**Supplementary Information for**

Enhancing Watershed Management through Adaptive Source Apportionment under a Changing Environment

Wenzhuo Wang[1], Guowangcheng Liu[1], Yuhan Zhang[1], Mingjing Wang[1], Yan Pan[1], Xinyi Meng[1], Junfeng Xiong[2], Zhenyao Shen[1], and Lei Chen[1, *]

[a] *State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing 100875, P.R. China*

[b] *Nanjing Institute of Geography & Limnology Chinese Academy of Sciences, Nanjing 210008, P.R. China*

Corresponding to: Lei Chen, **Tel/fax**: +86 10 58807031, **Email:** chenlei1982bnu@bnu.edu.cn

## Supplementary Tables

**Supplementary Table 1 Hydrological years and socio-economic development conditions under the different scenarios**

| Number | Hydrological year | Socio-economic development conditions |
|---|---|---|
| 1 | Wet year | High (20s) |
| 2 | Dry year | High (20s) |
| 3 | Normal year | Low (80s) |
| 4 | Normal year | Middle (00s) |
| 5 | Normal year | High (20s) |

Note: 80s represents the period from 1985 to 1999; 00s represents the period from 2000 to 2009; 20s represents the period from 2010 to 2020.

**Supplementary Table 2 Contributions of the various pollution sources under the different scenarios**

| Source type | Scenarios 1 | | Scenarios 2 | | Scenarios 3 | | Scenarios 4 | | Scenarios 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Nitrogen | Phosphorus | Nitrogen | Phosphorus | Nitrogen | Phosphorus | Nitrogen | Phosphorus | Nitrogen | Phosphorus |
| Industrial sewage | 20.09 | 1.44 | 20.09 | 1.44 | 296.08 | 23.53 | 80.53 | 6.4 | 20.09 | 1.44 |
| Urban domestic | 336.36 | 18.73 | 336.36 | 18.73 | 4059.19 | 293.11 | 1104.09 | 79.73 | 336.36 | 18.73 |
| Urban nonpoint source pollution | 143.53 | 11.9 | 121.46 | 10.07 | 147.01 | 0.59 | 328.84 | 1.5 | 106.41 | 8.82 |
| Intensive livestock and poultry breeding | 1448.51 | 163.04 | 915.27 | 105.65 | 13.79 | 2.72 | 227.55 | 48.31 | 1495.06 | 95.03 |
| Rural domestic | 1246.4 | 93.06 | 1155.87 | 86.31 | 245.85 | 10.5 | 508.58 | 21.73 | 1206.04 | 90.05 |
| Planting industry | 7672.47 | 314.1 | 1904.55 | 50.28 | 17466.36 | 166.44 | 32278.28 | 556.56 | 2617.95 | 142.14 |
| Dispersed livestock and poultry breeding | 175.26 | 74.62 | 110.74 | 48.35 | 675.79 | 153.84 | 682.65 | 217.82 | 144.71 | 34.78 |
| Aquaculture | 663.81 | 130.6 | 667.11 | 131.25 | 20.09 | 1.9 | 116.09 | 11 | 669 | 131.62 |
| Background value pollution from natural land and river sediments | 102.02 | 27.89 | 34.25 | 11.12 | 66.44 | 5.44 | 172.27 | 16.38 | 55.34 | 12.29 |

**Supplementary Table 3 Proportions of the various pollution sources contributing to the different scenarios**

| Source type | Scenarios 1 | | Scenarios 2 | | Scenarios 3 | | Scenarios 4 | | Scenarios 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Nitrogen | Phosphorus | Nitrogen | Phosphorus | Nitrogen | Phosphorus | Nitrogen | Phosphorus | Nitrogen | Phosphorus |
| Industrial sewage | 0.17% | 0.17% | 0.38% | 0.31% | 1.29% | 3.58% | 0.23% | 0.67% | 0.30% | 0.27% |
| Urban domestic | 2.85% | 2.24% | 6.39% | 4.04% | 17.66% | 44.54% | 3.11% | 8.31% | 5.06% | 3.50% |
| Urban nonpoint source pollution | 1.22% | 1.42% | 2.31% | 2.17% | 0.64% | 0.09% | 0.93% | 0.16% | 1.60% | 1.65% |
| Intensive livestock and poultry breeding | 12.27% | 19.52% | 17.38% | 22.81% | 0.06% | 0.41% | 0.64% | 5.04% | 22.48% | 17.77% |
| Rural domestic | 10.56% | 11.14% | 21.95% | 18.63% | 1.07% | 1.60% | 1.43% | 2.26% | 18.13% | 16.83% |
| Planting industry | 64.97% | 37.60% | 36.17% | 10.85% | 75.97% | 25.29% | 90.93% | 58.01% | 39.36% | 26.57% |
| Dispersed livestock and poultry breeding | 1.48% | 8.93% | 2.10% | 10.44% | 2.94% | 23.38% | 1.92% | 22.70% | 2.18% | 6.50% |
| Aquaculture | 5.62% | 15.63% | 12.67% | 28.34% | 0.09% | 0.29% | 0.33% | 1.15% | 10.06% | 24.61% |
| Background value pollution from natural land and river sediments | 0.86% | 3.34% | 0.65% | 2.40% | 0.29% | 0.83% | 0.49% | 1.71% | 0.83% | 2.30% |

**Supplementary Table 4 Trend terms of the equations for each pollution source emission during 1985-2020**

| Source type | Time series | Trend | |
|---|---|---|---|
| | | Nitrogen | Phosphorus |
| Urban domestic | 1985-2003 | $Y_t$= 7950.63-1.813GDP ($R^2$=0.81) | $Y_t$=574.12-0.131GDP ($R^2$=0.81) |
| | 2004-2020 | $Y_t$= 999.58-0.027GDP ($R^2$=0.76) | $Y_t$= 72.18-0.002GDP ($R^2$=0.76) |
| Industrial sewage | 1985-2003 | $Y_t$= 579.92-0.132GDP ($R^2$=0.81) | $Y_t$= 46.10-0.011GDP ($R^2$=0.81) |
| | 2004-2020 | $Y_t$= 72.91-0.002GDP ($R^2$=0.76) | $Y_t$= 5.79-0.00016GDP ($R^2$=0.76) |
| Rural domestic | 1989-1991 | $Y_t$= -0.15+6.22$POP_r$ ($R^2$=0.99) | $Y_t$= -0.007+0.266$POP_r$ ($R^2$=0.99) |
| | 1992-2020 | $Y_t$= 15.576+5.775$POP_r$ ($R^2$=0.95) | $Y_t$= 0.691+0.246$POP_r$ ($R^2$=0.94) |
| Intensive livestock and poultry breeding | 1985-2020 | $Y_t$= (0.94pig+4.4cattle+4.4sheep+696 poultry) *scale-rate ($R^2$=0.99) | $Y_t$= (0.20pig+0.5cattle+0.5sheep+194 poultry) *scale-rate ($R^2$=0.99) |
| Dispersed livestock and poultry breeding | 1985-2020 | $Y_t$= (0.94pig+4.4cattle+4.4sheep+696 poultry) *(1-SR) ($R^2$=0.99) | $Y_t$= (0.20pig+0.5cattle+0.5sheep+194 poultry) *(1-SR) ($R^2$=0.99) |
| Aquaculture | 1985-2001 | $Y_t$=0.351+0.002AP ($R^2$=0.99) | $Y_t$=0.033+ 0.000204AP ($R^2$=0.99) |
| | 2002-2022 | Yt= -3.082+ 0.002AP ($R^2$=0.99) | $Y_t$= -0.292+0.000214AP ($R^2$=0.99) |
| Planting industry | 1985-2020 | $Y_t$=-28201219.04+19637.62AREAf-20998.65Pt+115.84NF ($R^2$=0.70) | $Y_t$=9879043.99-5379.47AREAf +56870.69Pt-1.18PF ($R^2$=0.76) |
| Urban nonpoint source pollution | 1985-2020 | $Y_t$ =-31548.91+291.04P+ 3039.87AREAur (R2=0.93) | $Y_t$=-2268.04+1.89P+15.88AREAur (R2=0.96) |
| Background value pollution from natural land and river sediments | 1985-2020 | $Y_t$=-1334281.10+1061.69P+ 167.84AREAna ($R^2$=0.71) | $Y_t$=-131338.36+89.88P+ 27.53AREAna ($R^2$=0.72) |

