Cite this work:

Cheng Lai, Xizhi Nong, Lihua Chen, Chi Zhang, Luiza C. Campos, Kourosh Behzadian, Ronghui Li, Variability and driving effect of aquatic gross primary productivity across long-distance inter-basin water diversion project, Journal of Cleaner Production, Volume 468, 2024, 143020, ISSN 0959-6526, https://doi.org/10.1016/j.jclepro.2024.143020.

Variability and driving effect of aquatic gross primary productivity across long-distance inter-basin water diversion project

Cheng Lai [a], Xizhi Nong [a, b, *], Lihua Chen [a], Chi Zhang [b], Luiza C. Campos [c, *], Kourosh Behzadian [d], Ronghui Li [a, *]

[a] College of Civil Engineering and Architecture, Guangxi University, Nanning 530004, China

[b] State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan 430072, China

[c] Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E 6BT, UK

[d] School of Computing and Engineering, University of West London, London W5 5RF, UK

*Corresponding author:

Xizhi Nong, E-mail address: nongxizhi@gxu.edu.cn

Luiza C. Campos, E-mail address: l.campos@ucl.ac.uk

Ronghui Li, E-mail address: lironghui@gxu.edu.cn

Abstract

Long-distance water diversion projects significantly affect regions' water resource cycle and allocation. However, many unknowns still exist in water ecosystem functionality and energy flow in large-scale inter-basin water diversion projects. This study focused on the Gross Primary Production (GPP) in the Middle-Route of the South-to-North Water Diversion Project of China (MRSNWDPC), i.e., the world's longest inter-basin water diversion project. The spatiotemporal distribution, driving factors, and pathways of GPP were comprehensively analyzed based on four years of high-frequency water quality monitoring and satellite re-analysis data from 11 national stations, coupling the Bayesian hierarchical models and multivariate statistical methods. The results showed that the daily average GPP in the main canal of the MRSNWDPC over the years was 2.650 g O_2 m^2 d^{-1}, with seasonal peak GPP occurring in summer and generally increasing with the distance along the canal. Five structural equation modeling (SEM) of GPP variations were built in the main canal, revealing the surface pressure (PS) and surface carbon dioxide concentrations (CO_2) and pH value being the main driving factors. The surface pressure showed significant negative impacts on GPP changes in the canal, while the CO_2 and pH showed different direction effects in different sections. The carbon equivalent GPPs in the MRSNWDPC is 0.828 g C m^{-2} d^{-1}, ranging from 0.600 – 1.028 g C m^{-2} d^{-1}, close to the Yangtze River and the East Sea of China. Frequently, hydraulic regulation may impact ecosystem energy flow. This study could provide a scientific basis for a deeper understanding and analysis of the energy flow mechanisms in water ecosystems of mega inter-basin water diversion

projects.

Keywords: Gross primary productivity; Inter-basin water diversion project; Aquatic ecosystem evaluation; Spatiotemporal dynamic analysis; Drivers path analysis.

1 Introduction

Aquatic metabolism is one of the most integrative measurements of aquatic ecosystem functioning and impairment and is highly sensitive to many anthropogenic and natural stressors at different levels of ecological organization (Bunn et al., 1999). Aquatic metabolism is also a crucial part of the global carbon cycle (Val et al., 2016), which is represented by the photosynthetic autotrophic carbon fixation process (i.e., gross primary productivity, GPP), which converts inorganic carbon into organic carbon and its dissipation process (i.e., ecosystem respiration, ER) (Battin et al., 2023; Shen et al., 2015). Among them, GPP is the most significant carbon flux and one of the primary energy inputs in aquatic ecosystems. It plays a vital role in maintaining the aquatic food web balance and the water body's health. GPP plays a significant role in determining the minimum and maximum daily dissolved oxygen levels in riverine water bodies (Genzoli and Hall, 2016; Quinn and McFarlane, 1989) and provides complementary information about the ecosystem's function as water quality only reflects the ecosystem structure (Sabater et al., 2000). Furthermore, GPP is sensitive to the impacts of disturbances or management actions on water bodies, and even minor variations of GPP can significantly affect the carbon balance of aquatic ecosystems (Palmer and Febria, 2012). Therefore, GPP has the potential to become a novel and unique indicator for future engineering water quality regulation. Understanding and quantifying the

67 influencing factors and patterns of GPP is crucial for ecosystem carbon change
68 monitoring and regulation, as well as policy formulation under the climate change
69 background.

70 Previous studies on GPP in aquatic ecosystems have mainly focused on natural
71 water bodies such as rivers, lakes, estuaries, and oceans (Gao et al., 2023; Olson et al.,
72 2020; Spilling et al., 2019; Zhang and Ye, 2021). However, with the increasing demands
73 for water allocation and utilization in human society, one of the most efficient hydro-
74 projects for water pressure mitigation, long-distance water diversion projects, have
75 been widely built in recent years and have become a complex artificial water system
76 with water-air interfaces, and achieved significant economic and social benefits (Duan
77 et al., 2022; Rodriguez-Castillo et al., 2019; Woodford et al., 2013). Those projects are
78 widely featured by concrete-lined canals, simple hydro-ecological structures, and
79 frequent hydrodynamic condition variations, which are different from natural rivers'
80 seasonal variations in water level and flow velocity. Thus, gaining a systematic
81 understanding of the variance mechanism of GPP in a long-distance water diversion
82 canal is not only beneficial to mitigating the potential ecological risks but also provides
83 insights into fundamental ecosystem functions, thereby facilitating the forecasting and
84 regulation of ecosystem development. However, the metabolic rate of canal ecosystems
85 is higher than natural water bodies under high water levels and significant flow rate
86 regulations, and the environmental factors and processes affecting the ecosystem
87 become more complex than natural water bodies (Aristi et al., 2014; Prichard and Scott,
88 2014). Furthermore, the GPP of small-medium streams and rivers is reasonably well

known (Hoellein et al., 2013; Yu et al., 2018), but estimates for large rivers or canals are much rarer. Therefore, it is unclear if and how GPP patterns change from small rivers to truly large and long canals, which we define as long canals as canal length is greater than 1000 km. The GPP regimes of such canals are virtually unknown.

The primary determinants of GPP activity in rivers are light, temperature, and hydrologic disturbance, with nutrients and organic matter that may accelerate the GPP response to each of these factors (Appling et al., 2018b). These drivers vary naturally and can show strong annual and seasonal patterns (i.e., light, thermal, and hydrologic regimes), which greatly differ canal continuum. Additionally, these drivers respond to global changes in climate and greenhouse gas. Long-distance inter-basin water diversion projects experience diverse climates, geology, and high taxonomic diversity of organisms. For example, the water quality and hydrology conditions are significantly different upstream and downstream (Bernhardt et al., 2018). Furthermore, GPP is sensitive to multiple stressors, which act independently or in combination with other stressors, often presenting a complex interplay of controls (Heathwaite, 2010; Nong et al., 2020). Ultimately, these various changes in environmental controls modify the rates and timing of ecosystem metabolism with potentially detrimental consequences for water quality and biological communities. The patterns of ecological function changes in channels under artificial regulation remain unknown. Additionally, the same water body exhibits spatial heterogeneity under different regional conditions, influenced by geographical and natural environmental factors. This is a challenge faced when studying large-scale or medium-scale water bodies, and therefore, previous research

findings can only be cautiously referenced and cannot be directly applied. Understanding the GPP dynamic processes and their drivers in inter-basin water diversion projects is of importance to provide insights and guidance in setting artificial hydrologic regulations under the impacts of climate change.

Considering the research gaps mentioned above, this study took the Middle Route of the South-to-North Water Diversion Project of China (MRSNWDPC), i.e., the longest inter-basin water diversion project in the world, as the study case. As a novel artificial water system which officially operated since 2014, the ecosystem of MRSNWDPC has yet to achieve the ecological balance. That is to say, the MRSNWDP is vulnerable to ecological anomalies and susceptible to external interference (von Schiller et al., 2017). Therefore, this presents a suitable and pressing case for investigating the mechanisms of GPP for the water management department. This research aims to investigate the spatiotemporal variability of GPP and its driving effect in long-distance water diversion projects. The estimation of GPPs was conducted with variants of single-station models based on an open-channel metabolism approach. The relationships between environmental factors and GPP inter and inner locations were explored using the random forest method. The structural equation model (SEM) was applied to analyze the driving factors and pathways of GPP variations. The main objectives of this study were: 1) to estimate the spatiotemporal variations of GPPs in long-distance inter-basin water diversion projects; 2) to identify key driving factors for GPP changes in the long-distance water diversion canal; 3) to understand the effect pathways of GPP variations. This study could contribute to a better understanding of

the ecosystem characteristics of such mega inter-basin water diversion projects, enrich and complete our understanding of GPP variability characteristics and regulation patterns under different hydrological regimes and geographical environmental features, provide scientific references for the functional and environmental assessments, and support the development of water quality management strategies.

2 Material and methodology

2.1 Study area and data collection

The Middle Route of the South-to-North Water Diversion Project of China originates from the Danjiangkou Reservoir, China. It flows 1,276 km northward via an open canal crossing four provinces and municipalities to the Tuancheng Lake, Beijing, the capital of China. The main canal spans the subtropical and temperate monsoon climate zones, with the average yearly air temperature between 11.2 and 15.8 ℃ and the average annual rainfall ranging from 495.6 to 795.4 mm (from 1981 to 2022). The MRSNWDPC has delivered over 60 billion m^3 of fresh water to North China since December 2014, contributing as the major drinking water resource for more than 75% of the cities along the canal. The Construction and Administration Bureau of the MRSNWDPC has built 11 automatic water quality monitoring stations along the main canal based on the national environmental protection program to monitor the water quality regimes. The stations are Taocha (TC), Jianggou (JG), Liuwan (LW), Fuchengnan (FCN), Zhanghebei (ZHB), Nandaguo (NDG), Tianzhuang (TZ), Xiheishan (XHS), Fenzhuanghe (FZH), Zhongyishui (ZYS), and Huinanzhuang (HNZ), as shown in **Fig. 1**. Three canal sections, i.e., upstream, midstream, and downstream, was defined based on the distance between each station and the starting point of the canal. More details can be found in **Table S1**.

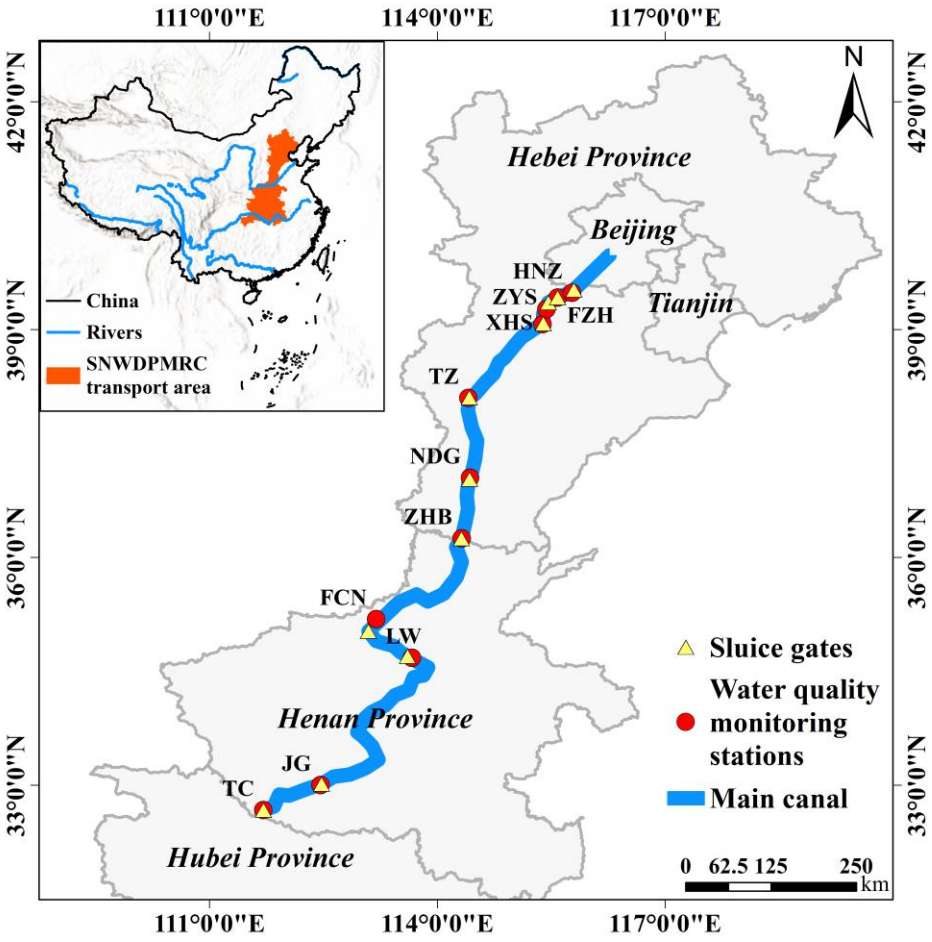


Fig. 1. Locations of the water quality monitoring stations along the main canal of the Middle Route of the South-to-North Water Diversion Project of China in this study (Note: TC to FCN are "upstream", ZHB to TZ are "midstream", and XHS to HNZ are "downstream").

A series of environmental factors influence GPP, and we collect and divide the factors set with two datasets in the current study based on the usage of factors, which are estimation dataset (including dissolved oxygen (DO), saturated DO (SDO), water temperature (WT), photosynthetically active radiation (PAR), and water depth (WD)) and analysis dataset (including pH, surface pressure (PS), wind speed (WS), precipitation (Pre), carbon dioxide (CO_2)). Factors in the estimation dataset were all selected based on the requirements of the streamMetabolizer approach to calculate GPP

levels in the main canal

. In the analysis dataset, pH represents the acidity or alkalinity of water, which affects phytoplankton photosynthesis by influencing the equilibrium of the carbonate system and controlling the partial pressure of carbon dioxide (Appling et al., 2018b; Jakobsen et al., 2015). PS, WS, and Pre are common meteorological factors widely adopted in GPP research. PS has a significant influence on SDO; WS is one of the most important determinants of the gas exchange coefficient (K600) (Antonopoulos and Gianniou, 2003; Jia et al., 2020b); and Pre may cause high flow event which inputs inorganic nutrients, dissolved organic carbon and suspended sediments, which can induce both positive and negative effects on primary production (Tang et al., 2015). CO_2 is one of the greenhouse gas which produced in the metabolism process of aquatic organisms. The relationship between greenhouse gases and oxygen was previously interpreted as an important metabolic role in dynamic gas production (e.g., respiration, methanogenesis, methane oxidation) and can reflect the oxygen concentration in surface water (Shen et al., 2015).

The water quality parameters were monitored from the automatic national monitoring stations along the main canal, including dissolved oxygen (DO, mg/L), water temperature (WT, ℃), and pH values, from January 2017 to December 2020, with a monitoring frequency of every six hours. The water depth (WD, m) data were recorded simultaneously in the nearest regulating sluice to each water quality monitoring station. Meteorological and environmental indicators included surface pressure (PS, hPa), shortwave radiation (SW, J/m^2), wind speed (WS, m/s), and

186 precipitation (Pre, mm/day). Of which, the PS, WS, and Pre data were obtained from
187 ERA5 (ECMWF Re-Analysis 5) hourly data (Campeau and Del Giorgio, 2014), while
188 SW data were obtained from CERES satellite re-analysis hourly data (Muñoz Sabater,
189 2019). Greenhouse gas, i.e., the surface carbon dioxide concentrations (CO_2, mg/kg)
190 was used from the Copernicus Atmosphere Monitoring Service (CAMS) and global
191 greenhouse gas re-analysis (EGG4) dataset, provided as hourly data
192 (Nasa/Larc/Sd/Asdc, 2017).

