Supplementary materials

S1. Distribution map of important cities

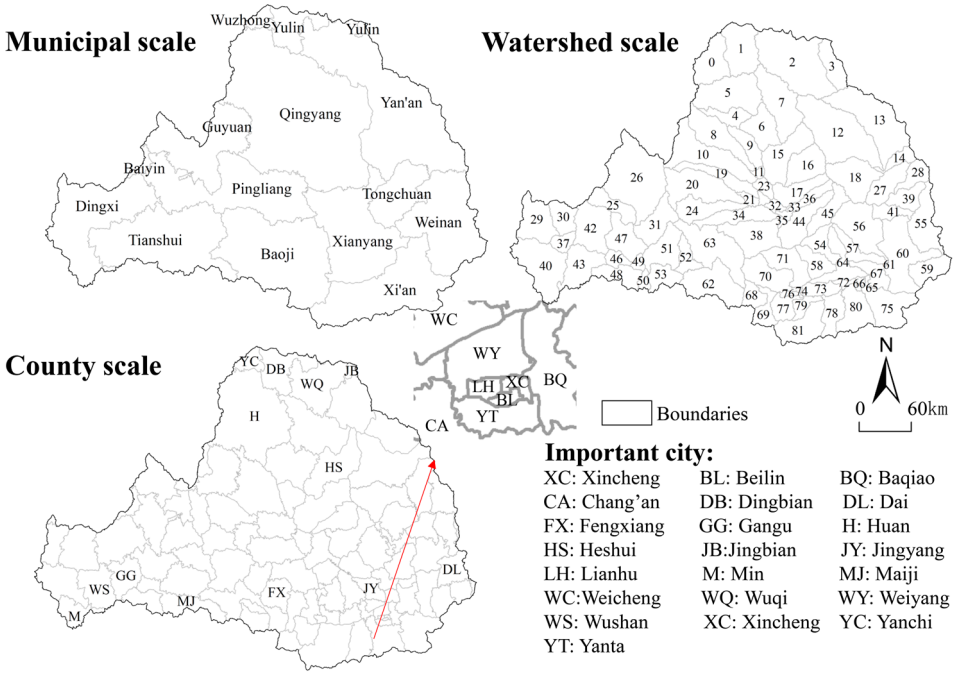

Figure S1. Distribution map of important cities and sub-watersheds at multiple scales.

S2. Calculation methods of ESSD

Table S1 provides the detailed calculation methods for the supply and demand of six types of ESs.

Table S1. Method for calculating the physical quantity of ESs.

ESs	Supply	Demand
FM	The FM supply reflects the capacity to reduce surface runoff and peak values [71, 72]: $FMS_i^t = P_i^t - Q_i^t$ $Q_i^t = \begin{cases} \frac{(P_i^t - 0.2S_{max,i}^t)^2}{P_i^t + (1-0.2)S_{max,i}^t}, P_i^t \geq 0.2S_{max,i}^t \\ 0, P_i^t < 0.2S_{max,i}^t \end{cases}$ $S_{max,i}^t = \frac{25400}{CN_i^t} - 254$ Where, FMS_i^t is the FM supply on the cell i in the t-th year (mm); Q_i^t is the surface runoff on the cell i in the t-th year; $S_{max,i}^t$ is the maximum soil water storage capacity on cell i in the t-th year; CN_i is an empirical parameter [74].	The construction land is densely populated and the urban assets are concentrated, thus there is a greater FM demand [73]. $FMD_i^t = \begin{cases} Q_i^t, in\ consturction\ land \\ 0, otherwise \end{cases}$ $Q_i^t = \begin{cases} \frac{(P_i^t - 0.2S_{max,i}^t)^2}{P_i^t + (1-0.2)S_{max,i}^t}, P_i^t \geq 0.2S_{max,i}^t \\ 0, P_i^t < 0.2S_{max,i}^t \end{cases}$ Where, FMD_i^t is FM demand on the cell i in the t-th year (mm); Q_i^t is the actual runoff on cell i in the t-th year (mm).
SR	SR represents the crucial function of natural landscapes in preventing sediment deposition [75]: $SRS_i^t = R_i^t \times K_i^t \times LS_i^t \times (1 - C_i^t \times P_i^t)$	The demand of SR is equivalent to the actual soil erosion as described in the RUSLE equation: $SRD_i^t = R_i^t \times K_i^t \times L_i^t \times S_i^t \times C_i^t \times P_i^t$