Note: AP: Aquatic production; $P_t$: Trend term in rainfall series; $P_p$: Periodic term in rainfall series; NF: Nitrogenous fertilizer; PF： Phosphate fertilizer; SR: scale-rate; GDP: Gross Domestic Product; $POP_r$ Rural Resident Population; AREAf: Farmland area; AREAur: Urban land area; Arena: Natural land area.

**Supplementary Table 5 Periodic and random terms of the equations for each pollution source emission during 1985-2020**

| Source type | Pollution | Term | Equation |
|---|---|---|---|
| Planting industry | Nitrogen | Periodic | $Y_P = -4.64 + 3990.21 \times \left( \frac{P_p - 0.37}{0.61} \times \cos\frac{0.7\pi}{1.5} - \sqrt{1 - \left(\frac{P_p - 0.37}{0.61}\right)^2} \times \sin\frac{0.7\pi}{1.5} \right)$ |
| | | Random | $X_t$=-0.71* $X_{t-1}$-0.54* $X_{t-2}$+$\varepsilon_t$+0.57*$\varepsilon_{t-1}$+0.95*$\varepsilon_{t-2}$ |
| | Phosphate | Periodic | $Y_P = -0.09 + 239.61 \times \left( \frac{P_p - 0.37}{0.61} \times \cos\frac{1.19\pi}{1.5} - \sqrt{1 - \left(\frac{P_p - 0.37}{0.61}\right)^2} \times \sin\frac{1.19\pi}{1.5} \right)$ |
| | | Random | Define X(t). Then, ΔY(t) = X(t+1) - X(t) represents the first-order difference of this function. $Y_t$=-0.65* $Yr_{-1}$ |
| Urban nonpoint source pollution | Nitrogen | Periodic | $Y_P = -0.03 + 62.02 \times \left( \frac{P_p - 0.37}{0.61} \times \cos\frac{1.7\pi}{1.5} + \sqrt{1 - \left(\frac{P_p - 0.37}{0.61}\right)^2} \times \sin\frac{1.7\pi}{1.5} \right)$ |
| | | Random | Define X(t). Then, ΔY(t) = X(t+1) - X(t) represents the first-order difference of this function. $Y_t$=$\varepsilon_t$-0.93*$\varepsilon_{t-1}$ |
| | Phosphate | Periodic | $Y_P = 0.22776 \times \left( \frac{P_p - 0.37}{0.61} \times \cos\frac{1.06\pi}{1.5} - \sqrt{1 - \left(\frac{P_p - 0.37}{0.61}\right)^2} \times \sin\frac{1.06\pi}{1.5} \right)$ |
| | | Random | $X_t$=-0.62* $X_{t-1}$-0.43* $X_{t-2}$-0.77* $X_{t-3}$-0.58* $X_{t-4}$ |

| Source type | Pollution | Term | Equation |
|---|---|---|---|
| Background value pollution from natural land and river sediments | Nitrogen | Periodic | $Y_P = -0.05 + 89.16 \times \left( \frac{P_p - 0.37}{0.61} \times \cos\frac{1.02\pi}{1.5} - \sqrt{1 - \left(\frac{P_p - 0.37}{0.61}\right)^2} \times \sin\frac{1.02\pi}{1.5} \right)$ |
| | | Random | Define X(t). Then, ΔY(t) = X(t+1) - X(t) represents the first-order difference of this function. $Y_t = \varepsilon_t + -1.0 * \varepsilon_{t-1}$ |
| | Phosphate | Periodic | $Y_P = -0.01 + 5.41 \times \left( \frac{P_p - 0.37}{0.61} \times \cos\frac{0.68\pi}{1.5} + \sqrt{1 - \left(\frac{P_p - 0.37}{0.61}\right)^2} \times \sin\frac{0.68\pi}{1.5} \right)$ |
| | | Random | Define X(t). Then, ΔY(t) = X(t+1) - X(t) represents the first-order difference of this function. $Y_t = -0.67 * Y_{t-1}$ |

**Supplementary Table 6 Trend terms of the equations fitting the change in contribution through yearly changes during 1985-2020**

| Source type | Time series | Trend | |
|---|---|---|---|
| | | Nitrogen | Phosphorus |
| **Urban domestic** | 1985-2020 | y=-57.83x+116466($R^2$=0.76) | y=-4.6992x+9463.4 ($R^2$=0.78) |
| **Industrial sewage** | 1985-2020 | y=-13.875x+27944 ($R^2$=0.76) | y=-0.4332x+872.11 ($R^2$=0.78) |
| **Rural domestic** | 1989-2010 | y=3.6628x-7241.6 ($R^2$=0.95) | y=0.0493x-97.439 ($R^2$=0.94) |
| | 2011-2020 | y=-1.878x+3879 ($R^2$=0.87) | y=-0.0312x+64.071 ($R^2$=0.89) |
| **Intensive livestock and poultry breeding** | 1989-2010 | y=3.3275x-6625.3 ($R^2$=0.65) | y=0.1766x-351.65 ($R^2$=0.61) |
| | 2011-2020 | y=-24.026x+48616 ($R^2$=0.94) | y=-1.3482x+2727.7 ($R^2$=0.94) |
| **Dispersed livestock and poultry breeding** | 1985-2020 | y=-2.5896x+5282.1 ($R^2$=0.79) | y=-0.067x+138.18 ($R^2$=0.59) |
| **Aquaculture** | 1985-2022 | y=0.7336x-1456.8 ($R^2$=0.97) | y=0.0274x-54.515 ($R^2$=0.96) |
| **Planting industry** | 1985-2013 | y=505.76x-997900 ($R^2$ = 0.89) | y=6.3815x-12631 ($R^2$ = 0.82) |
| | 2014-2020 | y=-3034.5x+6E+06 ($R^2$ =0.99) | y=40.203x-80672 ($R^2$ = 0.99) |
| **Urban nonpoint source pollution** | 1985-2020 | y=0.2538x-500.49 ($R^2$ = 0.91) | y=0.0004x-0.8343 ($R^2$ = 0.63) |
| **Background value pollution from natural land and river sediments** | 1985-2020 | y=0.0923x-180.55 ($R^2$ = 0.32) | y=0.0016x-3.0652 ($R^2$ = 0.32) |

Note: The x represents the year, and the y represents the contribution amount to the watershed export (t/a).