193 **2.2 Data processing**

194 High-quality data is essential for reliable data analysis (Inness, 2019). This study
195 pre-processed the collected data by data management, cleaning, and denoising methods.
196 Different intelligent techniques were applied to eliminate and reduce outliers and data
197 noise as follows:

198 (1) Data pre-processing, including removing and replacing significant outliers in
199 the raw data. The outliers were detected and removed using quartile and moving median
200 methods, and the missing data were imputed using linear interpolation. Additionally, to
201 make the distribution of rainfall conform to the normal distribution, a logarithm
202 transformation $Pre^* = \log_{10}\left(0.1 + \sqrt{Pre}\right)$ was applied for the raw precipitation data,
203 where Pre^* and Pre represent the transformed and raw data, respectively (Ehrlinger
204 and Woess, 2022). Finally, all the indicators in the estimation dataset were resampled
205 to a frequency of every six hours using the time series resampling method, generating
206 processed data after the basic data cleaning step.

207 (2) Deep denoising. In this study, the Complete Ensemble Empirical Mode

Decomposition with Adaptive Noise (CEEMDAN), Sample Entropy (SamEn), and Density-based Spatial Clustering of Applications with Noise (DBSCAN) were applied for deep denoising of the cleaned data. Detailed information about CEEMDAN, SamEn, and DBSCAN can be found in sections S1, S2, and S3, respectively. The steps were as follows: 1) the original sequences were decomposed into multiple Intrinsic Mode Functions (IMFs) and a residual term using CEEMDAN; 2) the SamEn method was employed to divide the IMF signals from complex to simple, categorizing them into high-frequency, mid-frequency, and low-frequency signals. These signals were mapped to the feature space of high-frequency, mid-frequency, and low-frequency characteristics, and then DBSCAN was applied to cluster the 3D feature space data. In the clustering results, clusters with a small number of data points less than 50% of the total samples, and with a variance more significant than the variance of the original sequence were defined as noise clusters; 3) the data was processed to obtain the final denoised data by filtering out the noise. The CEEMDAN was implemented using the MATLAB toolbox provided by Torres (Feng et al., 2020), with the parameters set as follows: the noise standard deviation, the number of realizations, and the maximum sifting iterations were set to 0.2, 700, and 5000, respectively. We developed a program based on the principle of sample entropy, with reconstruction dimension and threshold set as program parameters, and classified the complexity of IMF signals using 2 and 0.15*S.D. (·), respectively. "(·)" stands for the inputted signal in SamEn, i.e., the IMF from CEEMDAN. DBSCAN was implemented using relevant functions provided by MATLAB, with minPts set to 10, and eps calculated as the mode of the Euclidean

distances between all data points. All data processing procedures were performed on the MATLAB R2021a platform (Torres et al., 2011).

Finally, the processed data were divided into an estimation dataset (including DO, Saturated DO (SDO), WT, photosynthetically active radiation (PAR), and WD) and an analysis dataset (including pH, PS, WS, Pre, CO_2), respectively.

2.3 GPP estimation method

In this study, the GPP based on measurements from individual water sampling stations was calculated using the single-station method. We used data such as dissolved oxygen, water temperature, photosynthetically active radiation, and water depth to calculate GPP using the streamMetabolizer program in R language. The core equation in the program for the DO variations at each time step was based on **eq.(1)** as follows (The MathWorks, 2022):

$$\frac{dO_{i,d}}{dt} = \left(\frac{GPP_d}{\overline{z_{i,d}}} \times \frac{PPFD_{i,d}}{\overline{PPFD_d}}\right) + \left(\frac{ER_d}{\overline{z_{i,d}}}\right) + f_{i,d}(K600_d)\left(Osat_{i,d} - O_{i,d}\right) \quad (1)$$

$$f_{i,d}(K600_d) = K_{600} \times \left(\frac{s_A + s_B T_t + s_C T_t^2 + s_D T_t^3}{600}\right)^{s_E} \quad (2)$$

Where $O_{i,d}$ is the modeled oxygen concentration on day d at time index i, and $dO_{i,d}/dt$ is the rate of concentration change. GPP_d, ER_d, and $K600_d$ are the three daily parameters fitted by the model: GPP_d and ER_d are daily average rates of gross primary productivity and ecosystem respiration, respectively (g O_2 m^{-2} d^{-1}), while $K600_d$ is a daily average value of the standardized gas exchange rate coefficient (d^{-1}, scaled to a Schmidt number of 600). The other variables are model inputs: $\overline{z_{i,d}}$ is the average water depth (m) over the width and length of the upstream; $PPFD_{i,d}$ is the

photosynthetic photon flux density (μmol photons m^{-2} d^{-1}); $\overline{PPFD_d}$ is the daily average observed $PPFD_{i,d}$; $f_{i,d}(K600_d)$ is a function that converts daily average $K600_d$ to an O_2-specific, temperature-specific gas exchange coefficient ($KO2_{i,d}$, d^{-1}) based on eq. (2) with T_t is the water temperature in °C, the Schmidt number coefficients are s_A = 1568, s_B = −86.04, s_C = 2.142, and s_D = −0.0216, and the scaling exponent s_E = −0.5 (Raymond et al., 2012); $Osat_{i,d}$ is the theoretical saturation concentration of O_2 if the water and air were in equilibrium. More detailed information can be found in (Jähne et al., 1987).

The streamMetabolizer can estimate GPP based on the Bayes hierarchical model, Monte Carlo Markov Chain (MCMC) method, and Euler's differential equations. The model outputs posterior probability distributions for daily GPP and K600. The reliability of the results can be assessed by examining the convergence of the model outputs, the fit between observed and predicted dissolved oxygen values, model errors, and the daily variation in GPP (Appling et al., 2018a). This study used a model called "b_np_oi_eu_psrckm.stan". Each Bayesian metabolism model ran on 4 MCMC chains, including 500 burn-in steps and 2000 save steps. The convergence of the daily estimated data was assessed using the Gelman-Rubin $\hat{R}$ criterion, with values below 1.1 indicating convergence. Data with $\hat{R}$ values higher than 1.1 were removed, which stands for whether there is still similar variation within or between chains after discarding the Warmup samples (Brooks and Gelman, 1998). The variation in GPP was modeled as a saturating function of light, as this setting typically provides robust and accurate results (Gelman and Rubin, 1992). The SDO and PAR were calculated using

the "calc_do_sat" and "calc_light" functions provided by the package, respectively. The average water depth of the neighboring control structures was used as the mean depth of the river. In the output results, GPP was constrained to be always positive, and any data points that did not meet this criterion were removed and interpolated. Additionally, a moving window with a width of 15 was used to apply a moving average filter to the GPP output, reducing noise and estimation errors. The data pre-processing and GPP prediction framework can be seen in **Fig. 2**.

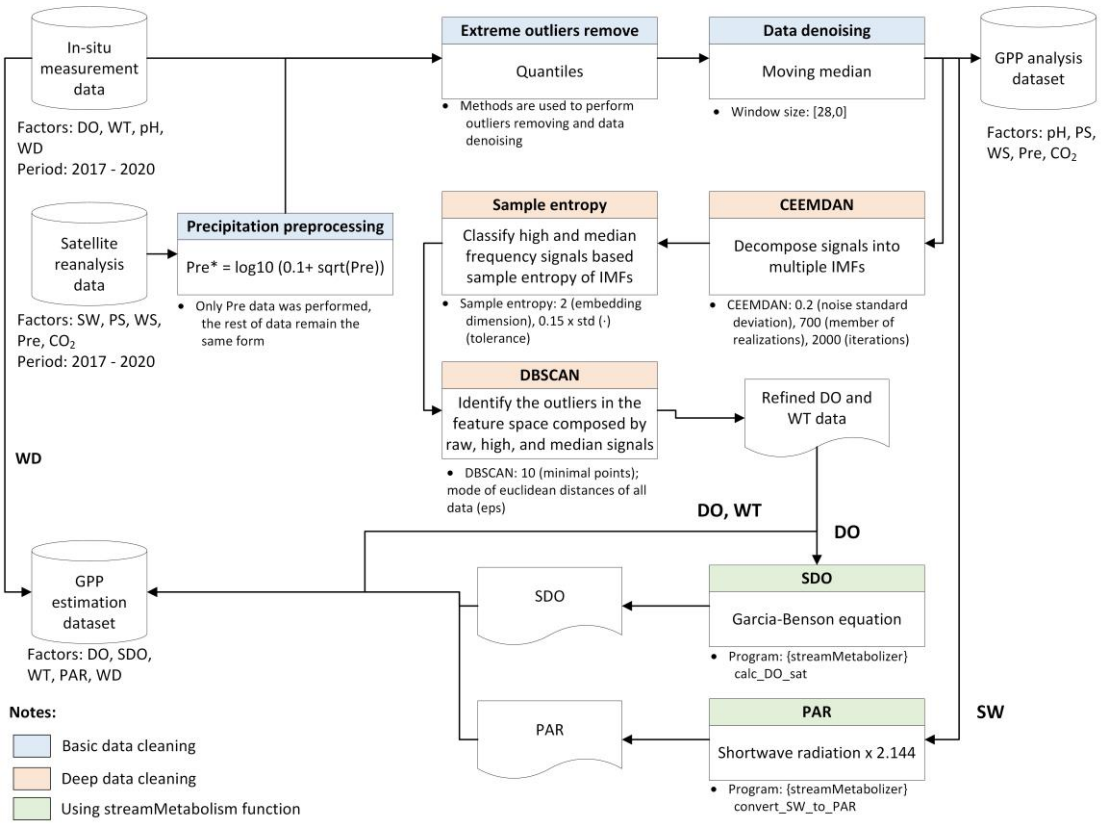


Fig. 2. In-situ monitoring and satellite re-analysis data processing scheme diagram in this study (Note: In basic data cleaning, the first and second numbers of "window sizes" stand for the backward and forward window sizes respectively; In deep data cleaning, numbers, and bracketed words are values and names of the algorithm parameters, "(·)" imply the input signal of sample entropy; In

using streamMetabolizer function, enclosed content and subsequent text are the name of the R package and specific function respectively).

2.4 Statistical approaches

2.4.1 Random Forest

In this study, random forest regression and variable importance methods were applied to identify the driving factors of GPPs. Random forest is a supervised ensemble learning algorithm widely used for classification and regression problems (Hall et al., 2015). It trains multiple weak learners by repeatedly bootstrap sampling and combines their output using voting (for classification problems) or averaging (for regression problems) to generate robust and accurate predictions. In this study, random forest regression was applied to identify the key drivers of GPP in analysis dataset, and to assess the explanatory ability of environmental factors on GPP variations. Additionally, the relative importance of factors in the analysis dataset was evaluated based on the "percent increase in mean square error (%IncMSE)" and the significance *P* values. The "%IncMSE" calculates the MSE increase when a feature is replaced. A higher %IncMSE indicates a feature has higher explanatory capacity on the dependent variable (Breiman, 2001). This calculation used the R package "randomForest" for random forest regression with 500 trees, and the "rfPermute" R package was used to calculate the "%IncMSE" and *P* values.

2.4.2 Structural Equation Model

The casual relationships and interactions between GPP and the physicochemical factors in different sections were identified through the structural equation modeling

(SEM) in this study. SEM is widely used in environmental and ecology science, which can provide a causal analysis framework to explore the relationship network and direct and indirect driving path among multivariate empirical data and theoretical models (Burpee et al., 2022; Song et al., 2021). Each introduced causal path in the model structure is quantified using a univariate regression model to assess if it is supported by empirical data. Furthermore, direct and indirect pathways can be combined to calculate the variables' total effect on the model's response by simple addition and multiplication of standardized path coefficients in the individual regression models (Bai and Cotrufo, 2022).

This study employed the SEM to analyze the causal pathways between environmental factors and GPP variations. To avoid the mixture of causal path patterns of different monitoring locations, the main canal was further divided into five segments based on the spatial distances inter-station in Section 2.1, i.e., TC to LW, LW to ZHB, ZHB to TZ, TZ to ZYS, and ZYS to HNZ, respectively. Path coefficients (partial multiple regression coefficients) were standardized to a common metric, allowing for the comparison of the relative importance of each effect. Greater effect values indicate a stronger influence on GPP. Additionally, four model goodness-of-fit evaluation indicators, including chi-square (χ^2), goodness of fit index (GFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR), were selected to assess the fit of the SEMs. The feasibility of SEM models was tested by p-value, and $p > 0.05$ indicated a reliable simulation of SEM. Notably, the dataset that has a large sample size could yield a model with $p > 0.05$, even though the model

is well-fitted. Therefore, the evaluation of path models should consider different evaluation metrics. The ranges and recommended values of the above indicators can be found in **Table S2**. This study applied the maximum likelihood estimation method to the SEM estimation and all response variables were previously standardized in order to eliminate the effect of unit. The calculation and evaluation of the SEM were conducted with the "lavaan" R package (Palt et al., 2022).

3 Results

3.1 Spatial and temporal distribution of environment parameters

The average DO concentrations and pH of the main canal ranged from 9.10 to 11.02 mg/L and 7.90 to 8.18, respectively (**Fig. 3a and 3c**). The DO, SDO, and pH values were increased along the canal with maximum and minimum average DO at XHS at TC station, respectively. The PS and WS varied across stations, whereas the PAR exhibited minor spatial differences. CO_2 was mostly consistent across stations, with sudden increases noted at specific locations. WD falls along the canal from 8.15 m to 3.86 m (Table S1). The temporal variations of each parameter in the main canal are shown in **Fig. S2**, and most of the indicators showed a significant seasonal variations pattern from 2017 to 2020.

The results of the Spearman correlation coefficients and the Mantel test showed the relationships among different factors in the entire canal (Fig. 3b). The highest correlations were the PS and CO_2 in the main canal. However, the correlation between indicators varied among different canal sections. For instance, the pH values and CO_2 showed no significance with PAR in the upstream, whereas they exhibited a significant

correlation in the midstream and downstream (Fig. S3). To sum up, different indicators showed complex spatial variation patterns in interstation, indicating the complex impacts on GPP and energy flux variations in the ecosystem of the main canal.

