Supplementary information for

Substantial increase of organic carbon storage in Chinese lakes

Dong Liu[1,2], Kun Shi[1,3]*, Peng Chen[4], Nuoxiao Yan[1], Lishan Ran[5], Tiit Kutser[6], Andrew N. Tyler[2], Evangelos Spyrakos[2], R. Iestyn Woolway[7], Yunlin Zhang[1,3], Hongtao Duan[1]*

[1] Key Laboratory of Lake and Watershed Science for Water Security, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China.

[2] Biological and Environmental Sciences, School of Natural Sciences, University of Stirling, Stirling, UK.

[3] Ecosystem Research Station of Lake Qiandaohu, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China.

[4] State Key Laboratory of Satellite Ocean Environment Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou, China.

[5] Department of Geography, The University of Hong Kong, Pok Fu Lam Road, Hong Kong.

[6] Estonian Marine Institute, University of Tartu, Mäealuse 14, Tallinn, Estonia.

[7] School of Ocean Sciences, Bangor University, Menai Bridge, Anglesey, UK.

* Corresponding author: Kun Shi and Hongtao Duan, Email: kshi@niglas.ac.cn and htduan@niglas.ac.cn

Contents of this file

Supplementary notes

Supplementary Note 1. Matching Landsat reflectance

This study adopted atmospherically corrected surface reflectance of TM/Landsat-5, ETM+/Landsat-7, OLI/Landsat-8, and OLI/Landsat-9 to remotely retrieve DOC and POC concentrations during 1984-2023. For the *in-situ* DOC or POC concentrations (Supplementary Table 1), we matched Landsat reflectance at the blue (R_{blue}), green (R_{green}), red (R_{red}), and near-infrared (R_{nir}) bands within a spatial window of ± 1 pixel and a time window of ± 7 days[1]. Then, the matched non-water reflectance was removed according to the pixel quality flag provided by Google Earth Engine (GEE). For a specific measured DOC or POC concentration, if more than one match-up were found, we just kept the one which was closest to the sampling time. Finally, we obtained 2,324 and 1,498 match-ups for the *in-situ* DOC and POC concentrations, respectively. Moreover, we divided the matched surface reflectance by pi (π) to get remote sensing reflectance.

Supplementary Note 2. Identifying freshwater and saline lake types

This study developed a weighted decision tree method to identify freshwater and saline lake types using eight lake basin properties (Datasets VII-IX in Supplementary Table 1). The weighted decision tree was developed via the following three steps:

[1] Binarization of lake salinity. Based on the national lake survey[2] and our sampling data (Supplementary Table 1), we got measured conductivies for 1,550 Chinese lakes. According to the conductivities, we distinguished and defined binary values for freshwater lakes (≤ 2,000 μs cm^{-1}, value = 1) and saline lakes (> 2,000 μs cm^{-1}, value = 2)[3].

[2] Determination of segmentation thresholds. For the measured conductivities, we matched mean basin property values in the sampling year and randomly selected 70% match-ups as the training data. Then, through a stepwise search

method from the minimum to maximum, the optimal segmentation threshold of each basin property was determined to obtain the highest identification accuracy for the remaining 30% testing match-ups.

[3] Weighted calculation of the decision tree. The final identification result was obtained through the weighted calculation using Supplementary Equation (1).

$$\text{Type} = \text{round}(\sum_{i=1}^{N}(P_i \times T_i) / N) \quad (1)$$

where Type = 1 for freshwater lakes and Type = 2 for saline lakes; N = 8; P_i denotes the identification type when using each basin property; and T_i indicates the identification accuracy for each basin property. T_i has values of 92.9%, 80.2%, 90.8%, 90.5%, 88.5%, 92.8%, 81.0%, and 88.6% for the factrors of evaporation, LAI_HighVeg, LAI_LowVeg, runoff, Temp2m, precipitation, wind speed, and DEM, respectively. Supplementary Equation (1) got an overall accuracy of 94.0% and was then applied to identify lake types of 24,366 Chinese lakes during 1984-2023 (Supplementary Fig. 1).

Supplementary Note 3. Delineating lake basin boundaries

This study delineated lake basin boudaries for 24,366 Chinese lakes of the HydroLAKES dataset. For the 584 large lakes (> 20 km^2), three steps were included to delineate basin boundaries:

[1] Water flow direction was calculated from the DEM data using the D8 algorithm[4], and rivers inflowing into lakes were visually selected by overlaying the flow direction results and the river networks of the HydroRIVERS dataset[5];

[2] Basin boundaries for the rivers inflowing into lakes were automatically delineated using a published software with a user-friendly graphical interface[6];

[3] All river basins for a specific lake were merged to one. The basin boundaries for the 584 large lakes are shown in Supplementary Fig. 2.

For other small lakes (≤ 20 km^2, N = 23,782), 79.6% of which have areas < 1 km^2, basin boundaries were determined using the HydroBASINS dataset at level 12 by