**Supplementary Table 7 Pareto fronts (first five order) for dynamic source apportionment in the Hangbu River**

| Pareto front | Nitrogen | | Phosphate | |
|---|---|---|---|---|
| | **Dynamic Key Source Area** | **Accumulated contribution proportion** | **Dynamic Key Source Area** | **Accumulated contribution proportion** |
| Pareto 1st-order frontier | Planting-2; Planting-3; Planting-6; Planting-8; Planting-13; Planting-14; Planting-15; Planting-19; Planting-21; Planting-27 and Planting-28 | 87.83% | Intensive livestock and poultry breeding-8; Intensive livestock and poultry breeding-28; Planting-2; Planting-3; Planting-6 and Urban domestic-2 | 57.68% |
| Pareto 2nd-order frontier | Intensive livestock and poultry breeding-3; Planting-10; Planting-16; Planting-18; Planting-20; Urban domestic-2 and Urban domestic-14 | 93.86% | Intensive livestock and poultry breeding-3; Intensive livestock and poultry breeding-14; Intensive livestock and poultry breeding-27; Planting-4; Planting-8; Planting-10; Planting-13; Planting-14; Planting-16; Planting-27; Urban domestic-14 | 84.52% |
| Pareto 3rd -order frontier | rural domestic-8; Intensive livestock and poultry breeding-10; Planting-1; Planting-24 and Urban domestic-18 | 95.05% | Intensive livestock and poultry breeding-10; Intensive livestock and poultry breeding-19; Intensive livestock and poultry breeding-24; Planting-15; Planting-21; Planting-28 | 89.76% |
| Pareto 4th-order frontier | Intensive livestock and poultry breeding-7; Planting-17; Planting-25; Urban domestic-8; Urban domestic-20 and Urban domestic-28 | 96.05% | Dispersed livestock and poultry breeding-6; Intensive livestock and poultry breeding-5; Intensive livestock and poultry breeding-7; Intensive livestock and poultry breeding-13; Intensive livestock and poultry breeding-21; Planting-19; Planting-24 | 92.76% |
| Pareto 5th-order frontier | Intensive livestock and poultry breeding -8; Planting-22 and Urban domestic-24 | 96.48% | Dispersed livestock and poultry breeding-14; Intensive livestock and poultry breeding-6; Intensive livestock and poultry breeding-16; Planting-1; Urban domestic-8 | 93.81% |

Note: Planting-2 represents the planting industry in sub-watershed No.2.

**Supplementary Table 8 Transfer matrix between different land use types from 1985 to 2020**

| 2020 / 1985 | Farmland | Forest land | Grassland | Water | Construction land | Wetland | Bare land |
|---|---|---|---|---|---|---|---|
| Farmland | 59.24% | 4.06% | 0.00% | 0.33% | 2.75% | 0.00% | 0.00% |
| Forest land | 1.18% | 28.78% | 0.00% | 0.00% | 0.03% | 0.00% | 0.00% |
| Grassland | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| Water | 0.85% | 0.02% | 0.00% | 1.68% | 0.04% | 0.00% | 0.00% |
| Construction land | 0.09% | 0.002% | 0.00% | 0.06% | 0.88% | 0.00% | 0.00% |
| Wetland | 0.00% | 0.01% | 0.00% | 0.01% | 0.01% | 0.00% | 0.00% |
| Bare land | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

**Supplementary Table 9 Data sources of the Hangbu River**

| Data type | Data name | Index | Year | Accuracy | Data sources |
|---|---|---|---|---|---|
| **Spatial data** | Digital Elevation Model | Elevation information | 2018 | 30m*30m | National Geomatics Center of China (http://www.ngcc.cn/ngcc/) |
| | Types of land use | Land cover distribution | 1990-2020 | 30m*30m | China Land Cover Dataset (https://zenodo.org) |
| | Soil type | Distribution of soil types | 1995 | 1:1000，000 | Harmonized World Soil Database version 1.2 |
| | Population distribution | Distribution of Population | 1985-2020 | 1km*1km | LandScan Global Population Data (https://landscan.ornl.gov/) |
| **Attribute data** | Hydrological data | Flow | 2010-2020 | Monthly Scale | Chaohu Lake Administration Bureau |
| | Water quality data | Various forms of phosphorus, nitrogen | 2010-2020 | Monthly Scale | Chaohu Lake Administration Bureau |
| | Meteorological data | Precipitation, temperature (including the highest and lowest temperature), relative humidity, wind speed and solar radiation | 1953-2020 | Diurnal scale | China Meteoro-logical Assimilation Driving Dataset and Chaohu Lake Administration Bureau |
| | Soil attribute type | Soil physical properties (density, hydraulic conductivity, field capacity, soil available water) and soil chemical properties (contents of soil organic nitrogen, nitrate nitrogen, organic phosphorus and mineralized phosphorus) | 1995 | 1km*1 km | Harmonized World Soil Database version 1.2 |

| Data type | Data name | Index | Year | Accuracy | Data sources |
|---|---|---|---|---|---|
| **Attribute data** | Crop management data | Cultivation, sowing, fertilization, irrigation and harvesting during crop growth | 1985-2020 | Annual scale | Field investigation |
| **Pollution source data** | Agricultural pollution sources | Planting, centralized and decentralized livestock and poultry breeding, rural decentralized life, aquaculture, etc. | 1985-2020 | Annual scale | Field investigation and Chaohu Lake Administration Bureau |
| | Urban pollution sources | Sewage-treatment plant pollution, urban NPS, etc. | 1985-2020 | Annual scale | Field investigation and Chaohu Lake Administration Bureau |

**Supplementary Table 10 Meteorological station information in the Chaohu Lake basin**

| Number | Station Name | Latitude/° | Longitude/° | Elevation/m | Index |
|---|---|---|---|---|---|
| 1 | Luan | 31.75 | 116.48 | 60.50 | Precipitation, temperature, relative humidity and wind speed |
| 2 | Huoshan | 31.40 | 116.33 | 68.10 | Precipitation, temperature, relative humidity and wind speed |
| 3 | Tongcheng | 31.05 | 116.95 | 44.90 | Precipitation, temperature, relative humidity and wind speed |
| 4 | Hefei | 31.85 | 117.28 | 23.60 | Precipitation, temperature, relative humidity, wind speed and solar radiation |
| 5 | Chaohu | 31.62 | 117.87 | 22.40 | Precipitation, temperature, relative humidity, wind speed and solar radiation |
| 6 | Maanshan | 31.68 | 118.52 | 23.40 | Precipitation, temperature, relative humidity and wind speed |
| 7 | Wuhu | 31.13 | 118.52 | 23.20 | Precipitation, temperature, relative humidity and wind speed |
| 8 | Tongling | 30.95 | 117.80 | 37.10 | Precipitation, temperature, relative humidity and wind speed |