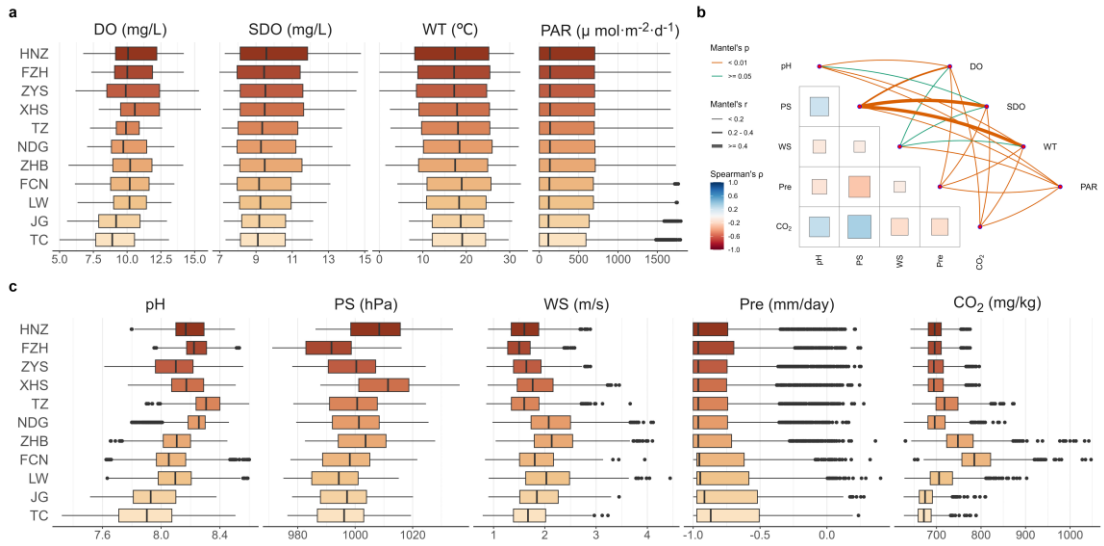


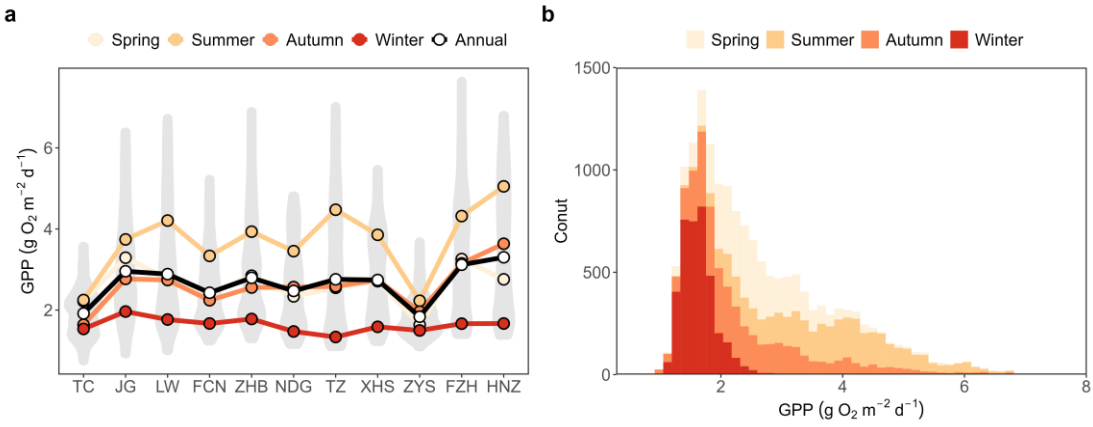
Fig. 3. Statistical summary of environmental variables in the main canal of MRSNWDPC (Note: (a) statistical summary of the GPP estimation dataset; (b) Spearman correlation matrix of the GPP analysis dataset and the Mantel test between the GPP estimation dataset and analysis dataset; (c) statistical summary of the GPP analysis dataset; Detailed information about the variables' monthly variance, section-scale mantel tests, and stations with WD data included can be found in Fig. S2, Fig. S3, and Table S1).

3.2 Spatial and temporal distribution of GPP

The spatial distributions and seasonal variations of GPPs at different stations can be seen in Fig. 4. The average daily GPP in the MRSNWDPC is 2.650 g O_2 m^{-2} d^{-1}. Of which, the highest (3.289 g O_2 m^{-2} d^{-1}) and lowest (1.910 g O_2 m^{-2} d^{-1}) GPP were observed at the HNZ and TC stations, respectively. Additionally, the highest seasonal GPP was observed in summer (3.713 g O_2 m^{-2} d^{-1}), followed by spring (2.614 g O_2 m^{-2}

365 d^{-1}), autumn (2.598 g O_2 m^{-2} d^{-1}), and winter (1.630 g O_2 m^{-2} d^{-1}). The histograms showed
366 that the GPPs have the narrowest variation range in winter (range from ~1.0 to ~2.5 g
367 O_2 m^{-2} d^{-1}) and have the most significant variation ranges in summer and autumn.
368 However, the peak modes of the GPP in each season showed surprising consistency
369 (~1.8 g O_2 m^{-2} d^{-1}).

370 Temporal variations of GPP in the main canal and different canal sections can be
371 found in **Fig. 5**. The monthly average GPPs usually rose gradually at the beginning of
372 each year, reaching their peak in June or July and then decreasing. Based on the
373 interannual maximum GPP variations, the GPP in the main canal has been increasing
374 over the years, and the midstream showed more significant change (range roughly from
375 3 to 5 g O_2 m^{-2} d^{-1}) compared to the upstream and downstream. Additionally, the Mann-
376 Kendall trend test results (**Table S3**) showed all the canal sections of the MRSNWDPC
377 have significant GPP-increasing trends.



378

379 **Fig. 4.** Spatiotemporal variations of the GPP in the main canal of MRSNWDPC from 2017 to 2020
380 (Note: (a) The GPP distributions in different stations and periods based on the violin plots; (b) The
381 mode distribution of GPPs in different seasons).

382

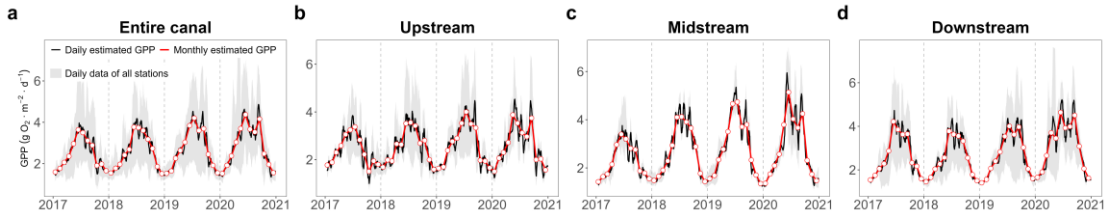


Fig. 5. Time series of GPP estimation in the entire canal and different canal sections from 2017 to 2020 (Note: the "black line" represents daily average GPPs, the "red dotted" line represents monthly average GPPs, and the "shaded area" represents the intraday daily GPP variation range in all stations).

3.3 Relationships between GPP and environmental driving effect

The RF and Spearman correlation are both used to analyze the relationship between environmental factors and GPP at the station scale, whose results can be found in **Fig. 6**. For most stations, the explained variances exceeded 60%, except for the JG station at 55.88% and the ZYS station at 39.32%. The PS showed the highest explanatory capacity of GPP variation in most stations, followed by the pH values and CO_2. The Spearman's correlation exhibits significant spatial difference patterns between different environmental factors and GPP. For instance, the pH-GPP pairs may present negative or positive correlations in different stations, while the PS-GPP pair kept strong negative relationships in all stations.

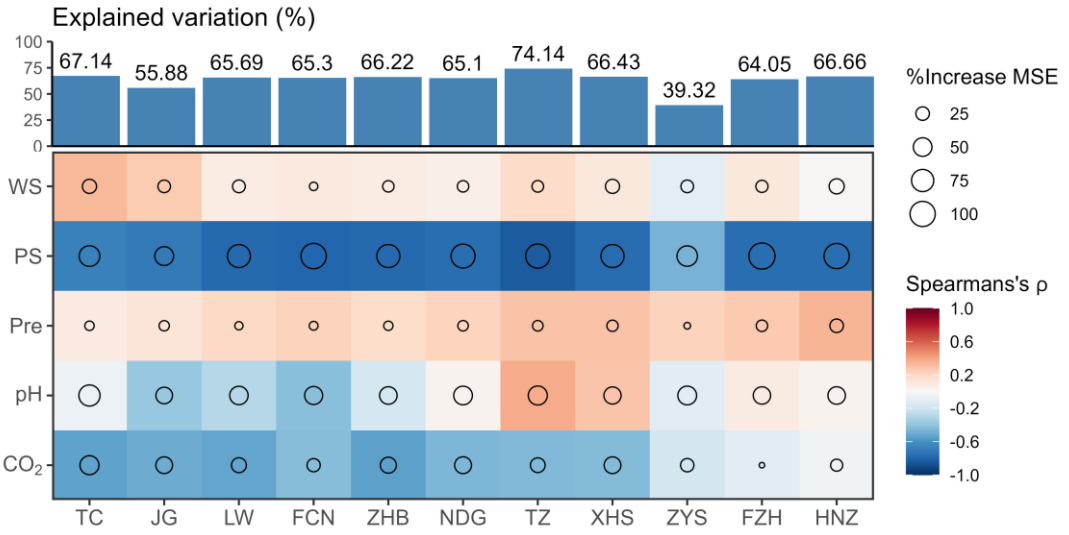


Fig. 6. Driving factors analysis of GPP variation in the MRSNWDPC based on the Spearman correlation matrix and random forest-based explained variances (Note: different rectangles with circles represent the combinations of Spearman's ρ and "percent increase in mean square error (%IncMSE)" based on the random forest regression predictions between GPP and the corresponding environmental factors).

The effect pathways of different factors on GPP variations were explored by the SEM and shown in **Fig. 7**. WS has both significant direct and indirect effects on GPP variations upstream (**Fig. 7a, b**). However, those pathways showed slightly or no significant relationships in midstream and downstream. The CO_2 and PS were the most important environmental factors of GPP variation with high and significant factor loadings in both direct and indirect effect pathways. PS showed negative direct effects on GPP variations with factor loadings ranging from -0.40 to -0.71, whereas CO_2 showed significant effects on GPP changes with spatial differences from 0.06 to -0.29. CO_2 shows little impact on GPP in TC to LW (**Fig. 7a**), ZHB to TZ (**Fig. 7c**), and ZYS to WHH section (**Fig. 7e**), and the rest of the SEMs showed significant positive effects on GPP. From the explanatory capacities of the SEMs, the models at the upstream (Fig.

409 7a) and downstream (Fig. 7e) of the canal showed relatively low explained variances
410 of GPP with R^2 of 0.29 and 0.27, respectively. The canal section from LW to ZYS
411 showed significant R^2 ranging from 0.49 to 0.57, indicating reasonable driving effects
412 on GPP changes.

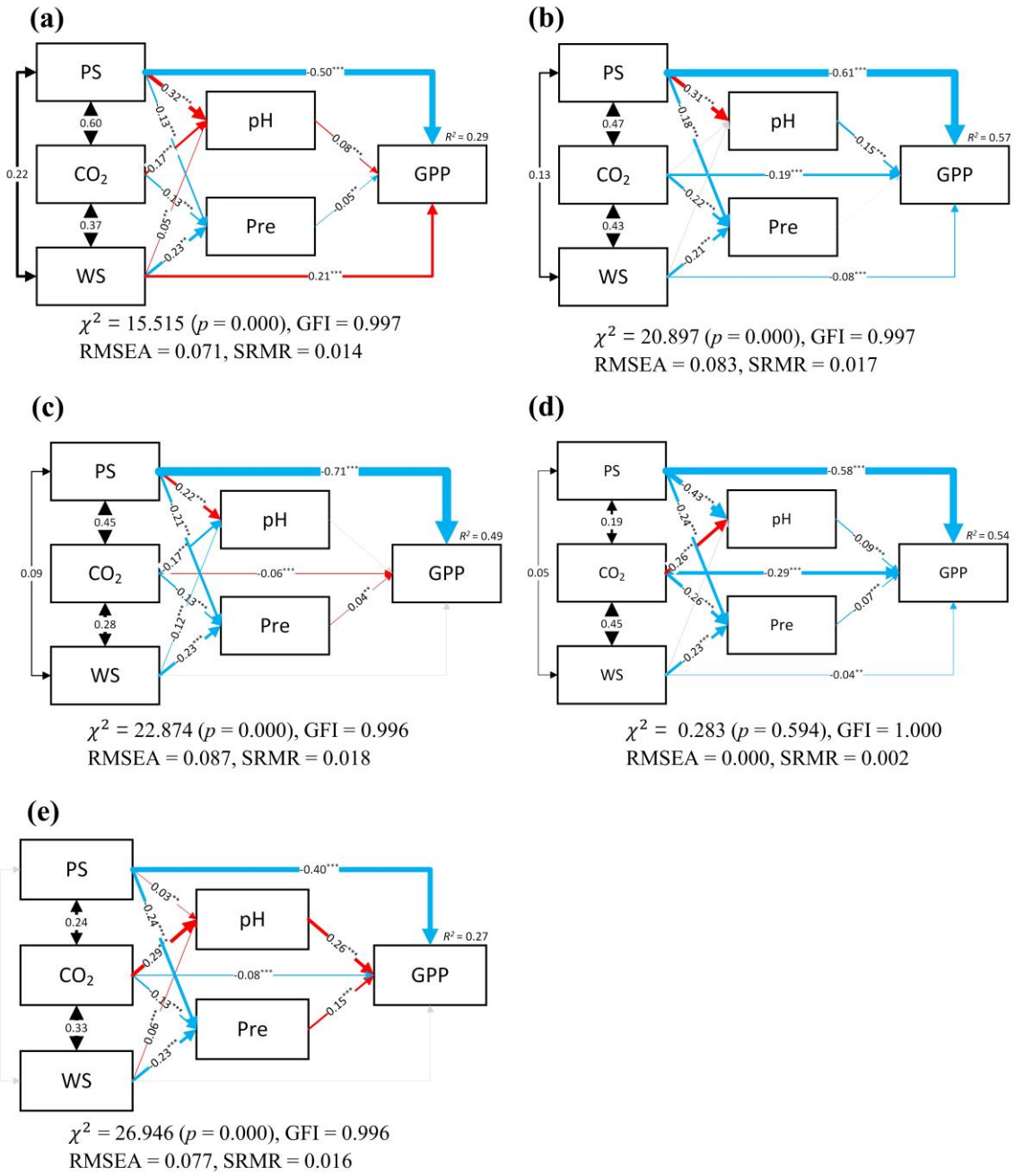


Fig. 7. Standard driving path analysis of GPPs in different sections of the main canal (Note: (a) TC – LW; (b) LW – ZHB; (c) ZHB – TZ; (d) TZ – ZYS; (e) ZYS – HNZ; the red arrow, blue arrow, black, and grey arrow were defined as the positive, negative, correlation, and not significant effect; the "***" and "**" represents the significance level of 0.001 and 0.01 for the factor loadings; the explained variances (R^2) of GPP variations from (a) to (e) are 0.29, 0.57, 0.49, 0.54, and 0.27, respectively).

4 Discussions

4.1 Spatiotemporal variance of GPP analysis in the main canal

Based on the spatial differences of GPPs in the MRSNWDPC canal case, the lowest average GPPs were observed at the starting point of the canal (TC). However, the maximum and minimum monthly average GPPs were detected in the midstream and downstream. Additionally, the highest monthly average GPPs occurred in summer and autumn, consistent with previous studies on the spatial distributions of algal cell density in the MRSNWDPC canal (Rosseel, 2012). These spatiotemporal variation characteristics of GPP can be attributed to the algal, as well as the primary producers, which photosynthesize and absorb atmospheric CO_2 and lead to organic carbon stock in the water bodies (Wang et al., 2022).