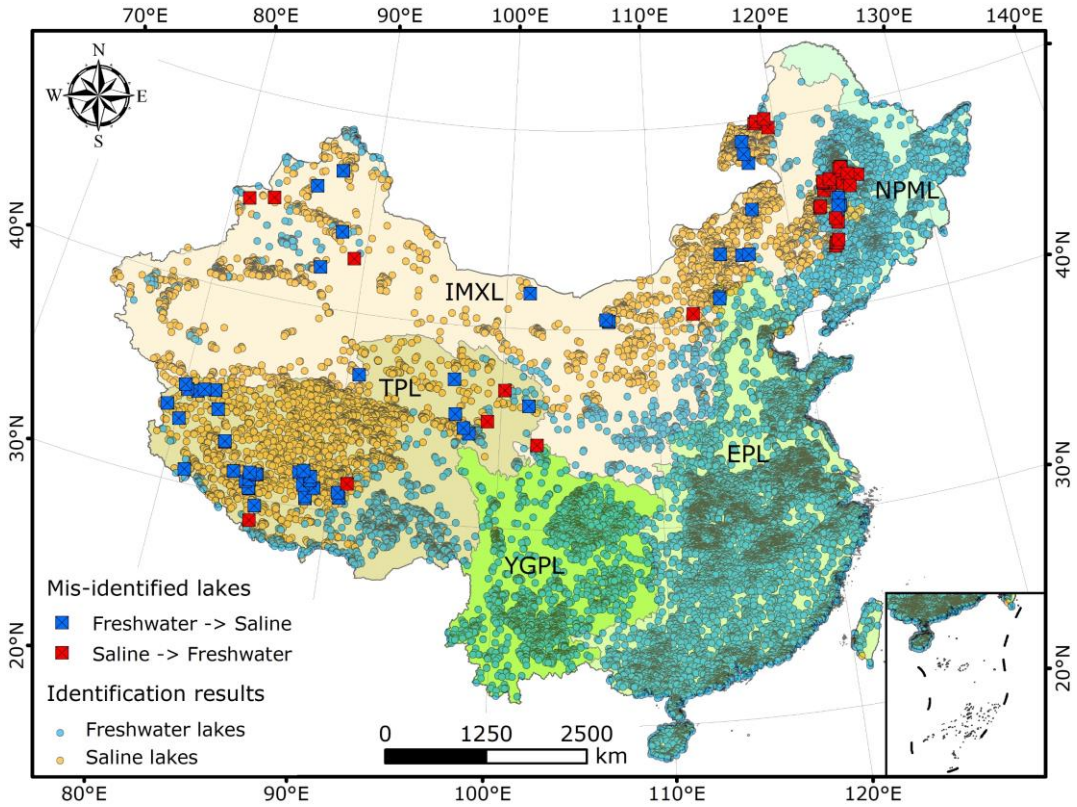
referring to Kuhn and Butman (2021)[7]. Then, the lake basin boundaries were used to calculate the mean values of different basin property variables (Datasets VII-IX in Supplementary Table 1).

Supplementary Note 4. Calculating relative contributions of different factors

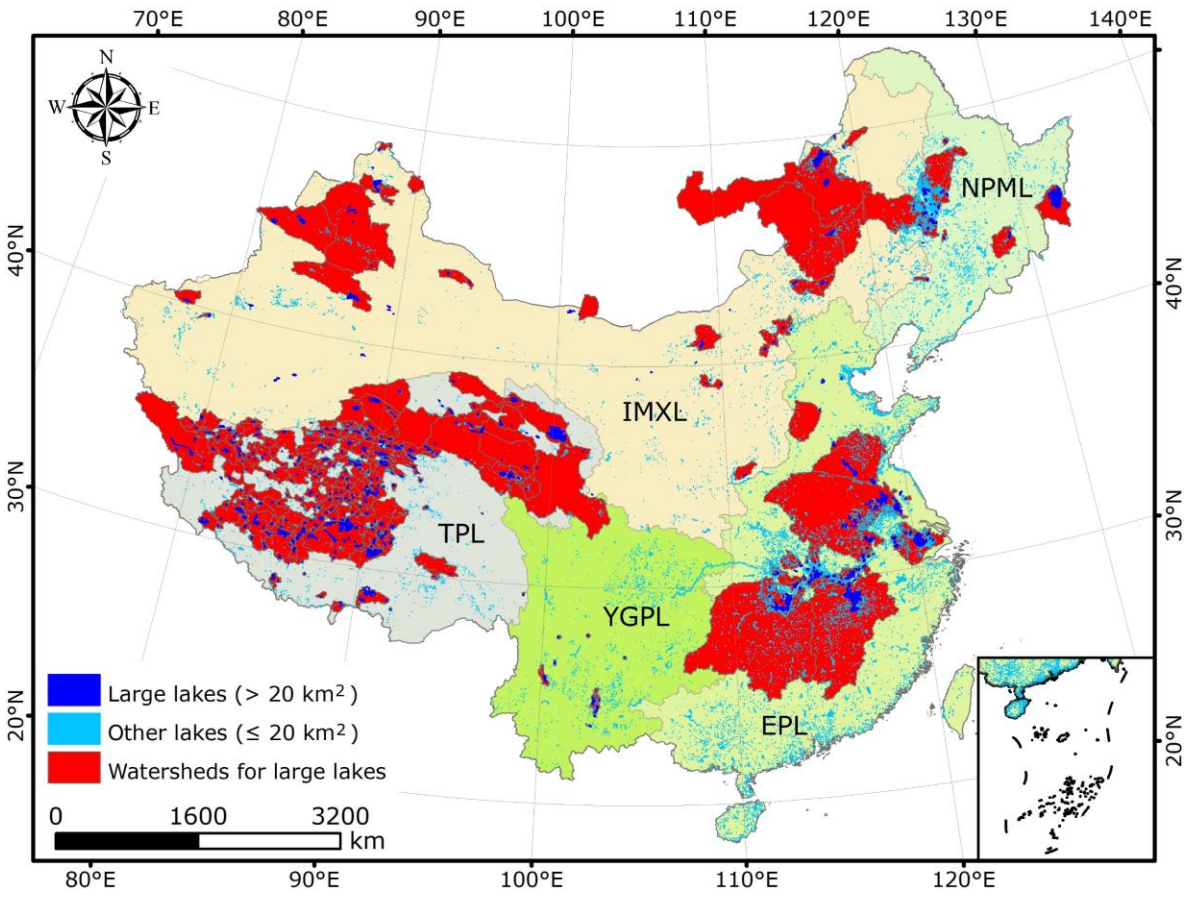
OC concentration is determined by lake features and basin properties[8,9]. Based on nonlinear Random Forest analysis, this study calculated relative contributions of eight lake/basin features to drive the spatiotemporal variations in DOC and POC concentrations of 24,366 Chinese lakes. The considered impact factors included human activities (population density), vegetation coverage (LAI_HighVeg and LAI_LowVeg), water input (evaporation, runoff, and precipitation), wind speed, and air temperature (Temp2m). ① To explain the spatial variations in DOC and POC concentrations, the relative contributions of the eight factors were calculated using the climatologically mean values. ② To understand the temporal variations in DOC and POC concentrations, the relative contributions of the eight factors were calculated using the annual mean values. According to the relative contributions of different factors, we determined the main driving factors which have the largest contributions for different lakes.