**Supplementary Table 11 Description of the parameter formula in the weather generator**

| Parameter name | Meaning | Formula |
|---|---|---|
| TMPMX | Monthly average maximum temperature /℃ | $\mu mx_{mon} = \sum_{d=1}^{N} T_{mx,mon} / N$ |
| TMPMN | Monthly average minimum temperature /℃ | $\mu mn_{mon} = \sum_{d=1}^{N} T_{mn,mon} / N$ |
| TMPSTDMX | Standard deviation of maximum temperature | $\sigma mx_{mon} = \sqrt{\sum_{d=1}^{N} (T_{mx,mon} - \mu mx_{mon})^2 / (N-1)}$ |
| TMPSTDMN | Standard deviation of minimum temperature | $\sigma mn_{mon} = \sqrt{\sum_{d=1}^{N} (T_{mn,mon} - \mu mn_{mon})^2 / (N-1)}$ |
| PCPMN | Average monthly rainfall /mm | $\overline{R}_{mon} = \sum_{d=1}^{N} R_{day,mon} / yrs$ |
| PCPSTD | Standard deviation of rainfall scalar | $\sigma_{mon} = \sqrt{\sum_{d=1}^{N} (R_{day,mon} - \overline{R}_{mon})^2 / (N-1)}$ |
| PCPSKW | Skewness coefficient of rainfall | $g_{mon} = N \sum_{d=1}^{N} (R_{day,mon} - \overline{R}_{mon})^3 / (N-1)(N-2)(\sigma_{mon})^3$ |
| PR_W1 | Number of dry days in a month /d | $P_i(W/D) = (days_{W/D,i}) / (days_{dry,i})$ |
| PR_W2 | Number of wet days in a month /d | $P_i(W/W) = (days_{W/W,i}) / (days_{wet,i})$ |
| PCPD | Average rainfall days /d | $\overline{d}_{wet,i} = day_{wet,i} / yrs$ |
| DEWPT | Dew point temperature /℃ | $\mu dew_{mon} = \sum_{d=1}^{N} T_{dew,mon} / N$ |
| SOLARAV | Monthly average solar radiation /[kJ(m$^2$·d)] | $\mu rad_{mon} = \sum_{d=1}^{N} H_{day,mon} / N$ |
| WNDAV | Average monthly wind speed /(m/s) | $\mu wnd_{mon} = \sum_{d=1}^{N} T_{wnd,mon} / N$ |

**Supplementary Table 12 Parameter of the SWAT model calibration**

| Parameter name | Trans form | Meaning | Parameter range | The best value |
|---|---|---|---|---|
| SURLAG.bsn | v | Surface runoff lag time | 0.05-24 | 13.64 |
| ADJ_PKR.bsn | v | Peak rate adjustment factor for sediment routing in the subcatchment | 0.5-2 | 1.75 |
| SLSUBBSN.hru | v | Average slope length | 10-150 | 32.75 |
| CN2.mgt | v | SCS runoff curve number f | 35-98 | 63.51 |
| BIOMIX.mgt | v | Biological mixing efficie | 0-1 | 0.30 |
| CH_COV1.rte | v | Channel erodibility factor | -0.05-0.6 | 0.06 |
| CH_ERODMO().rte | v | Jan. channel erodability factor | 0-1 | 0.77 |
| SOL_AWC().sol | r | Available water capacity of the soil layer | -1-1 | -0.23 |
| SOL_Z().sol | r | Depth from soil surface to bottom of layer | -1-1 | -0.071 |
| SOL_ZMX().sol | r | Maximum rooting depth of soil profile | -1-1 | 0.18 |
| SOL_K().sol | r | Saturated hydraulic conductivity | -1-1 | 0.67 |
| ESCO.hru | v | Soil evaporation compensation factor | 0-1 | 0.54 |
| ALPHA_BF.gw | v | Baseflow alpha factor | 0-1 | 0.65 |
| GW_DELAY.gw | v | Groundwater delay | 0-500 | 78.75 |
| RCHRG_DP.gw | v | Deep aquifer percolation fraction | 0-1 | 0.59 |
| GW_REVAP.gw | v | Groundwater "revap" coefficient | 0.02-0.2 | 0.06 |
| CH_K2.rte | v | Effective hydraulic conductivity in main channel alluvium | -0.01-5000 | 2712.50 |
| EPCO.hru | v | Plant uptake compensation factor | 0-1 | 0.55 |
| CH_N2.rte | v | Manning's "n" value for the main channel | -0.01-0.3 | 0.19 |
| GWQMN.gw | v | Treshold depth of water in the shallow aquifer required for return flow to occur | 0-5000 | 4637.50 |
| CANMX.hru | v | Maximum canopy storage | 0-100 | 33.75 |
| SOL_ORGN().chm | v | Initial organic N concentration in the soil layer | 0-100 | 67.00 |
| SOL_NO3().chm | v | Initial NO3 concentration in the soil layer | 0-100 | 25.00 |
| K_N.wwq | v | Michaelis-Menton half-saturation constant for nitrogen | 0.01-0.3 | 0.29 |
| RS3.swq | v | Benthic source rate for NH4-N in the reach at 20°C | 0-1 | 0.37 |
| RS4.swq | v | Rate coefficient for organic N settling in the reach at 20 °C | 0-1 | 0.01 |
| AI6.wwq | v | Rate of oxygen uptake per unit NO2-N oxidation | 0.01-0.02 | 0.02 |
| NPERCO.bsn | v | Nitrogen percolation coefficient | 0-1 | 0.19 |
| SOL_ORGP().chm | v | Initial organic P concentration in surface soil layer | 0-100 | 78.38 |
| SOL_SOLP().chm | v | Initial labile P concentration in surface soil layer | 0-100 | 9.70 |

| Parameter name | Trans form | Meaning | Parameter range | The best value |
|---|---|---|---|---|
| ERORGP.hru | v | Organic P enrichment ratio | 0-5 | 2.69 |
| AI2.wwq | v | Fraction of algal biomass that is phosphorus | 0.01-0.02 | 0.02 |
| BC4.swq | v | Rate constant for mineralization of organic P to dissolved P in the reach at 20°C | 0.01-0.7 | 0.32 |
| SPEXP.bsn | v | Exponent parameter for calculating sediment reentrained in channel sediment routing | 1-1.5 | 1.41 |
| SPCON.bsn | v | Linear parameter for calculating the maximum amount of sediment that can be reentrained during channel sediment routing | 0.0001-0.01 | 0.01 |
| USLE_P.mgt | v | USLE equation support pra | 0-1 | 0.73 |
| USLE_K().sol | r | USLE equation soil erodibility (K) factor | -1-1 | -0.82 |