However, it should be noted the GPP of this study did not show identically increasing characteristics like the algae density with the canal distance rose in ZYS. Previous studies have reported a significant increase in algal cell density from south to north in the main canal (Segatto et al., 2021). Indeed, the phenomenon of GPP-related variables decoupling is common. Previous research has found features such as the geomorphic characteristics of rivers (Nong et al., 2021), flow turbidity with bed particle composition (Segatto et al., 2021), and the local nutrient variance (Ledford et al., 2021) could exert the decoupling on GPP related variables. Even though fully unveiling and quantitating the complex and interacting controls on GPP requires a large number of additional stations and long data series, these observations show that GPP can serve as an indicator of the ecological environment to provide a comprehensive assessment of

habitat conditions, which were also reported in previous studies (Huang et al., 2018; Ledford et al., 2021; Levi and McIntyre, 2020). Therefore, water quality management agencies should pay more attention to GPP indicators' utility in assessing the ecological functionality of projects.

It should be noticed that the main canal still perseveres certain features similar to natural water bodies, even though it is an artificial water system. For instance, it shares the determining factors of phytoplankton community (water temperature, total phosphorus, ammonia nitrogen, flow speed, and flow discharge) with most surface water bodies and exhibits approximately 20 days long water residence time like lakes (Zhang et al., 2021b). Therefore, several major rivers and coastal water bodies in China were used as a comparison to evaluate the level of GPP in the MRSNWDPC by converting GPP to carbon flux based on the respiratory quotient between O_2 and CO_2 as shown in **Table 1** (Zhang et al., 2023). The average carbon equivalent GPP (0.828 g C m^{-2} d^{-1}) of the MRSNWDPC is closest to that of the Yangtze River (0.684 g C m^{-2} d^{-1}) and the East China Sea (0.873 g C m^{-2} d^{-1}). This may be because both of their water sources come from the Yangtze River resulting in similar physicochemical properties of the water bodies, even if the intake water of this project is sourced from a tributary of the Yangtze River. Overall, the GPP of MRSNWDPC is at a moderate level among China's major rivers in terms of productivity and there was no occurrence of high GPP due to abnormal algal proliferation during the study period.

462 **Table 1**

463 Average GPP comparison with the main canal of the MRSNWDPC and other water bodies in China.

Water bodies	Type	Average (Min-Max) GPP (g C m^{-2} d^{-1})	Length (km)	Surface area (km^2)	Number of sites
Main canal of MRSNWDPC	Open canal	0.828 (0.60 – 1.03)	1,179	-	11
Pearl River	River	0.460 (0.05 – 2.30)	2,320	452,000	8
Yangtze River	River	0.684 (0.07 – 1.35)	6,300	1,800,000	18
Yellow River	River	3.003 (0.001 – 10.66)	5,464	752,443	14
Haihe River	River	2.353 (0.01 – 5.75)	1,031	318,200	17
Liaohe River	River	1.002 (0.01 – 1.89)	1,345	219,600	10
Songhua River	River	3.020 (1.37 – 4.11)	2,309	556,800	18
South China Sea	Coastal zone	1.556 (0.01 – 5.98)	-	3,500,000	51
East China Sea	Coastal zone	0.873 (0.04 – 3.75)	-	770,000	58
Bohai Sea	Coastal zone	0.307 (0.01 – 0.65)	-	77,284	23
Yellow Sea	Coastal zone	0.722 (0.03 – 3.79)	-	380,000	31

464 Note: "-" stands for no data or invalid data. "Number of sites" only includes the monitoring sites located in the mainstream, and sites located in the estuary of the river

465 are not involved. The data on China's main river and coastal area are from (Zhang et al., 2023).

4.2 Driving factors identification of GPP

Random Forest and Spearman's correlation were conducted to identify the driving factors of GPP by analyzing the correlation and explained variance (**Fig. 6**). PS exhibited a higher Spearman correlation coefficient (-0.47 – -0.84) with GPP than other environmental factors. This may be attributed to PS can control the oxygen deficit levels by directly influencing the concentration of SDO, thereby exerting control over GPP. Previous studies have reported that precipitation events can impact GPP in rivers by increasing flow, altering the physicochemical parameters of the water environment, and reducing PAR through cloud cover (Nijboer and Verdonschot, 2004; O'Donnell and Hotchkiss, 2019). However, there is a relatively small correlation between GPP and Pre across different stations, indicating the water levels and flow rates in the canal depend on human regulation instead of increasing through surface runoff caused by Pre, like the natural rivers, resulting in a small impact of Pre on GPP. WS is considered an important factor influencing the gas transfer coefficient K600 (Shen et al., 2022). However, our study shows a slight correlation between GPP and WS, and not all canal sections are affected by WS. Therefore, using WS-based methods to estimate the GPP in water diversion projects may risk overestimating and exaggerating the effect of WS. CO_2 will be mainly discussed in the later path analysis section.

pH, as one of the essential water quality indicators, is acknowledged to have a significant effect on phytoplankton growth by influencing the equilibrium of the carbonate system and controlling the partial pressure of carbon dioxide (Jakobsen et al., 2015). In the case of photosynthesis alone, pH and GPP are positively correlated

because an increase in the intensity of photosynthesis, as one of the processes of gross primary production, simultaneously depletes dissolved CO_2 in the water column, leading to an increase in pH. However, a negative correlation was observed between pH with GPP before the NDG station, while the correlation reversed after this station. This result is consistent with previous research about pH variation with dissolved inorganic carbon (DIC, including CO_2 (aq), HCO_3^-, and CO_3^{2-}) in the water system (Shen et al., 2022). As the flow rate decreases from upstream to downstream (Bukaveckas et al., 2020), large flow disturbance dilution makes it difficult for primary producers to utilize DIC for photosynthesis before the NDG station, resulting in lower pH values. After the NDG station, the flow disturbance on the primary production process is negligible as flow decreases, DICs are mainly absorbed and consumed by primary producers, leading to higher GPP and pH values. This observation is consistent with Zhang et al. (2015) and Hall et al. (2023), which demonstrated that hydrological conditions control the nutrient structure and mass transfer efficiency in water bodies through flushing and stifle effects.

Apart from pH values, GPP is also driven by the nexus of water quality and ecological conditions in the main canal of SNWDPMRC. Zhang et al. (2021a) indicated that the primary producers in the main canal contain only two types, that is phytoplankton and epiphytic algae, due to the special characteristics of the water diversion canal. Therefore, the GPP is supplied by these two types of phytoplankton in the main canal, and changes in the community structure and quantity of the two groups will have a direct impact on the GPP status. Tang et al. (2020) investigated the

relationship between water quality and phytoplankton community based on the sampling data within the entire main canal. The results indicated that the WT affected the dynamic of phytoplankton cell density and dominant taxa in the main canal. Meanwhile, the nutrients can both influence the growth and population structure. On the one hand, phytoplankton absorbed a large amount of ammonia nitrogen for cell growth in the main canal. On the other hand, when nitrogen concentration increased and phosphorus concentration decreased, the diatoms in the main canal gradually succeeded in becoming the dominant taxa.

Overall, measuring GPP can serve as a pivotal way to understand the health of an ecosystem amidst various environmental factors. Enhancing our understanding of the interrelationship between GPP and its driving factors presents a novel ecological option. It offers a comprehensive approach to monitoring and regulating the aquatic environment in long-distance water diversion projects.

4.3 Path analysis of GPP's drivers

Previous studies on SEM of GPP mainly include water physicochemical factors, riparian vegetation cover, nutrient load, and algal biomass in natural rivers (Jia et al., 2020a; Marzolf and Ardón, 2021; Tan et al., 2021; Zhang et al., 2021b), while little research discusses the effects of greenhouse gases on GPP. Furthermore, the turbulent characteristics of water diversion projects could amplify the significant greenhouse gases when compared to the natural rivers. Because inter-basin water diversion projects require maintaining high flow velocity and significant flow rates, the turbulent water-air interface facilitates the entry of greenhouse gases from the surface to the water of

the canal, participating in the river's metabolic processes. Additionally, the heightened water flow induces the formation of bubbles could amplify the exchange between water and air (Ulseth et al., 2019). Therefore, the fast water flow allows more greenhouse gases to enter, providing more carbon supplementation for primary production. Numerous studies have revealed the complex exchange relationships of CO_2 at the water-air interface (Cao et al., 2020; Cole et al., 2007; Gong et al., 2021; Ulseth et al., 2019), such as CO_2 entering the river through photosynthesis and gas diffusion coefficient K (Crawford et al., 2014; Gomez-Gener et al., 2021; Ulseth et al., 2019) controlled the disturbance, mixing state, and intensity at the water-air interface. However, these investigations predominantly focused on natural river systems, and their applicability to canal-based water diversion projects remains partially limited. As such, more specific research on the characteristics and impact of CO_2 and other greenhouse gases in water diversion projects is still needed.

In this study, spatiotemporal variance, driving factors, and path analysis of GPP were performed in an inter-basin water diversion project using multiple data sources. While focusing on the GPP, this important ecological indicator, we could contribute to a better understanding of the artificial aquatic ecosystem in water diversion projects for the government and water quality management departments. It also promotes using GPP for evaluating artificial ecosystems and developing standardized ecological guidelines for inter-basin water diversion projects to manage water ecosystems. Similar research has been successfully applied in various water body restoration (Baattrup-Pedersen et al., 2022; Blersch et al., 2019; Gomez-Gener et al., 2021). Considering that

554 the project will continue to operate for decades, future research focusing on the
555 ecological assessment of inter-basin water diversion projects and carbon-related studies
556 should receive more attention.

557 ## 5 Conclusions

558 The spatiotemporal variations of GPP in the Middle Route of the South-to-North
559 Water Diversion Project of China were studied based on 11 water quality monitoring
560 stations and satellite re-analysis data from 2017 to 2020. The environmental driving
561 factors of GPP and effect pathways were analyzed using random forest and structural
562 equation modeling methods. The main findings of this study are as follows:

563 (1) The overall GPP level of the canal ranged from 1.920 to 3.290 g O_2 m^{-2} d^{-1},
564 which presents similar GPP levels with the Yangtze River and the East China
565 Sea while comparing to other major rivers and estuary areas of China. This
566 indicates the ecosystems and ecological service of the main canal of
567 MRSNWDPC are healthy and well-functioning.

568 (2) The GPP exhibited significant spatial differences in the main canal, with the
569 concentrations gradually increasing from upstream to downstream, and the
570 highest monthly average GPPs occurred in summer, while the lowest in winter.

571 (3) Path analysis of GPP factors in the canal revealed significant causal driving
572 relationships of CO_2, pH, and PS on GPP variations. The PS had significant
573 negative impacts on GPP changes in the canal, while CO_2 and pH showed
574 different direction effects in different sections. Additional attention should be
575 given to greenhouse gas emissions in water diversion project management.

This study provides a new ecological indicator for evaluating the aquatic ecosystem in the complex and unique context of long-distance inter-basin water diversion projects. It reveals the spatiotemporal variations of GPP, and the relationships with driving factors and offers a new perspective for water quality management in large-scale water projects. The findings of this study can also be applied to other large-scale inter-basin water diversion projects. In the current situation of sharp water resources supply conflicts, this study provides insights into the ecological status of artificial large-scale water projects from an ecosystem perspective. It also suggests that mega hydro-projects require specific attention when studying the carbon cycle contribution of water bodies.

Acknowledgments

This research was funded by the Specific Research Project of Guangxi for Research Bases and Talents (No.AD22035185), the Youth Science Foundation of Guangxi (No.2023GXNSFBA026296), the National Natural Science Foundation of China (No.52309016), the visiting scholars' fund at the WRHES (No.2021NSG02), the Belt and Road Special Foundation of the National Key Laboratory of Water Disaster Prevention (No.2022nkms06). The authors would like to thank the editors and the reviewers for their valuable suggestions and contributions which significantly improved this article, and would also need to acknowledge the Construction and Administration Bureau of the Middle Route of the South-to-North Water Diversion Project of China that supported the data collection.

Reference

Antonopoulos, V.Z. and Gianniou, S.K. 2003. Simulation of water temperature and dissolved oxygen distribution in Lake Vegoritis, Greece. Ecological Modelling 160(1), 39-53.

Appling, A.P., Hall, R.O., Yackulic, C.B. and Arroita, M. 2018a. Overcoming Equifinality: Leveraging Long Time Series for Stream Metabolism Estimation. Journal of Geophysical Research-Biogeosciences 123(2), 624-645.

Appling, A.P., Read, J.S., Winslow, L.A., Arroita, M., Bernhardt, E.S., Griffiths, N.A., Hall, R.O., Harvey, J.W., Heffernan, J.B., Stanley, E.H., Stets, E.G. and Yackulic, C.B. 2018b. The metabolic regimes of 356 rivers in the United States. Scientific Data 5, 180292.

Aristi, I., Arroita, M., Larranaga, A., Ponsati, L., Sabater, S., von Schiller, D., Elosegi, A. and Acuna, V. 2014. Flow regulation by dams affects ecosystem metabolism in Mediterranean rivers. Freshwater Biology 59(9), 1816-1829.

Baattrup-Pedersen, A., Alnoe, A.B., Rasmussen, J.J., Levi, P.S., Friberg, N. and Riis, T. 2022. Stream restoration and ecosystem functioning in lowland streams. Ecological Engineering 184, 106782.

Bai, Y. and Cotrufo, M.F. 2022. Grassland soil carbon sequestration: Current understanding, challenges, and solutions. Science 377(6606), 603-608.

Battin, T.J., Lauerwald, R., Bernhardt, E.S., Bertuzzo, E., Gener, L.G., Hall, R., Hotchkiss, E.R., Maavara, T., Pavelsky, T.M., Ran, L.S., Raymond, P., Rosentreter, J.A. and Regnier, P. 2023. River ecosystem metabolism and carbon

biogeochemistry in a changing world. Nature 613(7944), 449-459.

Bernhardt, E.S., Heffernan, J.B., Grimm, N.B., Stanley, E.H., Harvey, J.W., Arroita, M., Appling, A.P., Cohen, M.J., McDowell, W.H., Hall, R.O., Read, J.S., Roberts, B.J., Stets, E.G. and Yackulic, C.B. 2018. The metabolic regimes of flowing waters. Limnology and Oceanography 63, S99-S118.

Blersch, S.S., Blersch, D.M. and Atkinson, J.F. 2019. Metabolic Variance: A Metric to Detect Shifts in Stream Ecosystem Function as a Result of Stream Restoration. Journal of the American Water Resources Association 55(3), 608-621.

Breiman, L. 2001. Random forests. Mach. Learn. 45(1), 5-32.

Brooks, S.P. and Gelman, A. 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics 7(4), 434-455.

Bukaveckas, P.A., Tassone, S., Lee, W. and Franklin, R.B. 2020. The Influence of Storm Events on Metabolism and Water Quality of Riverine and Estuarine Segments of the James, Mattaponi, and Pamunkey Rivers. Estuaries and Coasts 43(7), 1585-1602.

Bunn, S.E., Davies, P.M. and Mosisch, T.D. 1999. Ecosystem measures of river health and their response to riparian and catchment degradation. Freshwater Biology 41(2), 333-345.

Burpee, B.T., Saros, J.E., Nanus, L., Baron, J., Brahney, J., Christianson, K.R., Ganz, T., Heard, A., Hundey, B., Koinig, K.A., Kopacek, J., Moser, K., Nydick, K., Oleksy, I., Sadro, S., Sommaruga, R., Vinebrooke, R. and Williams, J. 2022.