In the Random Forest analysis, the Gini, Permutation, and Boruta importance indexes are usually used to evaluate the importance of different features. Each of these three indexes has its advantages and disadvantages. According to Rodríguez-Pérez and Bajorath (2021)[10], we chose the Gini importance index because it is computationally efficient, suitable for large-scale analysis, and easy to understand and explain. For a given factor, the non-linear Random Forest analysis indicated its relative contribution using the Gini importance, which was equivalent to the mean decrease in Gini impurity calculated as the normalized sum of the impurity decrease values for all nodes[10]. The contribution of each factor had a value of 0 – 100%, and the sum of the contributions of the eight factors was 100%.

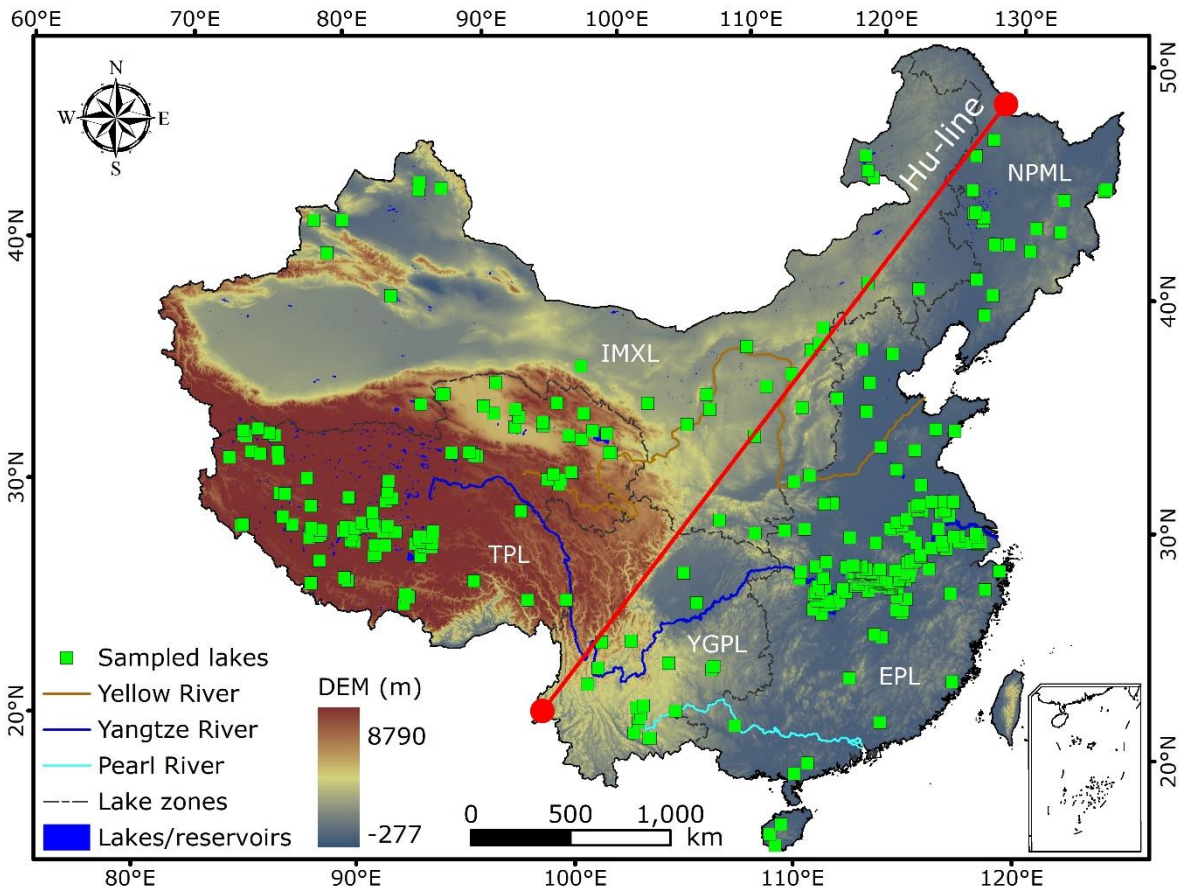
Supplementary figures



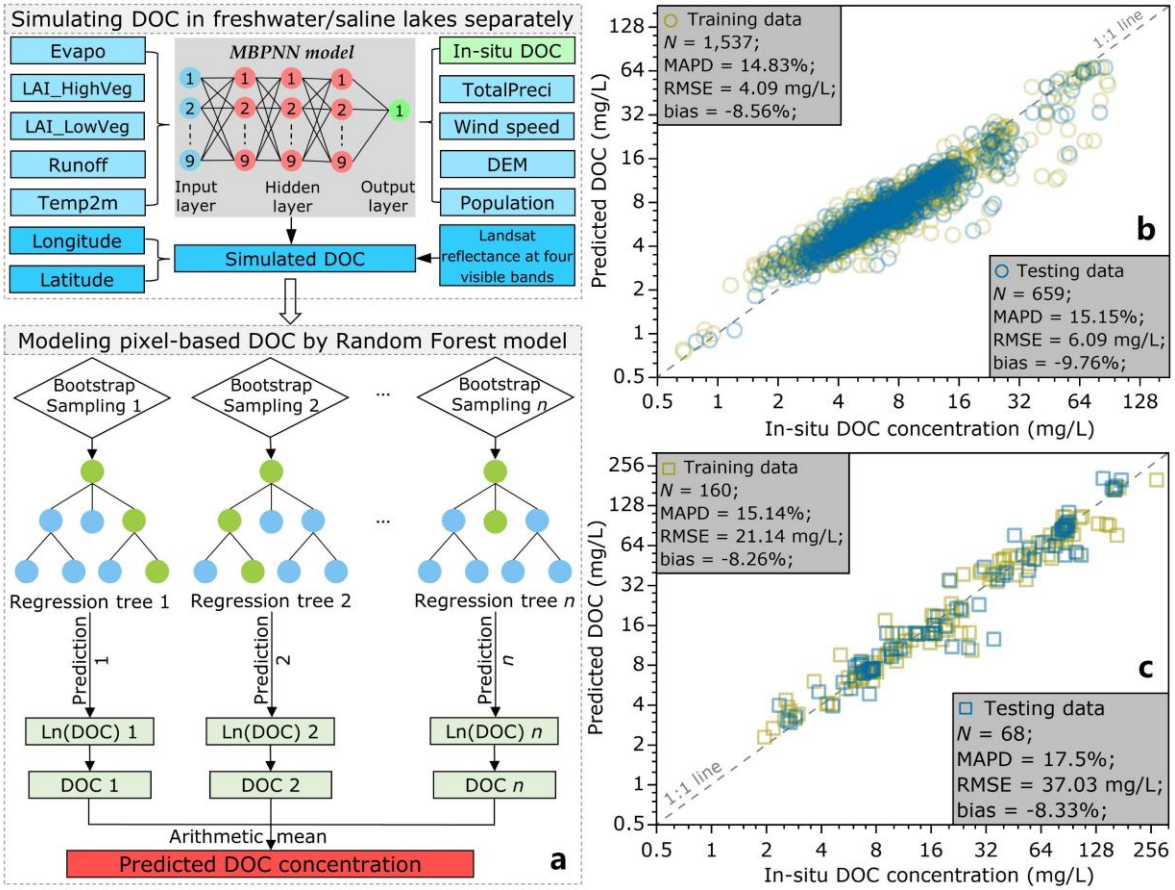
Supplementary Fig. 1. The identification types of 24,366 freshwater or saline lakes in 2015. The mis-identified results for 93 of the 1,550 modeling lakes are also shown.



