0.34)	(0.44)		(0.05)	(0.23)	(0.58)	(0.43)
SC5	98.1%	8.31	9.03	1.85	0.08	100.0%	7.75	8.51	1.94	0.10
		(0.02)	(0.16)	(0.55)	(0.86)		(0.05)	(0.22)	(0.27)	(0.84)
HB1	100.0%	75	8.49	1.89	0.13	100.0%	7.82	8.45	2.01	0.20
		(0.05)	(0.12)	(0.29)	(1.16)		(0.04)	(0.3)	(0.25)	(0.62)
HB2	100.0%	795	8.61	2.11	0.14	100.0%	7.12	6.59	2.09	0.10
		(0.04)	(0.09)	(0.14)	(0.3)		(0.02)	(0.16)	(0.18)	(0.44)
HeN1	100.0%	794	9.12	2.10	0.08	100.0%	7.49	8.34	2.55	0.05
		(0.05)	(0.12)	(0.13)	(0.48)		(0.03)	(0.15)	(0.17)	(1.5)
HuN1	98.1%	7.63	10	1.92	0.26	98.1%	6.86	7.20	4.59	0.26
		(0.04)	(0.26)	(0.57)	(15)		(0.05)	(0.16)	(0.22)	(0.75)
HuN2	96.2%	7.89	8.4	1.63	0.58	98.1%	7.64	7.39	2.75	0.32
		(0.07)	(0.13)	(0.3)	(0.47)		(0.02)	(0.15)	(0.2)	(0.46)
HuN3	84.9%	7.12	6.22	1.87	0.38	90.6%	7.31	7.54	2.54	0.73
		(0.06)	(0.2)	(0.37)	(1.02)		(0.02)	(0.15)	(0.33)	(0.48)
HuN4	75.5%	7.1	6.16	2.01	0.17	92.5%	7.52	8.97	2.87	0.23
		(0.05)	(0.25)	(0.28)	(0.55)		(0.06)	(0.21)	(0.34)	(0.4)
HuN5	100.0%	7.64	<i>7.7</i> 5	1.97	0.18	100.0%	7.80	7.91	2.63	0.28
		(0.04)	(0.14)	(02)	(0.3)		(0.07)	(0.17)	(0.32)	(0.58)
HB3	88.7%	7.61	8.51	2.42	0.17	100.0%	7.65	8.01	1.45	0.17
		(0.05)	(0.26)	(0.31)	(0.44)		(0.03)	(0.13)	(0.14)	(0.49)
AH1	100.0%	7.47	<i>77</i> 4	2.54	0.16	100.0%	8.01	7.73	2.05	0.21
		(0.03)	(0.18)	(0.18)	(0.38)		(0.03)	(0.17)	(0.22)	(0.49)
JS1	98.1%	791	7.8	2.54	023	92.5%	8.08	9.68	1.71	0.21
		(0.03)	(0.24)	(027)	(0.41)		(0.06)	(0.28)	(0.77)	(0.44)
JS2	94.3%	732	799	2.66	0.22	100.0%	7.42	7.69	2.64	0.24
		(0.02)	(0.22)	(031)	(0.69)		(0.08)	(0.25)	(0.41)	(1.07)

Note: Values in parentheses represent coefficients of variation calculated from the weekly means of each monitoring indicator.

Table S 6 Numbers and percentages of YRB sewage outlets in each EM cluster based on the yearly means of COD, NH<sub>3</sub>-N, and pH in wastewater discharged in 2016 and 2017

Year	Cluster	1	2	3	4	5	6	7	8	SUM
2016	EM_A	46	8	68	66	185	94	/	/	467
		(9.8%)	(1.7%)	(14.5%)	(14.0%)	(39.4%)	(20.0%)	/	/	
	EM_B	370	96	195	49	420	186	48	68	1432
		(25.7%)	(6.7%)	(13.6%)	(3.4%)	(29.2%)	(12.9%)	(3.3%)	(4.7%)	
2017	EM_C	8	6	146	78	87	66	26	111	528
		(1.5%)	(1.1%)	(27.5%)	(14.7%)	(16.4%)	(12.5%)	(4.9%)	(20.9%)	
	EM_D	461	272	93	155	506	62	262	/	1811
		(25.3%)	(14.9%)	(5.1%)	(8.5%)	(27.8%)	(3.4%)	(14.4%)	/	

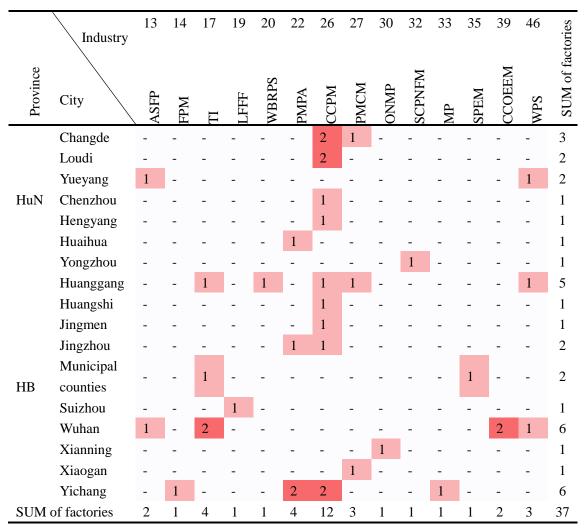
Note: Percentages in parentheses represent proportions of sewage outlets in a specific cluster.