641 Identifying factors that affect mountain lake sensitivity to atmospheric nitrogen
642 deposition across multiple scales. Water Research 209, 117883.
643 Campeau, A. and Del Giorgio, P.A. 2014. Patterns in CH₄ and
644 CO₂ concentrations across boreal rivers: Major drivers and
645 implications for fluvial greenhouse emissions under climate change scenarios.
646 Global Change Biology 20(4), 1075-1088.
647 Cao, Z.M., Yang, W., Zhao, Y.Y., Guo, X.H., Yin, Z.Q., Du, C.J., Zhao, H.D. and Dai,
648 M.H. 2020. Diagnosis of CO2 dynamics and fluxes in global coastal oceans.
649 National Science Review 7(4), 786-797.
650 Cole, J.J., Prairie, Y.T., Caraco, N.F., McDowell, W.H., Tranvik, L.J., Striegl, R.G.,
651 Duarte, C.M., Kortelainen, P., Downing, J.A., Middelburg, J.J. and Melack, J.
652 2007. Plumbing the global carbon cycle: Integrating inland waters into the
653 terrestrial carbon budget. Ecosystems 10(1), 171-184.
654 Crawford, J.T., Stanley, E.H., Spawn, S.A., Finlay, J.C., Loken, L.C. and Striegl, R.G.
655 2014. Ebullitive methane emissions from oxygenated wetland streams. Global
656 Change Biology 20(11), 3408-3422.
657 Duan, K., Caldwell, P.V., Sun, G., McNulty, S.G., Qin, Y., Chen, X.H. and Liu, N. 2022.
658 Climate change challenges efficiency of inter-basin water transfers in alleviating
659 water stress. Environmental Research Letters 17(4), 044050.
660 Ehrlinger, L. and Woess, W. 2022. A Survey of Data Quality Measurement and
661 Monitoring Tools. Front. Big Data 5, 30.
662 Feng, D., Fang, K. and Shen, C. 2020. Enhancing Streamflow Forecast and Extracting

663 Insights Using Long-Short Term Memory Networks With Data Integration at
664 Continental Scales. Water Resources Research 56(9), e2019WR026793.
665 Gao, S., Schwinger, J., Tjiputra, J., Bethke, I., Hartmann, J., Mayorga, E. and Heinze,
666 C. 2023. Riverine impact on future projections of marine primary production and
667 carbon uptake. Biogeosciences 20(1), 93-119.
668 Gelman, A. and Rubin, D.B. 1992. Inference from Iterative Simulation Using Multiple
669 Sequences. Statistical Science 7, 457-472.
670 Genzoli, L. and Hall, R.O. 2016. Shifts in Klamath River metabolism following a
671 reservoir cyanobacterial bloom. Freshwater Science 35(3), 795-809.
672 Gomez-Gener, L., Rocher-Ros, G., Battin, T., Cohen, M.J., Dalmagro, H.J., Dinsmore,
673 K.J., Drake, T.W., Duvert, C., Enrich-Prast, A., Horgby, A., Johnson, M.S., Kirk,
674 L., Machado-Silva, F., Marzolf, N.S., McDowell, M.J., McDowell, W.H.,
675 Miettinen, H., Ojala, A.K., Peter, H., Pumpanen, J., Ran, L.S., Riveros-Iregui,
676 D.A., Santos, I.R., Six, J., Stanley, E.H., Wallin, M.B., White, S.A. and Sponseller,
677 R.A. 2021. Global carbon dioxide efflux from rivers enhanced by high nocturnal
678 emissions. Nature Geoscience 14(5), 289-294.
679 Gong, C., Yan, W.J., Zhang, P.P., Yu, Q.B., Li, Y.Q., Li, X.Y., Wang, D.S. and Jiao, R.Y.
680 2021. Effects of stream ecosystem metabolisms on CO2 emissions in two
681 headwater catchments, Southeastern China. Ecological Indicators 130(97),
682 108136.
683 Hall, N., Testa, J., Li, M. and Paerl, H. 2023. Assessing drivers of estuarine pH: A
684 comparative analysis of the continental USA's two largest estuaries.

Limnology and Oceanography 68, 2227-2244.

Hall, R.O., Yackulic, C.B., Kennedy, T.A., Yard, M.D., Rosi-Marshall, E.J., Voichick, N. and Behn, K.E. 2015. Turbidity, light, temperature, and hydropeaking control primary productivity in the Colorado River, Grand Canyon. Limnology and Oceanography 60(2), 512-526.

Heathwaite, A.L. 2010. Multiple stressors on water availability at global to catchment scales: understanding human impact on nutrient cycles to protect water quality and water availability in the long term. Freshwater Biology 55(s1), 241-257.

Hoellein, T.J., Bruesewitz, D.A. and Richardson, D.C. 2013. Revisiting Odum (1956): A synthesis of aquatic ecosystem metabolism. Limnology and Oceanography 58(6), 2089-2100.

Huang, W., Liu, X.B., Peng, W.Q., Wu, L.X., Yano, S., Zhang, J.M. and Zhao, F. 2018. Periphyton and ecosystem metabolism as indicators of river ecosystem response to environmental flow restoration in a flow-reduced river. Ecological Indicators 92, 394-401.

Inness, A., Ades, M, Agustí-Panareda, A, Barré, J, Benedictow, A, Blechschmidt, A, Dominguez, J, Engelen, R, Eskes, H, Flemming, J, Huijnen, V, Jones, L, Kipling, Z, Massart, S, Parrington, M, Peuch, V-H, Razinger M, Remy, S, Schulz, M and Suttie, M 2019 Copernicus Atmosphere Monitoring Service (CAMS) global greenhouse gas reanalysis (EGG4), Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS), Accessed on <14-Aug-2022>.

Jähne, B., Münnich, K.O., Bösinger, R., Dutzi, A., Huber, W. and Libner, P. 1987. On

707 the parameters influencing air-water gas exchange. Journal of Geophysical
708 Research: Oceans 92(C2), 1937-1949.
709 Jakobsen, H.H., Blanda, E., Staehr, P.A., Højgård, J.K., Rayner, T.A., Pedersen, M.F.,
710 Jepsen, P.M. and Hansen, B.W. 2015. Development of phytoplankton
711 communities: Implications of nutrient injections on phytoplankton composition,
712 pH and ecosystem production. Journal of Experimental Marine Biology and
713 Ecology 473, 81-89.
714 Jia, J.J., Gao, Y., Zhou, F., Shi, K., Johnes, P.J., Dungait, J.A.J., Ma, M.Z. and Lu, Y.
715 2020a. Identifying the main drivers of change of phytoplankton community
716 structure and gross primary productivity in a river-lake system. Journal of
717 Hydrology 583, 13.
718 Jia, J.J., Gao, Y., Zhou, F., Shi, K., Johnes, P.J., Dungait, J.A.J., Ma, M.Z. and Lu, Y.
719 2020b. Identifying the main drivers of change of phytoplankton community
720 structure and gross primary productivity in a river-lake system. Journal of
721 Hydrology 583, 124633.
722 Ledford, S.H., Kurz, M.J. and Toran, L. 2021. Contrasting Raz–Rru stream
723 metabolism and nutrient uptake downstream of urban wastewater effluent sites.
724 Freshwater Science 40(1), 103-119.
725 Levi, P.S. and McIntyre, P.B. 2020. Ecosystem responses to channel restoration decline
726 with stream size in urban river networks. Ecological Applications 30(5), e02107.
727 Marzolf, N.S. and Ardón, M. 2021. Ecosystem metabolism in tropical streams and
728 rivers: a review and synthesis. Limnology and Oceanography 66(5), 1627-1638.

729 Muñoz Sabater, J. 2019 ERA5-Land hourly data from 1950 to present. Muñoz Sabater,
730 J. (ed), Copernicus Climate Change Service (C3S) Climate Data Store (CDS),
731 Accessed on <14-Aug-2022>.
732 Nasa/Larc/Sd/Asdc 2017 CERES and GEO-Enhanced TOA, Within-Atmosphere and
733 Surface Fluxes, Clouds and Aerosols 1-Hourly Terra-Aqua Edition4A.
734 Nasa/Larc/Sd/Asdc (ed).
735 Nijboer, R.C. and Verdonschot, P.F.M. 2004. Variable selection for modelling effects
736 of eutrophication on stream and river ecosystems. Ecological Modelling 177(1-2),
737 17-39.
738 Nong, X.Z., Shao, D.G., Shang, Y.M. and Liang, J.K. 2021. Analysis of spatio-
739 temporal variation in phytoplankton and its relationship with water quality
740 parameters in the South-to-North Water Diversion Project of China.
741 Environmental Monitoring and Assessment 193(9), 593.
742 Nong, X.Z., Shao, D.G., Zhong, H. and Liang, J.K. 2020. Evaluation of water quality
743 in the South-to-North Water Diversion Project of China using the water quality
744 index (WQI) method. Water Research 178, 115781.
745 O'Donnell, B. and Hotchkiss, E.R. 2019. Coupling Concentration- and Process-
746 Discharge Relationships Integrates Water Chemistry and Metabolism in Streams.
747 Water Resources Research 55(12), 10179-10190.
748 Olson, C.R., Solomon, C.T. and Jones, S.E. 2020. Shifting limitation of primary
749 production: experimental support for a new model in lake ecosystems. Ecology
750 Letters 23(12), 1800-1808.

751 Palmer, M.A. and Febria, C.M. 2012. The Heartbeat of Ecosystems. Science
752 336(6087), 1393-1394.
753 Palt, M., Le Gall, M., Piffady, J., Hering, D. and Kail, J. 2022. A metric-based analysis
754 on the effects of riparian and catchment landuse on macroinvertebrates. Science
755 of the Total Environment 816, 151590.
756 Prichard, A.H. and Scott, C.A. 2014. Interbasin water transfers at the US-Mexico
757 border city of Nogales, Sonora: implications for aquifers and water security. Int. J.
758 Water Resour. Dev. 30(1), 135-151.
759 Quinn, J.M. and McFarlane, P.N. 1989. Epilithon and dissolved oxygen depletion in
760 the Manawatu River, New Zealand: Simple models and management implications.
761 Water Research 23(7), 825-832.
762 Raymond, P.A., Zappa, C.J., Butman, D., Bott, T.L., Potter, J., Mulholland, P., Laursen,
763 A.E., McDowell, W.H. and Newbold, D. 2012. Scaling the gas transfer velocity
764 and hydraulic geometry in streams and small rivers. Limnology and Oceanography:
765 Fluids and Environments 2(1), 41-53.
766 Rodriguez-Castillo, T., Estevez, E., Gonzalez-Ferreras, A.M. and Barquin, J. 2019.
767 Estimating Ecosystem Metabolism to Entire River Networks. Ecosystems 22(4),
768 892-911.
769 Rosseel, Y. 2012. lavaan: An R Package for Structural Equation Modeling. Journal of
770 Statistical Software 48(2), 1 - 36.
771 Sabater, S., Armengol, J., Comas, E., Sabater, F., Urrizalqui, I. and Urrutia, I. 2000.
772 Algal biomass in a disturbed Atlantic river: water quality relationships and

773 environmental implications. Science of The Total Environment 263(1), 185-195.
774 Segatto, P.L., Battin, T.J. and Bertuzzo, E. 2021. The Metabolic Regimes at the Scale
775 of an Entire Stream Network Unveiled Through Sensor Data and Machine
776 Learning. Ecosystems 24(7), 1792-1809.
777 Shen, X.M., Cai, Y.P., Su, M.R., Wan, H., Shen, Y.M. and Yang, Z.F. 2022. High
778 discharge intensified low net ecosystem productivity, hypoxia, and acidification at
779 three outlets of the Pearl River Estuary, China. Water Research 214, 118171.
780 Shen, X.M., Sun, T., Liu, F.F., Xu, J. and Pang, A.P. 2015. Aquatic metabolism
781 response to the hydrologic alteration in the Yellow River estuary, China. Journal
782 of Hydrology 525, 42-54.
783 Song, C.G., Luo, F.L., Zhang, L.L., Yi, L.B., Wang, C.Y., Yang, Y.S., Li, J.X., Chen,
784 K.L., Wang, W.Y., Li, Y.N. and Zhang, F.W. 2021. Nongrowing Season CO2
785 Emissions Determine the Distinct Carbon Budgets of Two Alpine Wetlands on the
786 Northeastern Qinghai-Tibet Plateau. Atmosphere 12(12), 1695.
787 Spilling, K., Fuentes-Lema, A., Quemalinos, D., Klais, R. and Sobrino, C. 2019.
788 Primary production, carbon release, and respiration during spring bloom in the
789 Baltic Sea. Limnology and Oceanography 64(4), 1779-1789.
790 Tan, X., Hou, E.Q., Zhang, Q.F. and Bunn, S.E. 2021. Benthic metabolism responses
791 to environmental attributes at multiple scales and its linkage to algal community
792 structure in streams. Hydrobiologia 848(21), 5067-5085.
793 Tang, J., Xiao, X., Wang, Y., Hu, S. and Wang, Y. 2020. Ecosystem structure and
794 function of the main channel of the middle route of south-north water diversion

795 project (in Chinese). China Environmental Science 40(12), 5391-5402.
796 Tang, S., Sun, T., Shen, X., Qi, M. and Feng, M. 2015. Modeling net ecosystem
797 metabolism influenced by artificial hydrological regulation: An application to the
798 Yellow River Estuary, China. Ecological Engineering 76, 84-94.
799 The MathWorks, I. 2022. MATLAB version: 9.13.0 (R2022b).
800 Torres, M.E., Colominas, M.A., Schlotthauer, G. and Flandrin, P. 2011 A complete
801 ensemble empirical mode decomposition with adaptive noise, pp. 4144-4147.
802 Ulseth, A.J., Hall, R.O., Boix Canadell, M., Madinger, H.L., Niayifar, A. and Battin,
803 T.J. 2019. Distinct air–water gas exchange regimes in low- and high-energy
804 streams. Nature Geoscience 12(4), 259-263.
805 Val, J., Chinarro, D., Pino, M.R. and Navarro, E. 2016. Global change impacts on river
806 ecosystems: A high-resolution watershed study of Ebro river metabolism. Science
807 of The Total Environment 569, 774-783.
808 von Schiller, D., Acuña, V., Aristi, I., Arroita, M., Basaguren, A., Bellin, A., Boyero, L.,
809 Butturini, A., Ginebreda, A., Kalogianni, E., Larrañaga, A., Majone, B., Martínez,
810 A., Monroy, S., Muñoz, I., Paunović, M., Pereda, O., Petrovic, M., Pozo, J.,
811 Rodríguez-Mozaz, S., Rivas, D., Sabater, S., Sabater, F., Skoulikidis, N.,
812 Solagaistua, L., Vardakas, L. and Elosegi, A. 2017. River ecosystem processes:
813 A synthesis of approaches, criteria of use and sensitivity to environmental stressors.
814 Science of The Total Environment 596-597, 465-480.
815 Wang, C., Zhang, H., Lei, P., Xin, X.K., Zhang, A.J. and Yin, W. 2022. Evidence on
816 the causes of the rising levels of CODMn along the middle route of the South-to-

North Diversion Project in China: The role of algal dissolved organic matter. Journal of Environmental Sciences 113, 281-290.