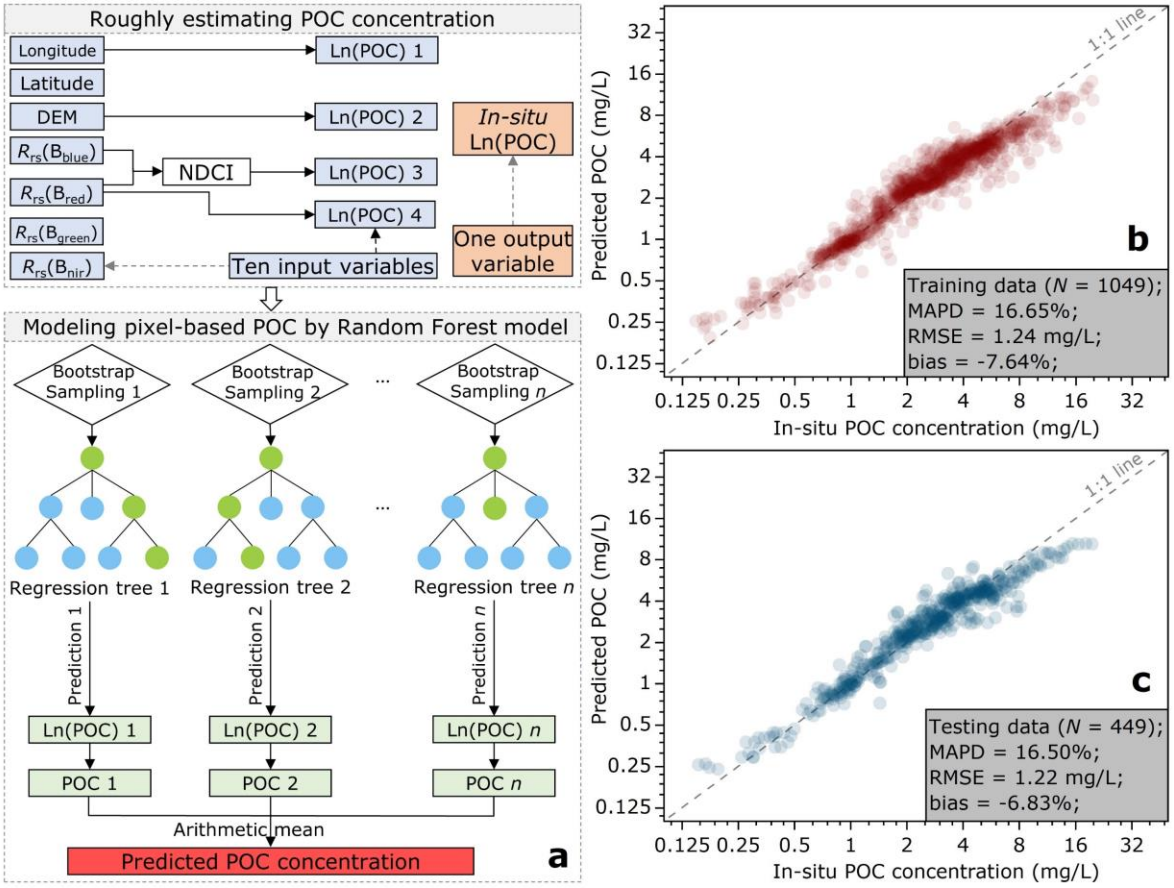
Supplementary Fig. 2. Basin boundaries for the 584 large lakes with areas > 20 km². IMXL: the Inner Mongolia-Xinjiang Lake zone; TPL: the Tibetan Plateau Lake zone; YGPL: the Yunnan-Guizhou Plateau Lake zone; NPML: the Northeast Plain and Mountain Lake zone; EPL: the Eastern Plain Lake zone.