Table S 7 YRB sewage outlet numbers in Clusters HPW (heavily polluted wastewater) and UPW (unpolluted wastewater) in different industries by EM clustering in 2016 and 2017

	2016									20	017			
Inclustry Code	Industry Acronym	HPW_2016	HPW_EM_A_1	HPW_EM_A_2	HPW_EM_B_8	HPW_EM_B_4	UPW_2016	HPW_2017	HPW_EM_C_7	HPW_EM_C_2	HPW_EM_D_3	HPW_EM_D_6	UPW_2017	ISICRev.4 Division ID
I_06	CMW	1	1	-	-	1	4	-	-	-	-	-	4	05&08
I_09	MBNFMO	-	-	-	-	-	2	-	-	-	-	-	4	07
I_13	ASFP	7	1	-	3	3	-	5	1	-	4	1	3	10
I_14	FPM	4	1	-	3	1	3	11	1	-	7	4	1	10
I_15	BM	11	1	-	9	1	8	16	-	-	12	4	11	10 & 11
I_16	TP	1	1	-	1	-	1	3	1	1	1	1	1	12
I_17	TI	12	10	1	3	2	1	20	3	2	5	15	1	13
I_19	LFFF	5	2	-	4	1	-	8	1	-	3	5	-	14 & 15
I_20	WBRPS	1	-	1	-	-	-	-	-	-	-	-	-	16
I_21	FM	-	-	-	-	-	1	-	-	-	-	-	-	31
I_22	PMPA	6	3	1	3	1	12	12	3	-	2	10	7	17
I_25	PPCNFP	1	-	-	1	-	3	1	-	-	1	-	2	19
I_26	CCPM	35	7	2	15	19	12	25	4	-	20	5	9	20
I_27	<b>PMCM</b>	7	3	-	6	-	3	7	2	-	5	2	5	21
I_28	CFM	-	-	-	-	-	-	3	-	-	2	1	-	20
I_30	ONMP	2	1	-	1	1	2	1	-	-	-	1	1	23
I_31	SCPFM	-	-	-	-	-	4	-	-	-	-	-	7	24
I_32	SCPNFM	2	1	-	-	2	3	2	1	-	2	-	4	24
I_33	MP	3	-	1	1	2	2	4	-	1	2	1	1	25
I_35	SPEM	2	2	-	1	1	-	1	1	-	1	-	-	28 & 32
I_38	<b>EMEM</b>	-	-	-	-	-	2	2	-	2	-	-	5	27
I_39	CCOEEM	12	6	-	7	2	6	15	6	-	14	2	6	26
I_41	OM	2	-	-	2	-	1	2	-	-	2	-	1	32
<u>I_44</u>	<b>EPHGS</b>	1	1	-	-	-	2	-	-	-	-	-	5	35
I_46	WPS	21	3	4	9	11	275	16	3	-	14	2	345	36&37
I_77	EPER	6	2	-	4	1	14	9	1	-	5	4	18	38 & 39 &91
I_81	OSA	-	-	-	-	-	1	1	-	-	-	1	1	81 & 96
SUM		142					362	164					442	

Note: Hyphen(-) represents zero. HPW\_EM\_A\_1 represents EM\_A\_1 cluster with high COD and NH<sub>3</sub>-N; HPW\_EM\_A\_2 represents EM\_A\_2 cluster with abnormal pH; HPW\_EM\_B\_8 represents EM\_B\_8 cluster with high COD and NH<sub>3</sub>-N; HPW\_EM\_B\_4 represents EM\_B\_4 cluster with high NH<sub>3</sub>-N; HPW\_EM\_C\_7 represents EM\_C\_7 cluster with high COD and NH<sub>3</sub>-N; HPW\_EM\_C\_2 represents EM\_C\_2 cluster with abnormal pH; HPW\_EM\_D\_3 represents EM\_D\_3 cluster with high COD and NH<sub>3</sub>-N; HPW\_EM\_D\_6 represents EM\_D\_6 cluster with high COD.

Table S 8 Numbers of the YRB factories with heavily polluted wastewater discharges (HPW) and their administrative regions and industrial classes in the heavily polluted zone (HPZ) in 2016



Note: Hyphen(-) represents zero.

Table S 9 Numbers of the YRB factories with heavily polluted wastewater discharges (HPW) and their administrative regions and industrial classes in the heavily polluted zone (HPZ) in 2017

	Industry	13	14	16	17	19	22	26	27	30	33	35	39	
Province	City	ASFP	FPM	TP	II	LFFF	PMPA	CCPM	PMCM	ONMP	MP	SPEM	CCOEEM	SUM of factories
	Changsha	-	1	1	-	-	-	-	-	1	-	-	-	3
	Changde	-	-	-	-	-	1	-	1	-	-	-	-	2
	Chenzhou	-	-	-	-	-	-	-	1	-	-	-	-	1
HuN	Huaihua	-	-	-	-	-	1	-	-	-	-	-	-	1
	Loudi	-	-	-	-	-	-	1	-	-	-	-	-	1
	Shaoyang	-	-	-	-	1	-	-	-	-	-	-	-	1
	Zhangjiajie	-	1	-	-	-	-	-	-	-	-	-	-	1
	Huanggang	-	-	-	1	-	-	1	2	-	-	-	-	4
	Jingmen	-	-	-	-	1	-	-	-	-	-	-	-	1
НВ	Municipal counties	-	-	-	1	-	-	-	-	-	-	1	-	2
	Wuhan	1	-	-	1	-	-	-	-	-	-	-	2	4
	Yichang	1	1	-	-	-	2	-	1	-	1	-	-	6
SUM o	of factories	2	3	1	3	2	4	2	5	1	1	1	2	27

Note: Hyphen(-) represents zero.