**Supplementary Table 13 Results of the SWAT model calibration**

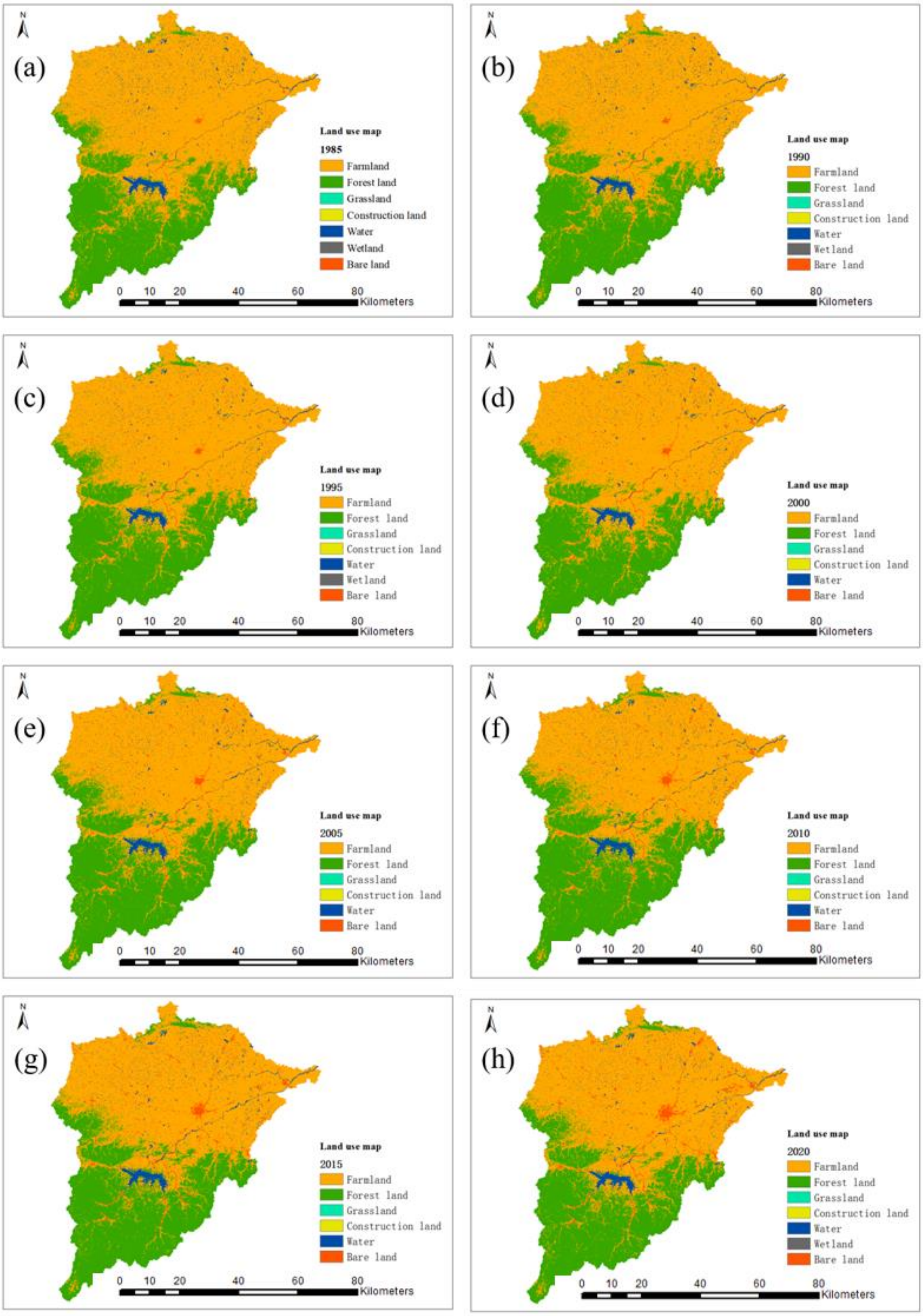
| Date | Simulation period | Data Type | $R^2$ | *NSE* | PBAIS |
|---|---|---|---|---|---|
| 2010.1-2015.12 | Calibration | Flow | 0.93 | 0.89 | -0.7% |
| | | Nitrogen | 0.72 | 0.68 | 17.2% |
| | | Phosphorus | 0.49 | 0.48 | -1.1% |
| 2016.1-2019.12 | Verification | Flow | 0.79 | 0.77 | -6.4% |
| | | Nitrogen | 0.62 | 0.58 | 18.4% |
| | | Phosphorus | 0.43 | 0.42 | -1.3% |

Note. $R^2$ is Pearson's correlation coefficient; *NSE* is the Nash-Sutcliffe efficiency; PBAIS is the Percent bias.

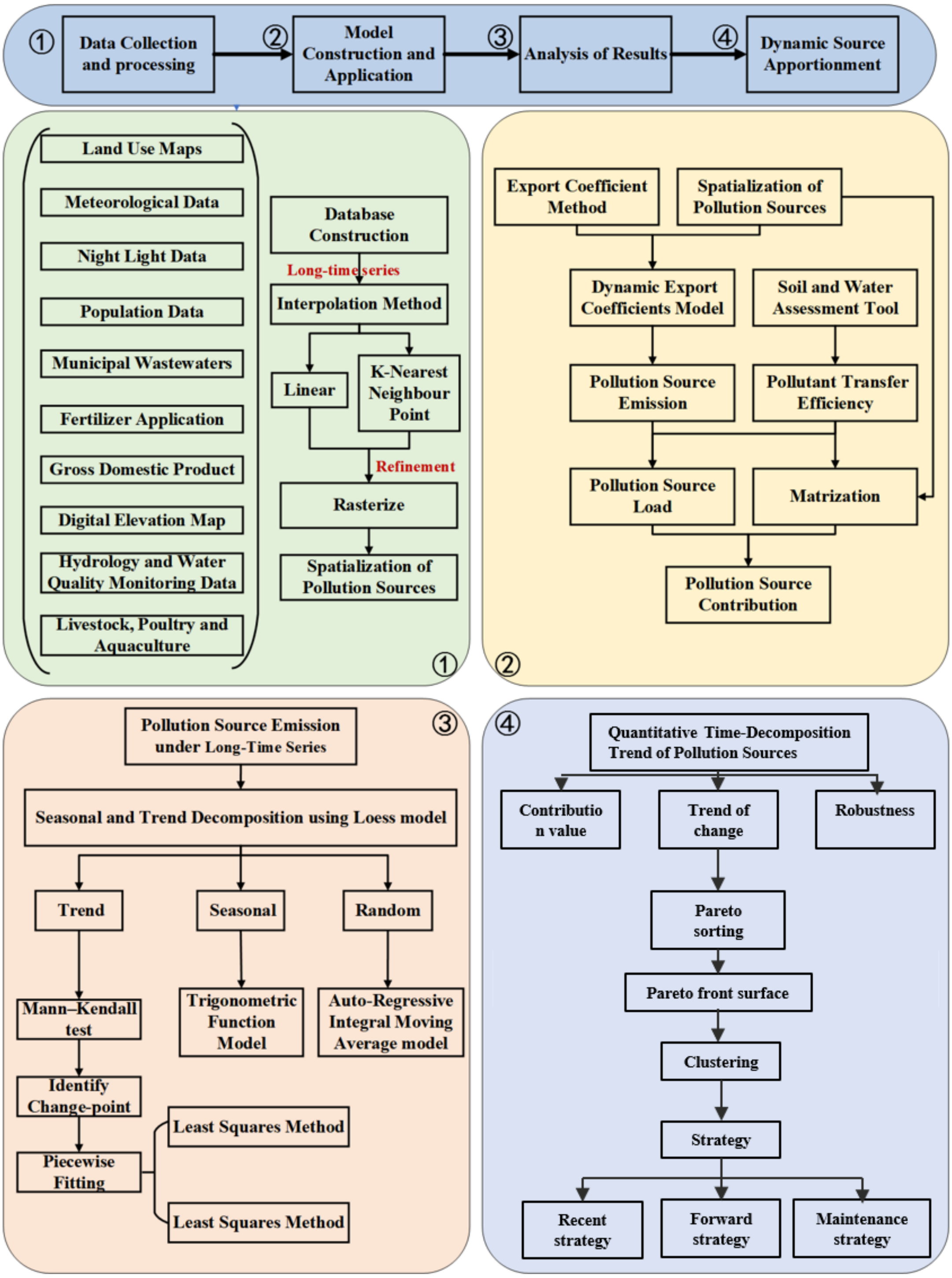
**Supplementary Table 14 Nomenclature Table**