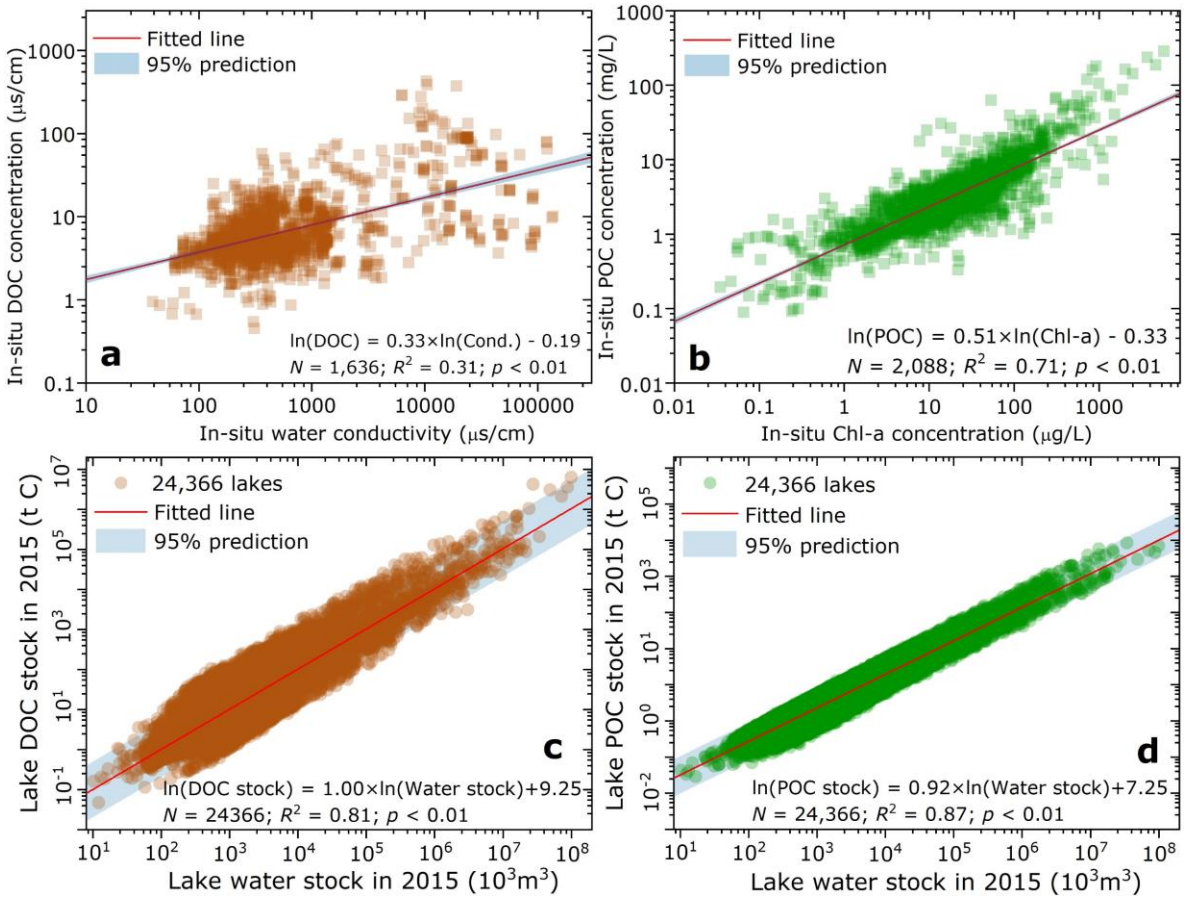
Supplementary Fig. 3. The spatial distribution of the 348 sampled lakes during 2004-2023. Please refer to Dataset I in Supplementary Table 1. The inserted global map was obtained from Google Earth.



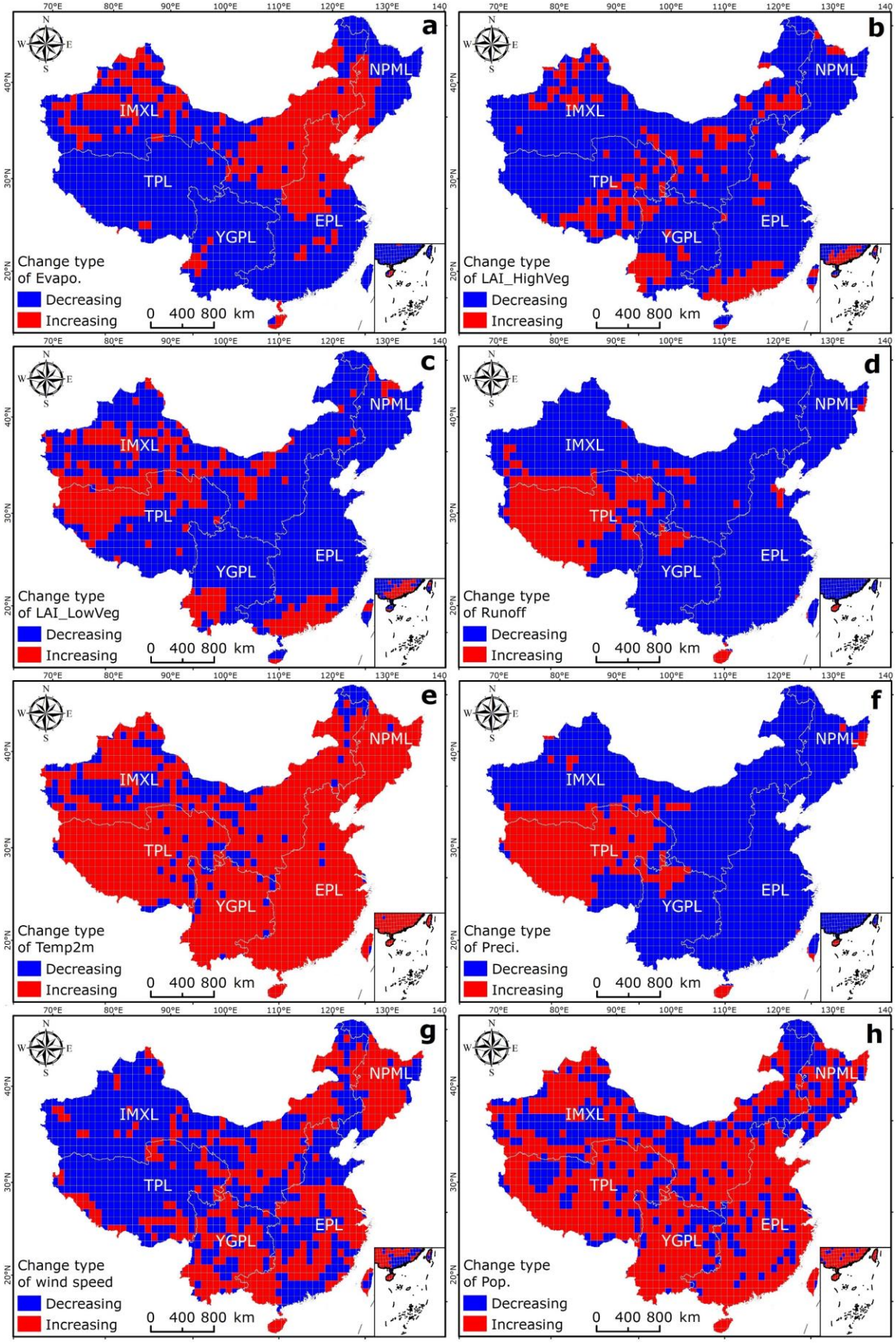
Supplementary Fig. 4. The model building and accuracy evaluation for predicting DOC concentration. **a** The Random Forest models for separately estimating DOC concentrations in freshwater and saline waters. **b** The validation results for freshwater lakes. **c** The validation results for saline lakes.