Woodford, D.J., Hui, C., Richardson, D.M. and Weyl, O.L.F. 2013. Propagule pressure drives establishment of introduced freshwater fish: quantitative evidence from an irrigation network. Ecological Applications 23(8), 1926-1937.

Yu, M., Wang, C., Liu, Y., Olsson, G. and Wang, C. 2018. Sustainability of mega water diversion projects: Experience and lessons from China. Science of the Total Environment 619, 721-731.

Zhang, C., Nong, X.Z., Shao, D.G., Zhong, H., Shang, Y.M. and Liang, J.K. 2021a. Multivariate water environmental risk analysis in long-distance water supply project: A case study in China. Ecological Indicators 125, 70-82.

Zhang, C., Zhu, Y., Song, G., Mi, W., Bi, Y., Wang, S., Liang, J. and Shang, M. 2021b. Spatiotemporal pattern of phytoplankton community structure and its determining factors in the channel of the middle route of South-to-North Water Diversion Project (in Chinese). Journal of Lake Sciences 33(3), 675-686.

Zhang, M., Francis, R.A. and Chadwick, M.A. 2023. A synthesis of ecosystem metabolism of China's major rivers and coastal zones (2000-2020). Wiley Interdisciplinary Reviews-Water 10(2), e1628.

Zhang, Q., Yuan, Y., Mi, W., Yang, Y., Bi, Y. and Hu, Z. 2015. Primary production and its influencing factors in Xiangxi River, Three-Gorges Reservoir (in Chinese). Journal of Lake Sciences 27(3), 436-444.

Zhang, Y.H. and Ye, A.Z. 2021. Would the obtainable gross primary productivity (GPP)

products stand up? A critical assessment of 45 global GPP products. Science of the Total Environment 783, 146965.

Figure captions

Fig. 1. Locations of the water quality monitoring stations along the main canal of the Middle Route of the South-to-North Water Diversion Project of China in this study (Note: TC to FCN are "upstream", ZHB to TZ are "midstream", and XHS to HNZ are "downstream").

Fig. 2. In-situ monitoring and satellite re-analysis data processing scheme diagram in this study (Note: In basic data cleaning, the first and second numbers of "window sizes" stand for the backward and forward window sizes respectively; In deep data cleaning, numbers and bracketed words are values and names of the algorithm parameters, "$(\cdot)$" imply the input signal of sample entropy; In using streamMetabolizer function, enclosed content and subsequent text are the name of the R package and specific function respectively).

Fig. 3. Statistical summary of environmental variables in the main canal of MRSNWDPC (Note: (a) statistical summary of the GPP estimation dataset; (b) Spearman correlation matrix of the GPP analysis dataset and the Mantel test between the GPP estimation dataset and analysis dataset; (c) statistical summary of the GPP analysis dataset; Detailed information about the variables' monthly variance, section-scale mantel tests, and stations with WD data included can be found in **Fig. S2**, **Fig. S3**, and **Table S1**).

Fig. 4. Spatiotemporal variations of the GPP in the main canal of MRSNWDPC from 2017 to 2020 (Note: (a) The GPP distributions in different stations and periods based on the violin plots; (b) The mode distribution of GPPs in different seasons).

Fig. 5. Time series of GPP estimation in the entire canal and different canal sections from 2017 to 2020 (Note: the "black line" represents daily average GPPs, the "red dotted" line represents monthly

863 average GPPs, and the “shaded area” represents the intraday daily GPP variation range in all
864 stations).

865 **Fig. 6.** Driving factors analysis of GPP variation in the MRSNWDPC based on the Spearman
866 correlation matrix and random forest-based explained variances (Note: different rectangles with
867 circles represent the combinations of Spearman’s ρ and “percent increase in mean square error
868 (%IncMSE)” based on the random forest regression predictions between GPP and the corresponding
869 environmental factors).

870 **Fig. 7.** Standard driving path analysis of GPPs in different sections of the main canal (Note: (a) TC
871 – LW; (b) LW – ZHB; (c) ZHB – TZ; (d) TZ – ZYS; (e) ZYS – HNZ; the red arrow, blue arrow,
872 black, and grey arrow were defined as the positive, negative, correlation, and not significant effect;
873 the “***” and “**” represents the significance level of 0.001 and 0.01 for the factor loadings; the
874 explained variances (R2) of GPP variations from (a) to (e) are 0.29, 0.57, 0.49, 0.54, and 0.27,
875 respectively).

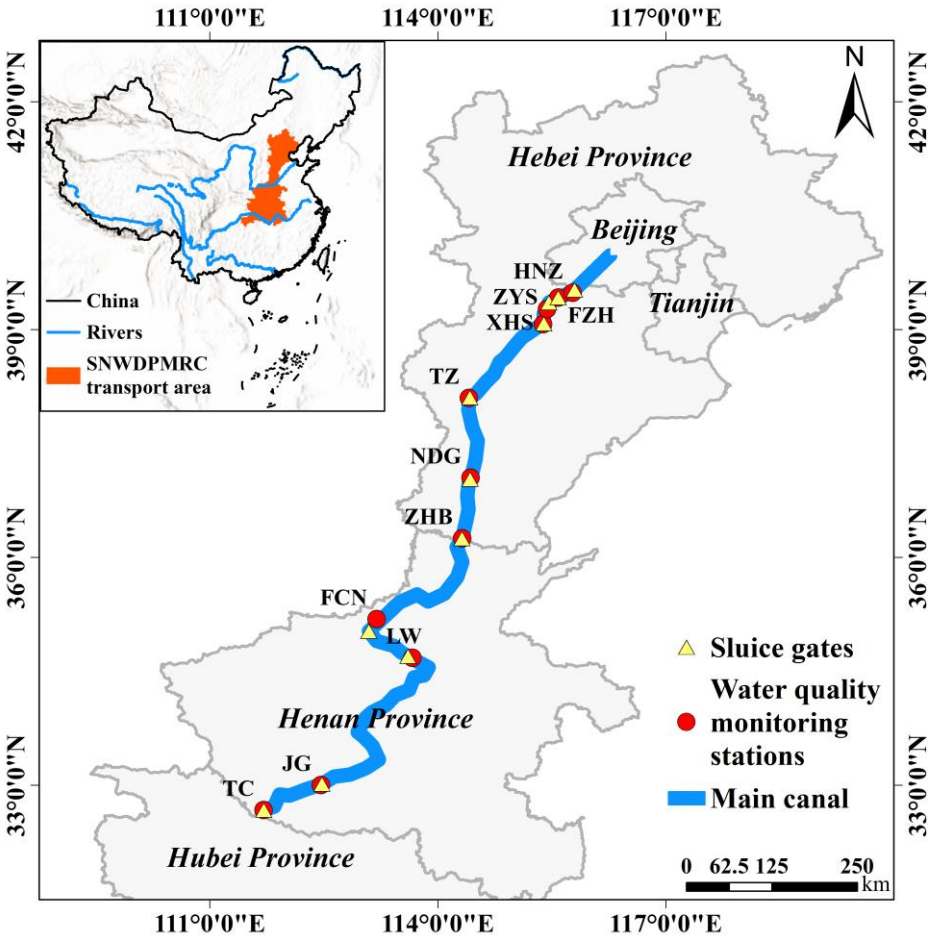


Fig. 1. Locations of the water quality monitoring stations along the main canal of the Middle Route of the South-to-North Water Diversion Project of China in this study (Note: TC to FCN are "upstream", ZHB to TZ are "midstream", and XHS to HNZ are "downstream").

876

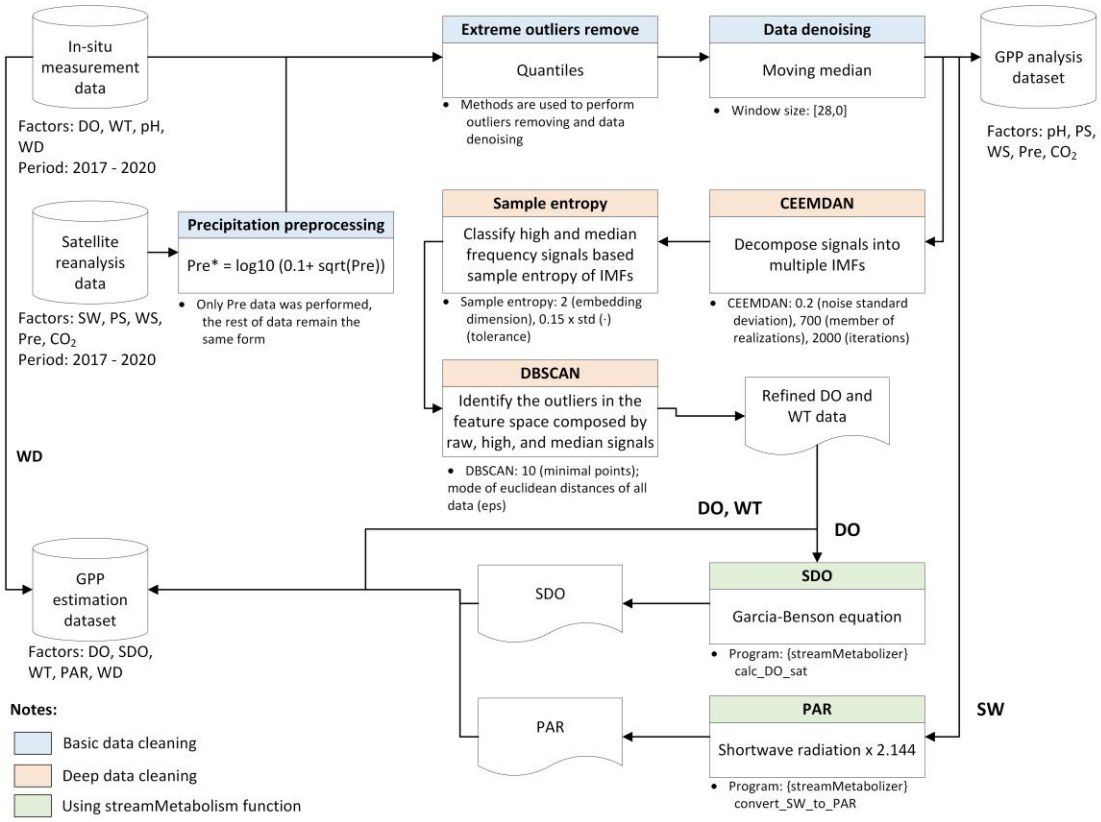
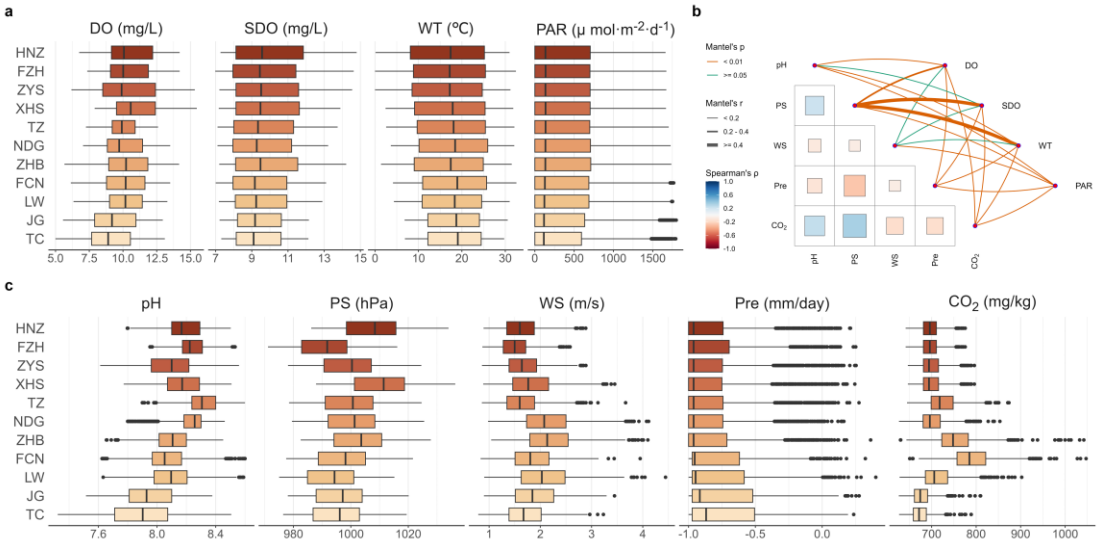


Fig. 2. In-situ monitoring and satellite re-analysis data processing scheme diagram in this study (Note: In basic data cleaning, the first and second numbers of "window sizes" stand for the backward and forward window sizes respectively; In deep data cleaning, numbers and bracketed words are values and names of the algorithm parameters, "(·)" imply the input signal of sample entropy; In using streamMetabolizer function, enclosed content and subsequent text are the name of the R package and specific function respectively).



884

885 **Fig. 3.** Statistical summary of environmental variables in the main canal of MRSNWDPC (Note: (a)

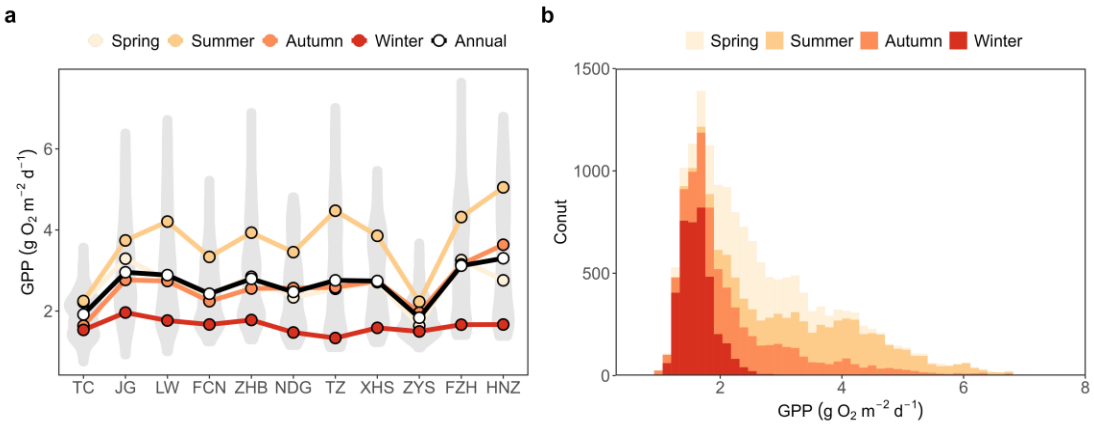
886 statistical summary of the GPP estimation dataset; (b) Spearman correlation matrix of the GPP

887 analysis dataset and the Mantel test between the GPP estimation dataset and analysis dataset; (c)

888 statistical summary of the GPP analysis dataset; Detailed information about the variables' monthly

889 variance, section-scale mantel tests, and stations with WD data included can be found in **Fig. S2**,

890 **Fig. S3**, and **Table S1**).



891

892 **Fig. 4.** Spatiotemporal variations of the GPP in the main canal of MRSNWDPC from 2017 to 2020

893 (Note: (a) The GPP distributions in different stations and periods based on the violin plots; (b) The

894 mode distribution of GPPs in different seasons).