| Name | Meaning |
|---|---|
| STL | Seasonal Trend Decomposition using locally weighted regression method |
| ARIMA | Autoregressive Integrated Moving Average |
| NPSs | Non-point sources |
| M-K | Mann-Kendall |
| $R^2$ | Coefficient of determination |
| DECM | Dynamic export coefficient model |
| SWAT | Soil and Water Assessment Tool |
| AR | Autoregressive |
| MA | Moving Average |
| NSGA | Non-dominated Sorting Genetic Algorithm |
| NSE | Nash-Sutcliffe model efficiency |
| PBIAS | Percent bias |
| PI | Phosphorus Index |
| MFA | Materials Flow Analysis |

## Supplementary Figures

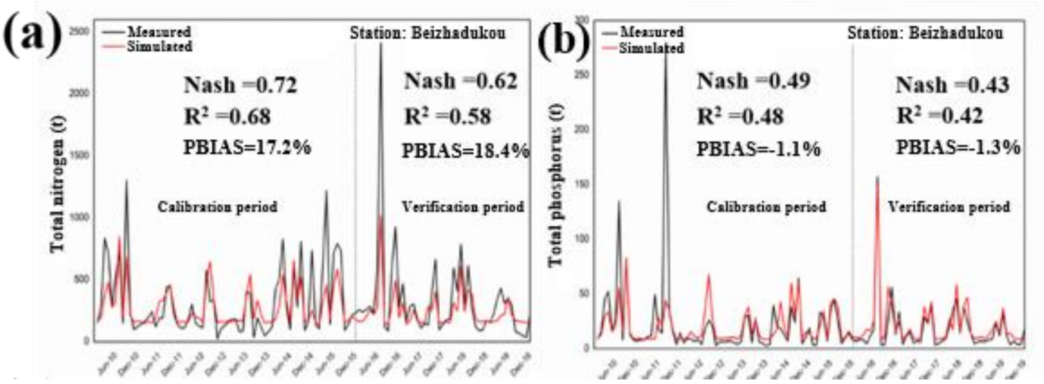



**Supplementary Figure 1 Temporal changes in land use distribution in the Hangbu River in (a) 1985; (b) 1990; (c) 1995; (d) 2000; (e) 2005; (f) 2010; (g) 2015; (h) 2020**

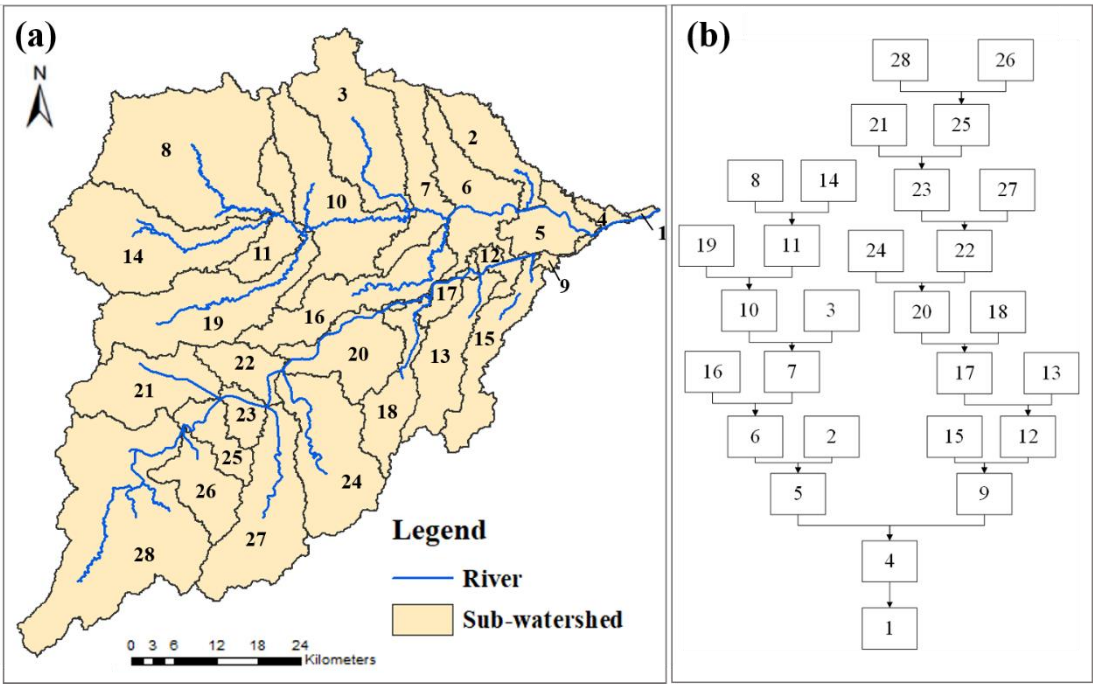



42

43 **Supplementary Figure 2 Conceptual overview (flow chart) of general study methods**




**Supplementary Figure 3 Comparison of simulated and measured monthly (a) nitrogen and (b) phosphorus at the outlet of the Hangbu River**




**Supplementary Figure 4 (a) Sub-watersheds of the Hangbu River; (b) upstream-downstream relationship in the Hangbu River**

Note: The encoding represents different sub-watersheds in the Hangbu River.

# Supplementary Notes

The surface area of Chaohu Lake spans 769 km$^2$, making it a renowned lake in China, and the basin surrounding it is home to a total population of 10.6 million people. Situated in the southwest region of the Chaohu Lake basin, the Hangbu River is an agricultural watershed covering an area of 4246 km$^2$. The Hangbu River lies within the northern subtropical humid monsoon climate zone and is characterized by considerable differences in annual rainfall. It is located between the coordinates 116°22′-117°22′ E and 30°59′-31°45′ N. The land use in the Hangbu River is classified into seven types: farmland, forestland, construction land, water, grassland, wetland, and bare land. The population data were obtained from LandScan Global Population Data (https://landscan.ornl.gov/). The downloaded remote sensing data from 1985 to 2020 had a resolution of 1 km×1 km. The land use data with a resolution of 30 km×30 km from 1990 to 2020 were obtained from the China Land Cover Dataset (https://zenodo.org). The division of the sub-watersheds was performed by geostatistical analysis using a digital elevation model, which was obtained from the National Geomatics Center of China (http://www.ngcc.cn/ngcc/). The daily temperature and rainfall data from 1953–2020 were obtained from the China Meteorological Data Service Center (http://data.cma.cn/). Eight meteorological stations are located in the watershed. Long-term measured data, including monthly nitrogen and phosphorus concentrations and monthly streamflow data from 2010–2019, were used here. In this study, township data, including livestock and poultry production, aquatic production, and fertilizer application conditions, were obtained from reliable statistical yearbooks. Data on wastewater treatment plants were derived from government documents, precise locations, literature, and expertise. In areas with impermeable ground, storm sewers convey polluted runoff directly into rivers or other receiving waters, defined as urban NPS pollution.