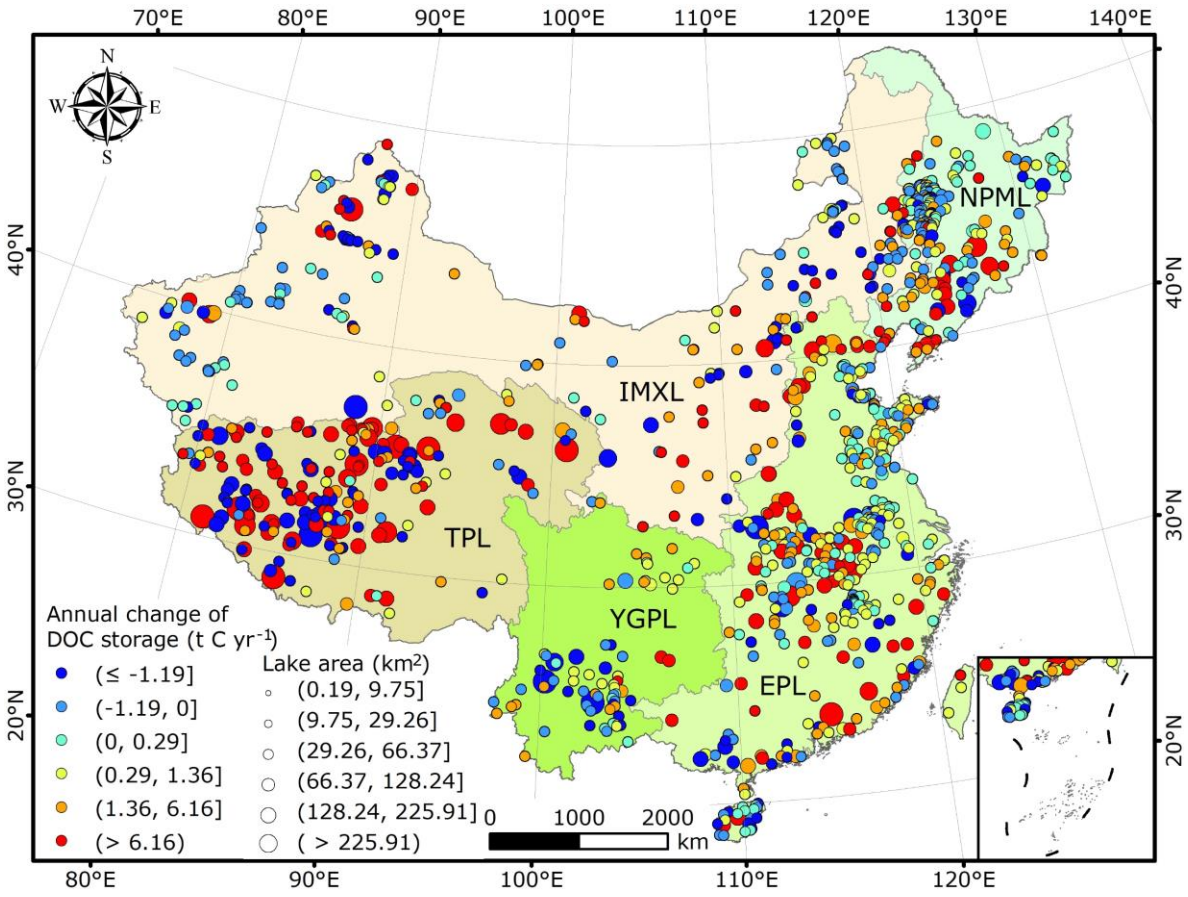
Supplementary Fig. 5. The model building and accuracy evaluation for predicting POC concentration. **a** The two-step Random Forest algorithm for estimating POC concentration. **b** The validation results for the 70% training data. **c** The validation results for the remaining 30% testing data.