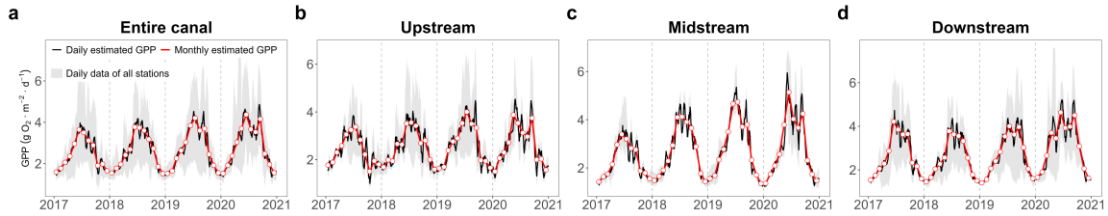


Fig. 5. Time series of GPP estimation in the entire canal and different canal sections from 2017 to 2020 (Note: the "black line" represents daily average GPPs, the "red dotted" line represents monthly average GPPs, and the "shaded area" represents the intraday daily GPP variation range in all stations).

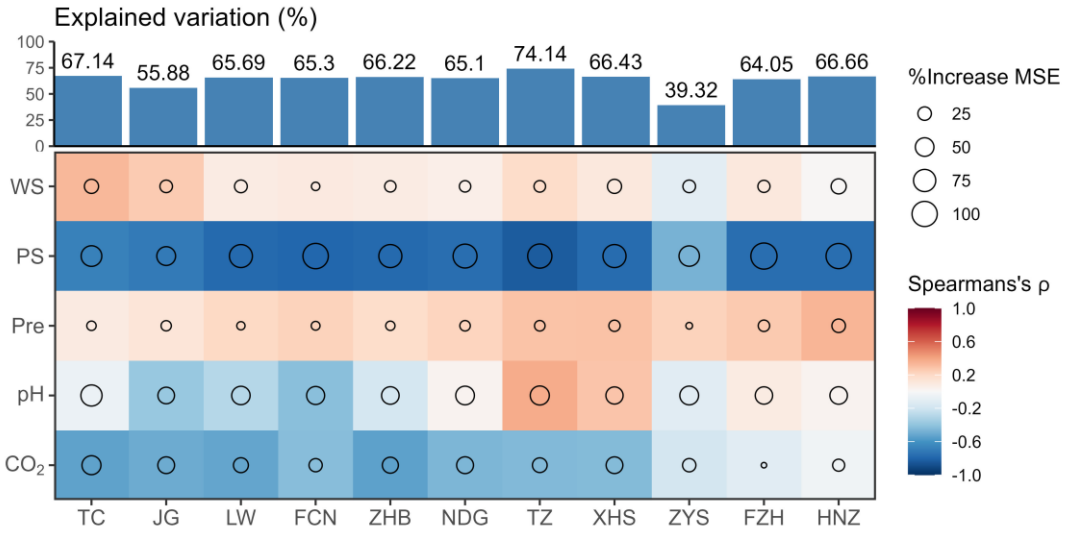


Fig. 6. Driving factors analysis of GPP variation in the MRSNWDPC based on the Spearman correlation matrix and random forest-based explained variances (Note: different rectangles with circles represent the combinations of Spearman's ρ and "percent increase in mean square error (%IncMSE)" based on the random forest regression predictions between GPP and the corresponding environmental factors).

900

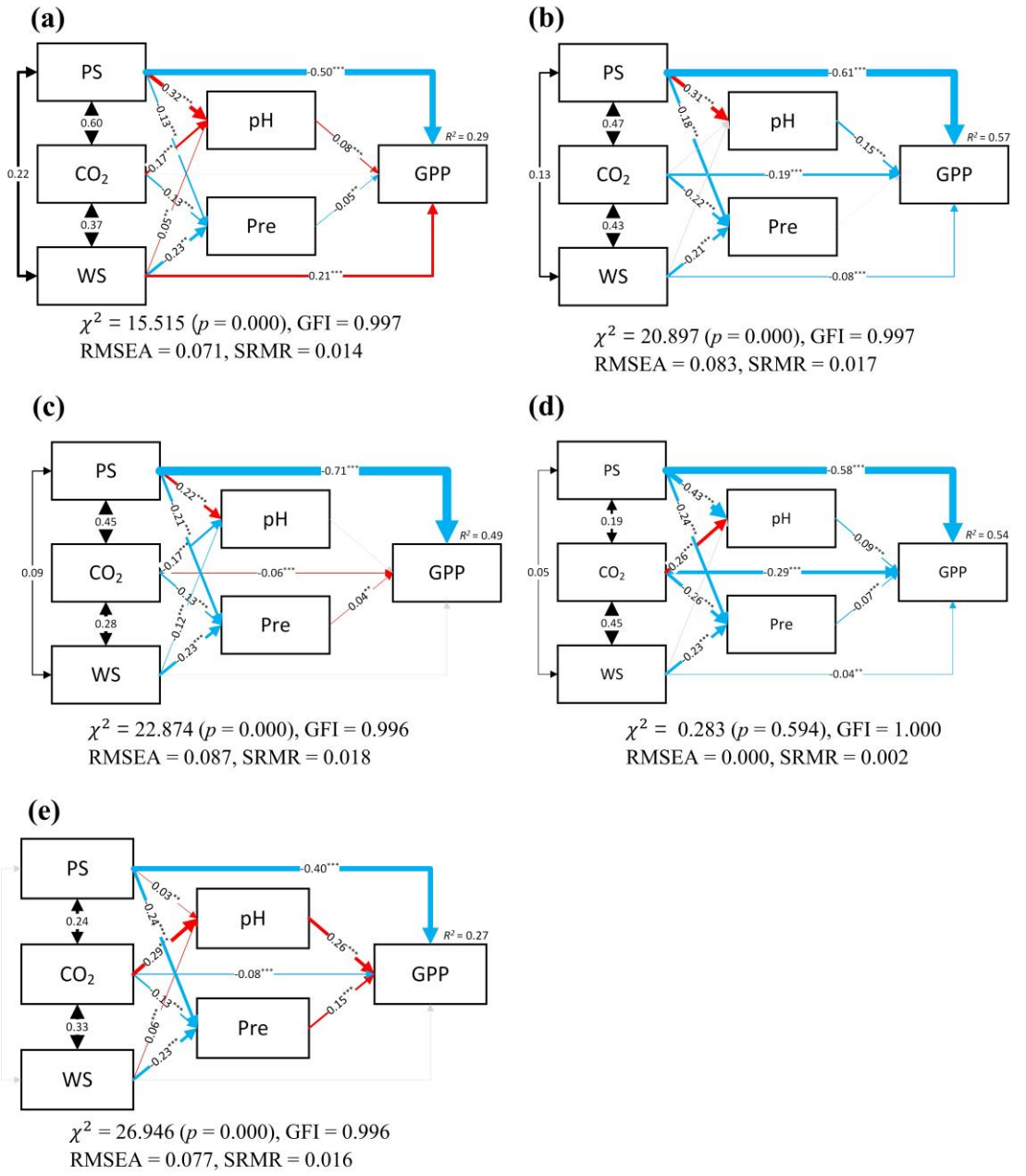


Fig. 7. Standard driving path analysis of GPPs in different sections of the main canal (Note: (a) TC – LW; (b) LW – ZHB; (c) ZHB – TZ; (d) TZ – ZYS; (e) ZYS – HNZ; the red arrow, blue arrow, black, and grey arrow were defined as the positive, negative, correlation, and not significant effect; the "***" and "**" represents the significance level of 0.001 and 0.01 for the factor loadings; the explained variances (R^2) of GPP variations from (a) to (e) are 0.29, 0.57, 0.49, 0.54, and 0.27, respectively).

Table

Table 1

Average GPP comparison with the main canal of the MRSNWDPC and other water bodies in China.

Water bodies	Type	Average (Min-Max) GPP (g C m^{-2} d^{-1})	Length (km)	Surface area (km^2)	Number of sites
Main canal of MRSNWDPC	Open canal	0.828 (0.60 – 1.03)	1,179	-	11
Pearl River	River	0.460 (0.05 – 2.30)	2,320	452,000	8
Yangtze River	River	0.684 (0.07 – 1.35)	6,300	1,800,000	18
Yellow River	River	3.003 (0.001 – 10.66)	5,464	752,443	14
Haihe River	River	2.353 (0.01 – 5.75)	1,031	318,200	17
Liaohe River	River	1.002 (0.01 – 1.89)	1,345	219,600	10
Songhua River	River	3.020 (1.37 – 4.11)	2,309	556,800	18
South China Sea	Coastal zone	1.556 (0.01 – 5.98)	-	3,500,000	51
East China Sea	Coastal zone	0.873 (0.04 – 3.75)	-	770,000	58
Bohai Sea	Coastal zone	0.307 (0.01 – 0.65)	-	77,284	23
Yellow Sea	Coastal zone	0.722 (0.03 – 3.79)	-	380,000	31

Note: "-" stands for no data or invalid data. Number of sites only includes the monitoring sites located in the mainstream, and sites locate in the estuary of the river are not involved. The data on China's main river and coastal area are from (Zhang et al., 2023).

Supplementary Materials

Variability and driving effect of aquatic gross primary productivity across long-distance inter-basin water diversion project

Cheng Lai [a], Xizhi Nong [a, b, *], Lihua Chen [a], Chi Zhang [b], Luiza C. Campos [c, *], Kourosh Behzadian [d], Ronghui Li [a, *]

[a] College of Civil Engineering and Architecture, Guangxi University, Nanning 530004, China

[b] State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan 430072, China

[c] Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E 6BT, UK

[d] School of Computing and Engineering, University of West London, London W5 5RF, UK

*Corresponding author:

Xizhi Nong, E-mail address: nongxizhi@gxu.edu.cn

Luiza C. Campos, E-mail address: l.campos@ucl.ac.uk

Ronghui Li, E-mail address: lironghui@gxu.edu.cn

Complete ensemble empirical mode decomposition

Complete ensemble empirical mode decomposition (CEEMDAN) is a data-driven, non-linear, non-stationary adaptive signal decomposition method that has been developed based on the Empirical Mode Decomposition (EMD) technique (Torres et al., 2011). The core of the EMD-based method lies in its ability to reveal local oscillations of the time series data by considering high-frequency and low-frequency oscillatory signals at multiple decomposition levels, thereby decomposing the original data into a series of Intrinsic Mode Functions (IMFs) and a residue (Huang et al., 1998). The IMF possesses the following characteristics: (1) it exhibits the same number of zero-crossings and extrema as the original data, and (2) it possesses symmetric envelopes defined by the local maxima and minima, respectively. Taking the original data $x[n]$ as an example, this decomposition can be represented as follows:

$$x[n] = \sum_{i=1}^{I} IMF_i + R[n] \quad (1)$$

Herein, n represents the length of the signal，I denotes the total number of decomposed IMF components，IMF_i denotes i-th IMF，$R[n]$ is the residue obtained from the decomposition of the data $x[n]$.

The original EMD algorithm suffers from a problem known as mode mixing, which refers to an IMF consisting of oscillations with significantly disparate scales. This issue typically arises when a single IMF incorporates components with vastly different scales or when components with similar scales are distributed across

multiple IMFs (Lei et al., 2009). The problem of mode mixing not only results in the aliasing of the data signal in the time-frequency domain but also leads to the loss of physical interpretation of the IMFs (Wu and Huang, 2009). Several EMD-based methods have been developed to address this issue, such as Ensemble Empirical Mode Decomposition (EEMD) and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). The EEMD introduces white noise into decomposition, effectively adding a complete time-frequency space. Through multiple iterations, the added noise helps to counteract the effects of noise during the decomposition process, facilitating the natural separation of frequency scales and reducing the occurrence of mode mixing (Wu and Huang, 2009). However, in EEMD, when the number of ensemble trials is small, residual noise may still be present in the resulting IMF components. The interaction between the signal and noise can potentially generate additional modes. Furthermore, the high computational complexity of EEMD poses another challenge. In order to address these issues, Torres (Torres et al., 2011) introduced the CEEMDAN algorithm. The CEEMDAN improves upon EEMD by introducing a limited amount of adaptive white noise, effectively separating the IMF components and reducing residual noise, thereby minimizing reconstruction errors. This noise addition is optimized to strike a balance between noise suppression and preservation of signal features. By incorporating this adaptive noise, CEEMDAN achieves more efficient IMF separation while significantly reducing the computational requirements for the decomposition process. CEEMDAN is based

on the characteristic scales of extrema and is particularly sensitive to abrupt fluctuations in non-stationary signals. This makes it highly suitable for applications in environmental monitoring, such as water quality prediction (Wang et al., 2021).

The detail processes of CEEMDAN are as follows: (Torres et al., 2011):

Let us define the operator $Ej(\cdot)$, which, given a signal, produces the j-th mode obtained by EMD. Let wi be white noise with $\mathcal{N}(0, 1)$. If $x[n]$ is the targeted data, we can describe our method by the following algorithm:

1. decompose by EMD I realizations $x[n] + \varepsilon_0 w^i[n]$ to obtain their first modes and compute

$$\widetilde{IMF_1}[n] = \frac{1}{I}\sum_{i=1}^{I} IMF_1^i[n] = \overline{IMF_1}[n] \quad (2)$$

2. at the first stage ($k = 1$) calculate the first residue as in Eq. (1): $r_1[n] = x[n] - \widetilde{IMF_1}[n]$.

3. decompose realizations $r_1[n] + \varepsilon_1 E_1(w^i[n])$, $i = 1, ..., I$, until their first EMD mode and define the second mode:

$$\widetilde{IMF_2}[n] = \frac{1}{I}\sum_{i=1}^{I} E_1(r_1[n] + \varepsilon_1 E_1(w^i[n])) \quad (3)$$

4. for $k = 2, ..., K$ calculate the k-th residue:

$$r_k[n] = r_{k-1}[n] - \widetilde{IMF_k}[n] \quad (4)$$

5. decompose realizations $r_k[n] + \varepsilon_k E_k(w^i[n])$, $i = 1, ..., I$, until their first EMD mode and define the (k+1)-th mode as

$$\widetilde{IMF_{(k+1)}}[n] = \frac{1}{I}\sum_{i=1}^{I} E_1(r_k[n] + \varepsilon_k E_k(w^i[n])) \quad (5)$$

6. go to step 4 for next k. Steps 4 to 6 are performed until the obtained residue

is no longer feasible to be decomposed (the residue does not have at least two extrema). The final residue satisfies:

$$R[n] = x[n] - \sum_{k=1}^{K} \widetilde{IMF_k} \tag{6}$$

with K the total number of modes. Therefore, the given signal $x[n]$ can be expressed as:

$$x[n] = \sum_{k=1}^{K} \widetilde{IMF_k} + R[n] \tag{7}$$

Eq. (5) makes the proposed decomposition complete and provides an exact reconstruction of the original data.

Sample Entropy

Sample entropy (SamEn) is a computational method based on entropy that improves upon the estimation of approximate entropy by eliminating self-matches (Richman and Moorman, 2000)。SampEn is the exact value of the negative average natural logarithm of conditional probability, which represents the probability of generating new patterns in a nonlinear dynamical system (Deng et al., 2021), It is primarily used for quantitative descriptions of the regularity and complexity of a system. A higher value of SampEn indicates a higher complexity of the time series, while a lower value indicates lower complexity.