# Supplementary Methods

## Supplementary Section 1

The indicator data from multiple sets of statistical data are integrated into the matched standard administrative unit, and the fused indicator data are processed, including abnormal value processing and missing value processing.

The fused data are processed to identify outliers based on the coefficient of variation (CV).

$$C_V = \frac{SD}{MN} \times 100\% \tag{1}$$

where $C_V$ is the coefficient of variation, $SD$ is the standard deviation, and $MN$ is the average value.

The missing data after fusion are filled with data based on interpolation methods. The LIM approximates the missing value of a time series based on the proportional relationship between the time and observed value.

$$Y(t_j) = Y(t_i) + [Y(t_k) - Y(t_i)]/(t_k - t_i)\,(t_j - t_i) \tag{2}$$

where $t$ is a time series set, the observed values corresponding to the times of $t_i$ and $t_k$ are $Y(t_i)$ and $Y(t_k)$, respectively, and $Y(t_j)$ is the missing sample value at the time $t_j$ (i < j < k).

If the trend of the data series is stable, the linear interpolation effect is better. However, if the influencing factors are complex and variable, the accuracy is sharply reduced. The KNI searches for k samples of observation points closest to the points to be estimated and assigns the points to be estimated with the weighted sum of the observations of these sample points[1]. The neighbour relationship between sample points is:

$$D_i = \frac{1}{|t_j - t_i|} \tag{3}$$

$$Y(t_j) = \sum_{i=i}^{k} \frac{D_i Y(t_i)}{\sum D_i} \tag{4}$$

where $D_i$ is the neighbour relationship between sample points *i* and *j*, and $Y(t_j)$ is a set of k adjacent points adjacent to $t_j$.

The data were pre-normalized to facilitate the use of the DSA method. The normalization was denoted by:

$$\acute{x_i} = \frac{x_i - \bar{x}}{x_{sd}} \tag{5}$$

where the $\acute{x_i}$ is the normalized value of the observed $x_i$ and $\bar{x}$ and $x_{sd}$ are the mean and standard deviation of the observed data, respectively.

**Supplementary Section 2**

The export coefficient method was proposed in 1996 (Johnes, 1996), and this method was used to calculate pollutant emissions from different sources.

$$L = \sum_{j=1}^{n} E_i A_t(I_j) + Pre \tag{6}$$

where $L$ is the load intensity of the NPSs inflow water body, kg/a; $E_i$ is the export coefficient of the *i*-th source, kg/(hm$^2$·a), or the export coefficient of population, livestock, and poultry, kg/(ind·a); $A_t$ is the area, hm$^2$ or livestock, population, and aquaculture output of the *t*-th land use type; $I_i$ is the input of the *i*-th source, kg; and *Pre* is the pollution of precipitation input.

The performance of the SWAT model was evaluated with the coefficient of determination ($R^2$), the Nash-Sutcliffe model efficiency (*NSE*), and percent bias (PBIAS).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n}(\overline{y_i} - y_i)^2} \tag{7}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n}} \tag{8}$$

$$Pbias = \sum \frac{\widehat{y_i} - y_i}{y_i} \times 100\% \tag{9}$$

where $n$ is the total number of data points, $\widehat{y_i}$ and $y_i$ are the $i$-th predicted and observed values, respectively $\overline{y_i}$ is the mean of $y_i$, and $y_{max}$ and $y_{min}$ are the maxima and minima of $y_i$, respectively.

**Supplementary Section 3**

The Markov algorithm was used to define upstream and downstream relationships. $S_1$, $S_2$, … $S_n$ represent the different sub-watersheds, $n$ is the number of sub-watersheds, and the upstream-downstream relationship in a sub-watershed can be represented by a $n \times n$ matrix. When the $S_j$-th sub-watershed is the next downstream sub-watershed of the $S_i$-th sub-watershed, its one-step transition probability is defined as 1, which means that the river flows from the $S_i$-th sub-watershed to the $S_j$-th sub-watershed. The mathematical expression was as follows:

$$H\,(i,j) = \begin{cases} 1, When\ S_i\ and\ S_j\ have\ adjacent\ upstream\ and\ downstream\ relations \\ 0, Other\ scenarios \end{cases} \tag{10}$$

The $n \times n$ matrix is defined as follows, where the matrix element $r_{ii}$ represents the transfer proportion of pollutants in the $i$-th sub-watershed.

$$R = \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_n \end{bmatrix} \tag{11}$$

where $R$ is the pollutant transfer efficiency.

Based on the upstream-downstream relationship in the Hangbu River, the migration and transformation processes of pollutants were further defined.

$$r_i = \frac{\left(\sum_j^n U_j^i + P_{in}\right) - P_{out}}{P_{out}} \tag{12}$$

$$P_j^i = U_j^i \prod_{i=1}^{l} r_i \tag{13}$$

$$d_i = \left( \sum_{j=1}^{l} P_j^i \middle/ \sum \sum_{j=1}^{l} P_j^i \right) \times 100\% \tag{14}$$

where $r_i$ is the channel transfer coefficient of the $i$-th sub-watershed; $U_j^i$ is the $i$-th kind of sources emission in the $j$-th sub-watershed; $n$ is the number of sources; $P_{in}$ and $P_{out}$ are the total input and output load of the $i$-th sub-watershed based on SWAT; $P_j^i$ is $i$-th kind of sources load to the outlets from the $j$-th sub-watershed; $l$ is the number of reaches where the pollutant was transferred to the specific water bodies, and $d_i$ is the proportion of the load of the $i$-th source into the watershed outlet, %.

## Supplementary Section 4

**The Mann-Kendall test and least squares method.** The Mann-Kendall test and linear trend test approaches served to examine the trends for annual variables. The slope of linear models can generally quantitatively characterize the rise or decline of interest variables. Furthermore, the least-squares method fits a straight line to determine the changing trend of the sequence.

Given a series $x(t)$ with the length of n, the null hypothesis of no trend assumes that the series $x(t)$ are independently distributed. The M-K test is based on the test statistic S:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x(j) - x(i)) \tag{15}$$

With

$$\begin{cases} sgn(x) = 1 & if \ \ x > 0 \\ sgn(0) = 0 & \\ sgn(x) = -1 & if \ \ x < 0 \end{cases} \tag{16}$$

A positive (negative) value of S indicates an upward (downward) trend.