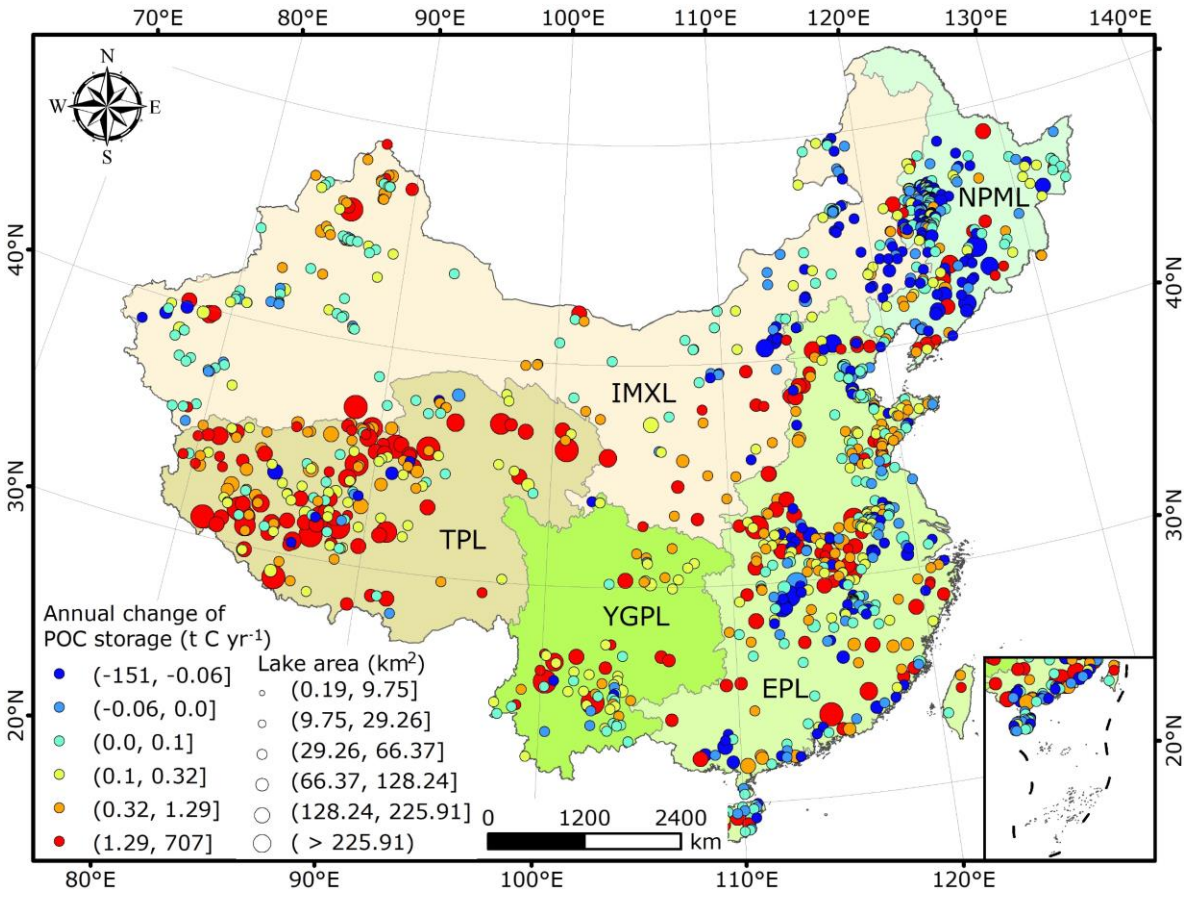
Supplementary Fig. 6. The linear fitting relationships among different variables. **a** between ln(DOC concentration) and ln(water conductivity); **b** between ln(POC concentration) and ln(Chl-a concentration); **c** between ln(DOC stock) and ln(water stock); **d** between ln(POC stock) and ln(water stock).



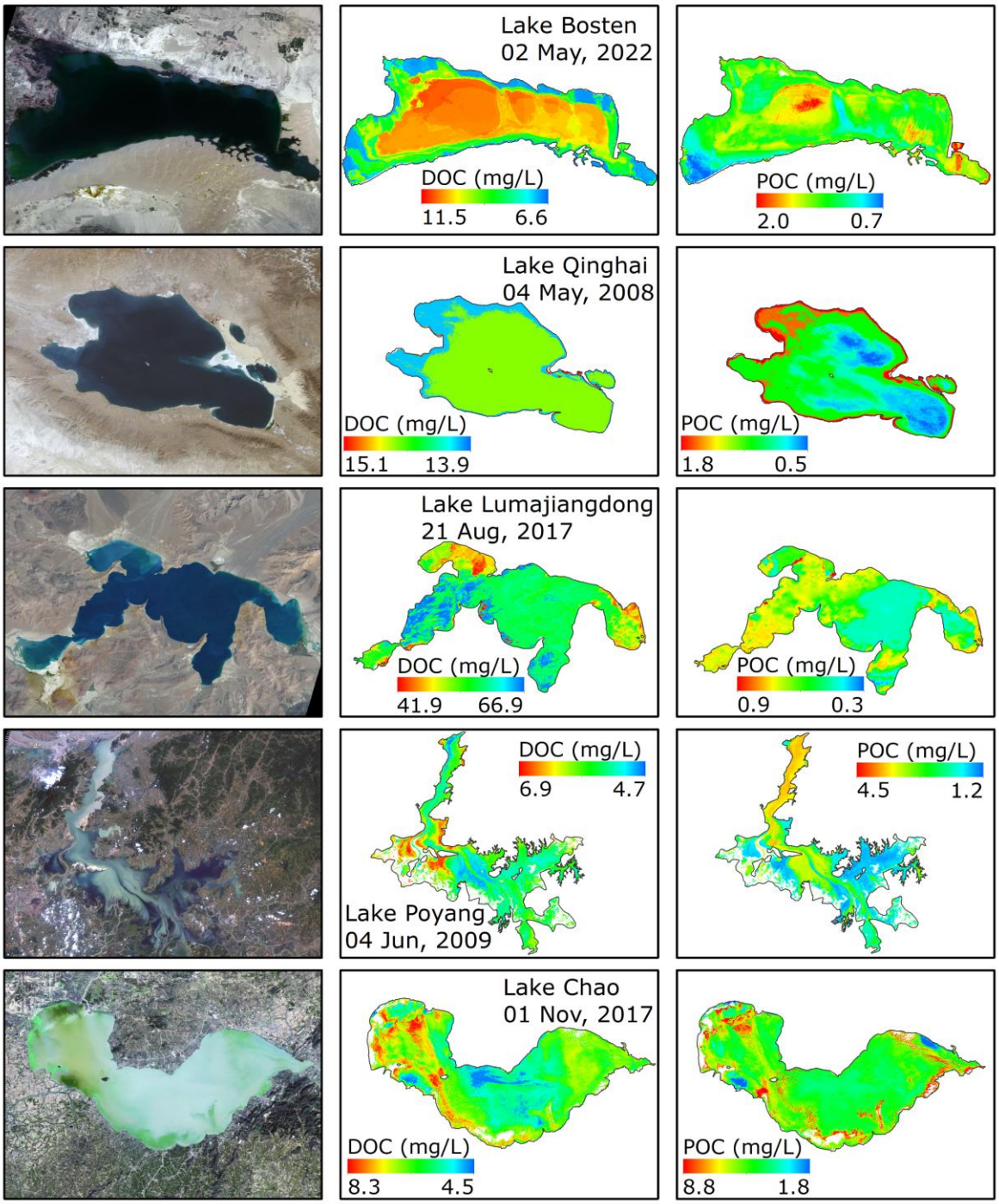
Supplementary Fig. 7. Temporal changes of different impact factors during 1984-2023. **a** Evaporation; **b** leaf area index of high vegetation; **c** leaf area index of low vegetation; **d** runoff; **e** air temperature at 2 m height; **f** precipitation; **g** wind speed; **h** population density.