The detail SE calculation processes as (Wang et al., 2021):

Define $X_{(n)} = x(1), x(2), \cdots, x(N)$ as a time series. The algorithm implementation of SE is as follows:

Step1: Mark off $(N - m + 1)$ sub-fragments named $X_m(i)$ from $X_{(n)}$

according to the m data points.

Step2: Calculate the distances between $X_m(i)$ and other $(N - m + 1)$ sub-fragments and select the largest distance value, written as $d[X(i), X(j)]$.

$$d[X_i, X_j] = \max_{k=0,\ldots,m-1}(|x(i+k| - x(j+k)|) \tag{8}$$

Step3: For the given $X_m(i)$, count the number of $j(1 \le j \le N - m, j \ne i)$ when $d[X(i), X(j)]$ is less than or equal to r and this number is written as B_i:

$$B_m^i = \frac{number\ of\ X(j)\ such\ that\ d[X_i, X_j] \le r}{N - m}, i \ne j \tag{9}$$

Step4: Calculate the mean value of B_i^m and record it as $B^m(r)$

$$B^m(r) = (N - m + 1)^{-1} \sum_{i=1}^{N-m+1} B_i^m \tag{10}$$

Step5: For the label k = m + 1, calculate $A^k(r)$ by repeating step 2 to 4.

Step6: According to the above calculation, the final sample entropy can be expressed as $SampEn(m, r, N)$:

$$SampEn(m, r, n) = -\ln \frac{A^k(r)}{B^m(r)} \tag{11}$$

In general, m = 1 or m = 2 and r = 0.1~0.25 SD, where SD represents the standard deviation of the original series.

Density-based spatial clustering of applications with noise

Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996) is a clustering method based on estimating the minimum density levels of samples. It clusters data points by considering the neighboring points and a threshold radius. Essentially, the DBSCAN algorithm identifies minimum density

regions that partition the data into different clusters within low-density areas (Schubert et al., 2017). DBSCAN requires the determination of two parameters: the minimum number of neighboring data points (minpts) and the radius (ε). Based on the neighboring points within different radius ranges, all data points can be classified into three categories: core points, boundary points, and noise. A data point is considered a core point if there are more than minpts data points within its radius range. If the points within the radius range of a core point are also core points, then the relationship between the two core points is called density directly reachable. Within the radius range of a core point, data points that are not core points are referred to as boundary points. The relationship between boundary points can be established through the transitivity of relationships between core points, which is known as density-reachable. Points that do not belong to any cluster, neither as core nor boundary points, are considered noise and are not part of any cluster. Fig. S1 illustrates the principal explanation diagram for minpts = 4 and ε as the radius range. In the figure, A is a core point, while B and C are boundary points, and N represents noise. A point is density-reachable to itself within the ε radius range, A and B have a density-reachable relationship, C and D have a density-connected relationship, and A, B, C, D, and N are not density-connected.

Figure captions

Fig. S1. Illustration of DBSCAN clustering principle. minpts = 4, ε as the radius of the circles. A, B, and the red points represent core points, while C and D are boundary points, and N denotes noise.

Fig. S2. Monthly data of estimation dataset (a – d) and analysis dataset (e – j) in main canal from 2017 to 2020. Points denote the average value and error bars stands for the value range of all sites in the same period.

Fig. S3. Spearman correlation analysis between different factors and Mantel test between the GPP analysis set and the estimation set at section-scale.

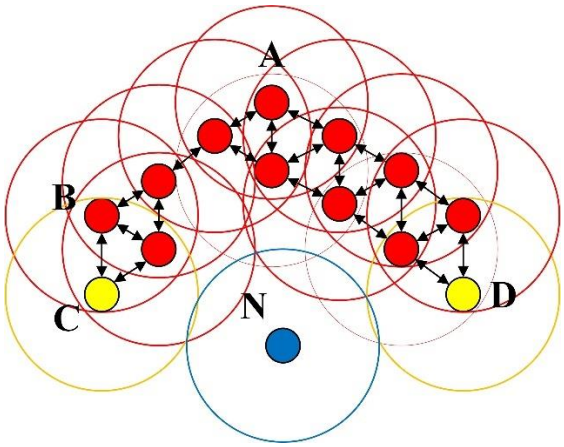


Fig. S1. Illustration of DBSCAN clustering principle. minpts = 4, ε as the radius of the circles. A, B, and the red points represent core points, while C and D are boundary points, and N denotes noise.

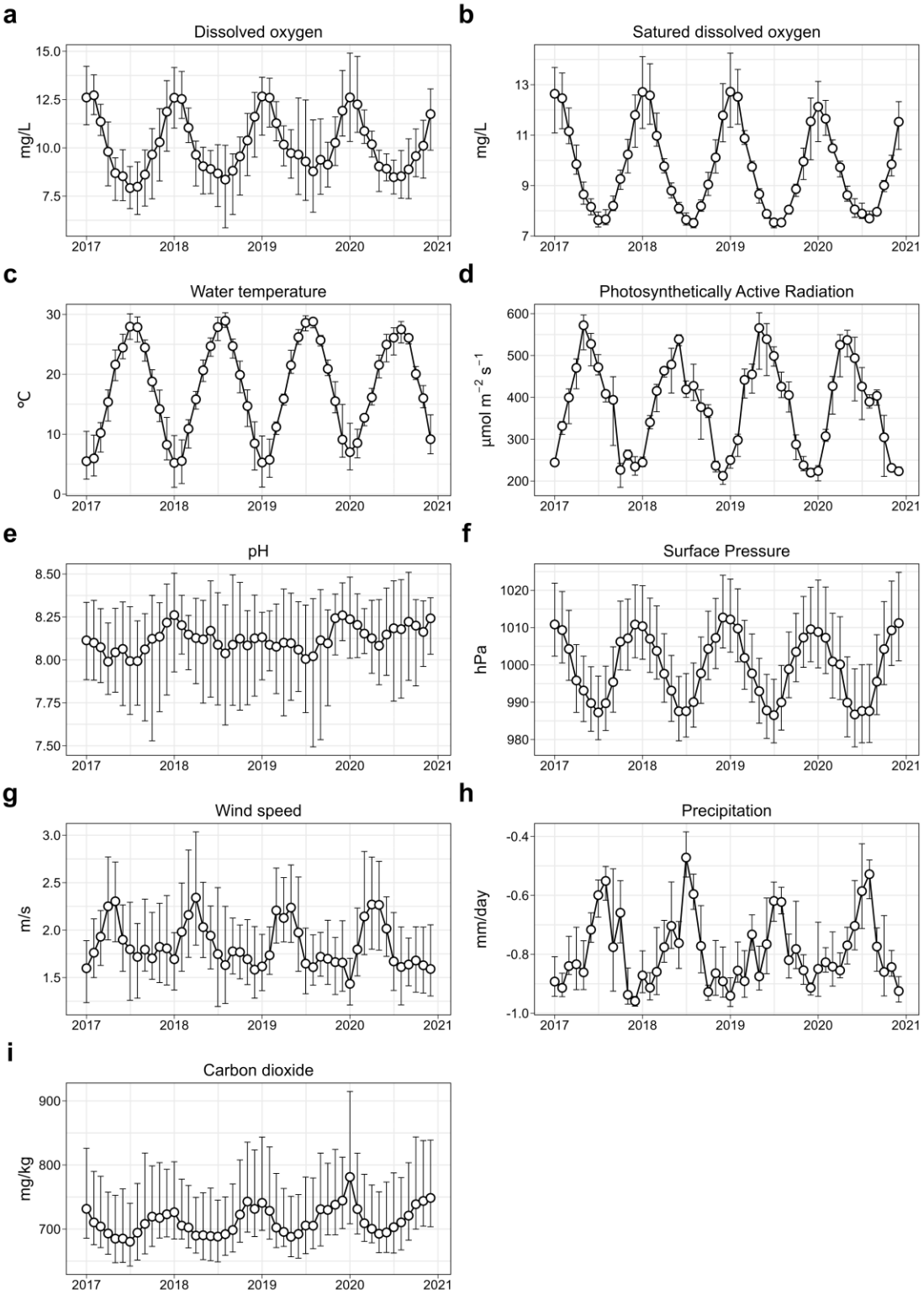


Fig. S2. Monthly data of estimation dataset (a – d) and analysis dataset (e – i) in main canal from 2017 to 2020. Points denote the average value and error bars stands for the value range of all sites in the same period.

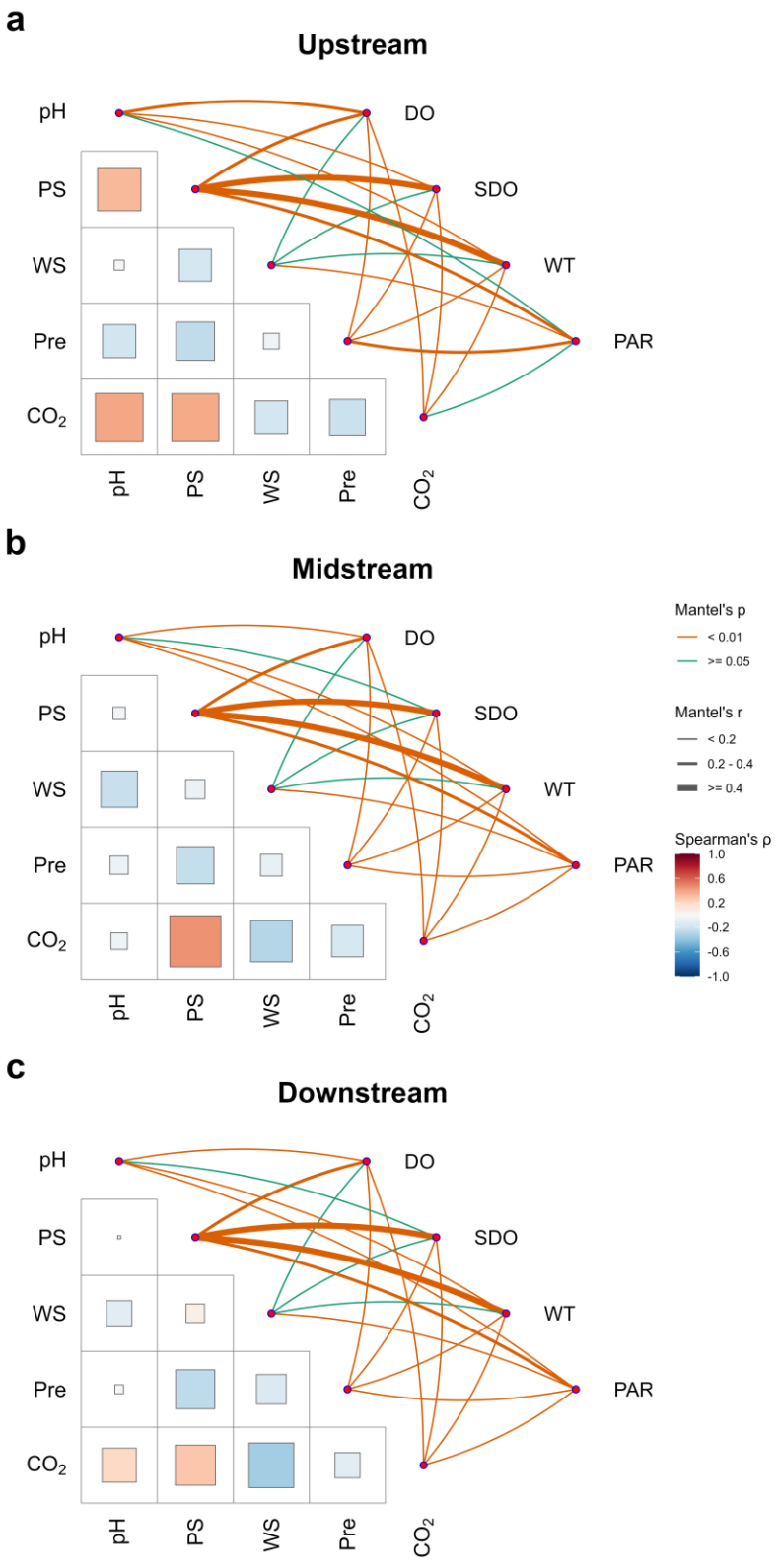


Fig. S3. Spearman correlation analysis between different factors and Mantel test between the GPP analysis set and the estimation set at section-scale.

Table S1

Detailed information and distances from the beginning station of all the main canal stations.

Canal sections	**Stations**	**Abbreviation**	**Designed Water depth of neatest sluice gates (m)**	**Distance away from the beginning station (km)**
Upper stream	Taocha	TC	8.15	0
	Jiangguo	JG	7.83	95
	Liuwang	LW	6.77	426
	Fuchengnan	FCN	8.54	524
Midstream	Zhanghebei	ZHB	5.96	731
	Nandaguo	NDG	6.03	837
	Tianzh	TZ	6.06	96

	uang		5	8
Down	Xiheis	XHS	4.7	11
stream	han		6	20
	Fenzhu	FZH	3.6	11
	anghe		1	72
	Zhong	ZYS	4.4	11
	yishui		0	94
	Huinan	HNZ	3.8	12
	zhuang		6	73

Table S2

Evaluation indicators and recommend values of model goodness-of-fit of the SEM.

Items	Abbreviation	Range	Recommend value
Chi-square minimum	χ^2	[0, +∞)	The less the better
Goodness of Fit Index	GFI	[0, 1]	>0.9 (excellent), >0.8 (acceptable)
Root Mean Square Error of Approximation	RMSEA	[0, +∞]	<0.08
Standardized Root Mean Square Residual	SRMR	[0, +∞]	<0.5

Table S3

Mann-Kendell test for GPP variations in different canal sections from 2017 to 2020.

Sections	Z value	P value
Entire main canal	5.223	$< 0.001^{***}$
Upper stream	2.476	$< 0.05^{*}$
Midstream	5.520	$< 0.001^{***}$
Downstream	6.913	$< 0.001^{***}$

Note: "***"represent significant in 0.001 level; "*" denotes data significant in 0.05 level.

Reference

Deng, C.N., Liu, L.S., Li, H.S., Peng, D.Z., Wu, Y.F., Xia, H.J., Zhang, Z.Q. and Zhu, Q.H. 2021. A data-driven framework for spatiotemporal characteristics, complexity dynamics, and environmental risk evaluation of river water quality. Science of the Total Environment 785, 17.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. 1996 A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C. and Liu, H.H. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 454(1971), 903-995.

Lei, Y., He, Z. and Zi, Y. 2009. Application of the EEMD method to rotor fault diagnosis of rotating machinery. Mech. Syst. Signal Proc. 23(4), 1327-1338.

Richman, J.S. and Moorman, J.R. 2000. Physiological time-series analysis using approximate entropy and sample entropy. Am. J. Physiol.-Heart Circul. Physiol. 278(6), H2039-H2049.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P. and Xu, X. 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Trans. Database Syst. 42(3), Article 19.

Torres, M.E., Colominas, M.A., Schlotthauer, G. and Flandrin, P. 2011 A complete ensemble empirical mode decomposition with adaptive noise, pp. 4144-4147.

Wang, J., Sun, X., Cheng, Q. and Cui, Q. 2021. An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting. Sci Total Environ 762, 143099.

Wu, Z. and Huang, N.E. 2009. Ensemble Empirical Mode Decomposition: a Noise-Assisted Data Analysis Method. Adv. Data Sci. Adapt. Anal. 1, 1-41.