Mann and Kendall found that the statistic $S$ is approximately normally distributed when $n$ > 8, with the mean $E(S)$ and variance $Var(S)$ as follows:

$$E(S) = 0$$

$$Var(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{i=1}^{m} t_i i(i-1)(2i+5)\right] \tag{17}$$

where, $t_i$ is the number of data in the tied group and $m$ is the number of tied groups. The standardized test statistic $Z$ follows the standard normal distribution:

$$Z = \begin{cases} (S-1)/\sqrt{Var(S)} & if \ \ S > 0 \\ 0 & if \ \ S > 0 \\ (S+1)/\sqrt{Var(S)} & if \ \ S > 0 \end{cases} \tag{18}$$

The null hypothesis of no trend is rejected if the absolute value of $Z$ is bigger than the theoretical value $Z_{1-\alpha/2}$, where $\alpha$ is the statistical significance level concerned. This article selects $\alpha$=0.05 confidence level.

$$y = kx + b \tag{19}$$

The least-squares method determines the trend change of the sequence using the first-order coefficient k of the straight line. If k>0, the sequence exhibits an upward trend. If k=0, then the sequence is a stationary sequence. When k<0, the sequence exhibits a downward trend.

ARIMA is a widely used statistical analysis method that helps in understanding and forecasting time series data by capturing the underlying patterns. The ARIMA model is based on the idea that if a variable in a time series exhibits random and non-stationary behavior, a mathematical model can be used to approximate the variable sequence[2]. The importance of using ARIMA modeling lies in its ability to identify and quantify temporal patterns and trends in environmental data, providing a nuanced understanding of pollution dynamics. In the analysis using ARIMA modeling, specific attention was paid to the parameters 'p', 'd', and 'q', which are crucial in understanding the dynamics of pollution source changes. The parameter 'p', indicating the order of the autoregressive part, was used to assess how previous values in the time series influence current values[3]. This is particularly important in environmental studies where past pollution levels can have a substantial impact on current conditions. The 'd' value, representing the degree of differencing, was crucial for transforming our environmental time series data into a stationary format, a necessary step for reliable ARIMA modeling. This parameter helped in identifying how many times data needed to be differenced to highlight patterns such as trends or seasonal effects. Lastly, 'q', the order of the moving average part, was used to account for the influence of past error terms on the current value of the series. This is particularly relevant when dealing with environmental data, which often contain random fluctuations that can obscure underlying trends. Due to its flexibility and rigor in constructing an appropriate model at each step, ARIMA models have been widely used in the environmental field to forecast future environmental problems, such as drought, flood, and eutrophication. The best-fit response model and ARIMA model were identified based on the evaluation indicators, such as the correlation coefficient ($R^2$), root-mean-square-error (*RMSE*) and normalized-root-mean-square-error (NRMSE).

**ADF Test Formula**:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \tag{20}$$

where $\Delta$ is the difference operator, $y_t$ is the time series data, $\alpha$, $\beta$, $\gamma$, and $\delta$ are parameters, and $\varepsilon_t$ is the error term.

**Autocorrelation Function (ACF)**:

$$ACF(k) = \frac{\sum_{t=k+1}^{T} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2} \tag{21}$$

where $k$ is the lag order, $T$ is the total length of the time series, $Y_t$ is the value of the time series at time $t$, and $\bar{Y}$ is the mean of the time series. The ACF is used to identify the order of the autoregressive model within the data.

**Partial Autocorrelation Function (PACF)**:

$$PACF(k) = Corr(Y_t - \widehat{Y}_t, Y_{t-k} - \widehat{Y}_{t-k}) \quad (22)$$

where $k$ is the lag order and $\widehat{Y}_t$ is the linear prediction of the data point at time $t$. The PACF is used to identify the order of the moving average model in the time series.

**General Form of the ARIMA Model**:

$$1 - \sum_{i=1}^{p} \emptyset_i L^i)(1 - L)^d y_t = (1 + \sum_{j=1}^{q} \theta_j L^j)\varepsilon_t \quad (23)$$

where $p$ is the order of the autoregressive terms, $d$ is the order of differencing, $q$ is the order of the moving average terms, $L$ is the lag operator, $\emptyset$ and $\theta$ are model parameters, $y_t$ is the time series data, and $\varepsilon_t$ is the error term.

**Formulas for AIC and BIC**:

- AIC=2k-2ln(L)
- BIC= ln(n)k-2ln(L)

where k is the number of parameters in the model, *L* is the likelihood of the model, and *n* is the sample size.

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \quad (24)$$

where $y_{max}$ and $y_{min}$ are the maxima and minima of $y_i$.

This involves selecting a subset of data from the vicinity of the predicted variable, and then performing linear or quadratic regression on the subset using the weighted least squares method to reduce the weight of values far from the estimated point. This process is repeated for each point in the time series to accurately decompose the entire curve. The STL model can be used with both additive and multiplicative time series models. The additive model assumes that each component's influence is independent of the others, and each component is expressed in absolute terms. In this study, the dynamic source emission model adopts the additive model because there is no influence of each factor on different components. The random terms of the time series from the planting industry, urban NPS and natural land pollution were constructed based on the ARIMA model.

**Supplementary Section 5**

**Fast Dominance Sorting.** Before performing the selection operation, the dominance and partial order relationships between each chromosome are examined. The determination of dominance relationship adopts the Pareto dominance method. If the objective function consists of *r* objective functions and can be expressed as:

$$F(x) = (f_1(\mathrm{x}), f_2(\mathrm{x}), \cdots, f_r(\mathrm{x})) \tag{25}$$

where $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_n) \in X \subset R^n$ is $n$-dimensional decision variables; $X$ is the decision space, which refers to all feasible solutions. For a given practical problem $\min f(x)$, if $\forall x_A, x_B \in X_f$, and:

$$\bigwedge_{i \in I} (f_i(x_A)) \geq (f_i(x_B)) \tag{26}$$

where $x_B$ is said to be dominant over $x_A$, denoted as $x_A \prec x_B$. Partially ordered set includes non-inferior grade and aggregation distance. The partial order relationship includes non-dominance ranking and crowding distance. During the sorting process, chromosomes are initially ranked based on the non-dominated front they belong to. For individuals within the same rank, their crowding distances are considered to select the more superior ones, ensuring the distribution of chromosomes.

In this study, data pre-processing and statistical analyses were completed in MATLAB 9.7 (MathWorks, USA), EViews 10 (Econometrics Views, USA), Excel 2019 (Microsoft Corporation, USA), SPSS 20 (Statistical Product Service Solutions, USA), Origin 9.0 (Origin Lab Corporation, USA), and ArcGIS 10.2 (Environmental Systems Research Institute, USA) software. The ARIMA modelling processes were performed in IBM SPSS statistics.

## Supplementary References


1. Balaji, R. & Upmanu, L. A *k*-nearest-neighbor simulator for daily precipitation and other weather variables. Water Resour. Res. 35(10), 3089-3101(1990).

2. Box, G., Jenkins, G.M. & Reinsel, G.C. Time series analysis forecasting and control. J. Time Ser. Anal. 31(2), 238-242 (1976).

3. Collar, N.M. et al. Linking fire-induced evapotranspiration shifts to streamflow magnitude and timing in the western United States. J. Hydrol. 612, 128242 (2022).