Supplementary Fig. 8. Annual changes of DOC storage for 1,125 lakes during 1984-2023. Time-series water volume data were sourced from the GloLakes dataset[11].



Supplementary Fig. 9. Annual changes of POC storage for 1,125 lakes during 1984-2023. Time-series water volume data were sourced from the GloLakes dataset[11].



Supplementary Fig. 10. Examples of Landsat-derived DOC and POC concentrations for six typical lakes. The typical lakes are freshwater Lake Bositeng (86.93°E, 41.98°N) by OLI/Landsat-9, saline Lake Qinghai (99.89°E, 36.94°N) with low salinity by TM/Landsat-5, saline Lake Lumajiangdong (81.52°E, 34.04°N) with high salinity by OLI/Landsat-8, freshwater Lake Poyang (116.04°E, 29.17°N) by TM/Landsat-5, and eutrophic Lake Chaohu (117.41°E, 31.55°N) by OLI/Landsat-8.

Supplementary tables

Supplementary Table 1. Basic information about the used data. HydroRIVERS, HydroLAKES, and HydroBASINS datasets were sourced from the HydroSHEDS project (https://www.hydrosheds.org/); GEE: Google Earth Engine (https://earthengine.google.com/); RESDP: Resource and Environment Science Data Platform (https://www.resdc.cn/); ECMWF: European Centre for Medium-Range Weather Forecasts (https://www.ecmwf.int/).

Datasets	Variables	Spatial coverage	Time span	Sources
I	DOC, DIC, POC, Chl-a, TSM, pH, temperature, conductivity	4201 stations of 348 lakes	2004-2023	This study
II	Surface reflectance of Landsat-5/7/8/9 satellites	China	1984-2023	GEE
III	Water storage	1,125 lakes	1984-2023	GloLakes[11]
IV	Lake polygon, water area, and mean water depth	24,366 lakes	2015	HydroLAKES
V	River network	China	2015	HydroRIVERS
VI	River basin	China	2015	HydroBASINS
VII	Population density	China (1,000 m)	1990-2020	RESDP
VIII	DEM	China (30 m)	2000	NASA
IX	Evaporation (Evapo), leaf area index of high vegetation (LAI_HighVeg), leaf area index of low vegetation (LAI_LowVeg), runoff, air temperature at 2 m height (Temp2m), total precipitation (Preci), wind speed	China (1.0°)	1984-2023	ECMWF
X	Organic carbon accumulation rate (OCAR)	115 lakes	After the 1950s	Meta-analysis (Supplementary Data file)

Supplementary Table 2. Hyper-parameters of the Random Forest models for remotely retrieving DOC and POC concentrations. n_estimators: the number of decision trees; max_depth: the maximum depth of decision trees; min_samples_split: the minimum number of samples that can be divided per node; min_samples_leaf: the minimum number of samples contained in leaf nodes. DOC conc.: DOC concentrstion. POC conc.: POC concentration.

No.	Variables	Search ranges	Search step	Hyper-parameters of Random Forest for modeling		
				DOC conc. in freshwater lakes	DOC conc. in saline lakes	POC conc. in lakes
1	n_estimators	[1, 200]	20	181	41	160
2	max_depth	[2, 20]	2	18	14	16
3	min_samples_split	[2, 10]	2	4	4	2
4	min_samples_leaf	[1, 15]	1	2	3	1

Supplementary references

1. Kloiber, S. M., Brezonik, P. L., Olmanson, L. G. & Bauer, M. E. A procedure for regional lake water clarity assessment using Landsat multispectral data. *Remote Sens. Environ.* **82**, 38-47 (2002).
2. Nanjing Institute of Geography and Limnology Chinese Academy of Sciences (NIGLAS). National lake survey report. *Science Press* (2019).
3. Mayer, X., Ruprecht, J. & Bari, M. Stream salinity status and trends in south-west Western Australia. *Department of Environment, Salinity and Land Use Impacts Series*, Report No. SLUI **38**., (2005).
4. O'Callaghan, J. F. & Mark, D. M. The extraction of drainage networks from digital elevation data. *Computer Vision, Graphics, and Image Processing* **28**, 323-344 (1984).
5. Lehner, B. & Grill, G. Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. *Hydrol. Process.* **27**, 2171-2186 (2013).
6. Xie, J., Liu, X., Bai, P. & Liu, C. Rapid watershed delineation using an automatic outlet relocation algorithm. *Water Resour. Res.* **58** (2022).
7. Kuhn, C. & Butman, D. Declining greenness in Arctic-boreal lakes. *P. Natl. Acad. Sci. USA* **118** (2021).
8. Tranvik, L. J. et al. Lakes and reservoirs as regulators of carbon cycling and climate. *Limnol. Oceanogr.* (2009).
9. Liu, D., Du, Y., Yu, S., Luo, J. & Duan, H. Human activities determine quantity and composition of dissolved organic matter in lakes along the Yangtze River. *Water Res.* **168**, 115132 (2020).
10. Rodriguez-Perez, R. & Bajorath, J. Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics. *Sci. Rep.* **11**, 14245 (2021).
11. Hou, J., Van Dijk, A. I. J. M., Renzullo, L. J. & Larraondo, P. R. GloLakes: water storage dynamics for 27 000 lakes globally from 1984 to present derived from satellite altimetry and optical imaging. *Earth Syst. Sci. Data* **16**, 201-218 (2024).