	SRS_i^t is the SR supply on the cell i in the t-th year (t). R_i^t, K_i^t, L_i^t, S_i^t, C_i^t, and P_i^t are respectively the precipitation erosion factor, soil erosion factor, slope length-gradient factor, vegetation management factor, and soil and water conservation measure factor.	SRD_i^t is the SR demand on the cell i in the t-th year (t). Other parameters are the same as those in SRS_i^t Equation.
WY	Compared to water retention, WY focuses more on the supply of water resources [76]: $$WYS_i^t = (1 - \frac{AET_i^t}{P_i^t}) \times P_i^t$$ Where, WY_i^t is the WY supply on the cell i in the t-th year (mm). AET_i^t is the actual evapotranspiration on the grid pixel i in the t-th year (mm).	The WY demand can be mainly classified into agricultural water, industrial water, and domestic water [17]: $$WYD_i^t = A_i^t \times Agr_i^t + G_i^t \times Ind_i^t + P_i^t \times Dom_i^t$$ WYD_i^t is the WY demand on the cell i in the t-th year (m^3). A_i^t, G_i^t, P_i^t are the cropland area, impervious surface area, and population of cell i in the t-th year. Agr_i^t, Ind_i^t, Dom_i^t are respectively the water consumption of agriculture, industry, and domestic use on the cell i in the t-th year.
WP	The output result of WP supply represents the total amount of nutrients retained by plants and soil, as well as the total amount of nutrients transferred through surface or groundwater flow [77]: $$WPS_i^t = load_i^t \times NDR_i^t$$ WP_i^t is the nitrogen and phosphorus output on the cell i in the t-th year (t). $load_i^t$ is the nutrient load. NDR_i^t is the nutrient transport ratio function for the uphill area and the downhill flow path.	The product of the allowable concentration of water pollution in the study area and the water discharge grid cells can be regarded as the allowable emission volume of pollutants [78]. Based on the third-class water quality standard for surface water environment, the allowable water pollution concentration values are set (the content of nitrogen should be less than or equal to 1 mg/L, and the content of phosphorus should be less than or equal to 0.2 mg/L).
CP	Relevant studies have shown that the crop production capacity had a positive correlation with the NDVI of the area [79]: $$CPS_i^t = G_{sum}^t \times \frac{NDVI_i^t}{NDVI_{sum}^t}$$ CPS_i^t is the crop supply on the cell i in the t-th year (kg). G_{sum}^t is the total crop yield. $NDVI_i^t$ is the NDVI value of the cultivated land. $NDVI_{sum}^t$ is the total NDVI value in the watershed.	The demand of CP is reflected by multiplying the per capita crop consumption by the population density [80]: $$CPD_i^t = D_{Crop,i}^t \times \rho_{pop,i}^t$$ CP_i^t is the crop demand on the cell i in the t-th year (kg). $D_{Crop,i}^t$ is the per capita crop consumption on the cell i in the t-th year. $\rho_{pop,i}^t$ is the population density on the cell i in the t-th year.
CS	CS is calculated by the InVEST model: $$CSS_i^t = C_{above,i}^t + C_{below,i}^t + C_{soil,i}^t + C_{dead,i}^t$$ CSD_i^t is the total carbon storage on the cell i in the t-th year (t). $C_{above,i}^t$, $C_{below,i}^t$, $C_{soil,i}^t$ and $C_{dead,i}^t$ are the carbon densities of aboveground, underground, soil and dead organic matter on the grid pixel i in the t-th year.	The CS demand is defined as the product of population density and per capita carbon emission [81]. $$CSD_i^t = D_{per,i}^t \times \rho_{pop,i}^t$$ CSD_i^t is the CS demand on the cell i in the t-th year (t). $D_{per,i}^t$ is the per capita carbon emission. $\rho_{pop,i}^t$ is the population density.

S3. Model validation

FM module: According to the data provided in the Shaanxi Province Water Resources Bulletin, we conduct a comparative assessment of the model using the runoff depth of the province as the object. From the simulation results of the SCS-CN model (see Table S2), it can be seen that the average relative error between the simulated values and the measured values is 7.51%. This error level indicates that the simulation accuracy of this model is high and the simulation effect is good.

Table S2. Comparison of simulation results of the SCS-CN model.

Runoff depth	2005	2010	2015	2020	2023
Shaanxi Province	210	235	150.4	187.6	247.4
Simulation values	224.7	248.9	162.9	172.6	226.8
Relative error	7.00%	5.92%	8.31%	8.00%	8.33%

SR module: LS is a topographic factor calculated by the soil erosion model topographic factor calculation tool (2.0) provided by the National Earth System Science Data Center. C is the vegetation cover and management factor, dimensionless, calculated according to the method proposed by Cai et al. (2000) [82]. P is the conservation support practice factor, dimensionless, calculated referring to Shao et al. (2023) [83].

WY module: According to the water resources bulletin (Gansu, Shaanxi and Ningxia Provinces), the total water resources in the Wei River Basin of the 14 cities are $1.58 \times 10^{11} m^3$ at 2005; $1.56 \times 10^{11} m^3$ at 2010; $1.25 \times 10^{11} m^3$ at 2015; $2.05 \times 10^{11} m^3$ at 2020 and $1.87 \times 10^{11} m^3$ at 2023. After repeated verification by this paper. The optimal Z parameters for the WY model are 9.5, 14.5, 12.1, 13.3, and 11.0 from 2005 to 2023.

CP module: We employed the spatial aggregation comparison method, which aggregated the results of the pixel-scale model to the county scale and compared them with the official statistical yields to evaluate the reliability of the model in estimating the regional total output. We aggregated the pixel CP values for each year from 2005 to 2023 to the county administrative units to obtain the "estimated total output" for each county. From the provincial and municipal statistical yearbooks of the corresponding years, we extracted the total crop output of each county as the benchmark value for the actual statistical output. We compared the data of 96 counties in the study area over 5 years, totaling 480 sample points. The main results are as follows:

Table S3. The CP model estimates the overall verification accuracy of the product at the county scale.

Precision indicators	Numerical value	Explanation and Interpretation
Coefficient of determination (R^2)	0.87	The model explains 87% of the temporal and spatial variations in the statistical output, with a high degree of goodness of fit.
Root Mean Square Error RMSE)	2.3×10^4 tons	The typical deviation between the estimated value and the statistical value is 2.3×10^4 tons.
Mean Absolute Error (MAE)	1.7×10^4 tons	The absolute error for each sample on average is 1.7×10^4 tons.
Mean Absolute Percentage Error (MAPE)	15.6%	The average relative error is 15.6%.

WP and CS module: The biophysical parameters for water purification and carbon sequestration services are mainly based on previous studies [84,85]. The biophysical parameters are as follows:

Table S4. Biophysical parameters in the WP module.

Land Cover	Cultivated land	Forest	Grassland	Waters	Construction land	Unused land
lucode	1	2	3	4	5	6
load_p	3.57	1.36	0.93	0	2.1	0.79
eff_p	0.48	0.67	0.6	0.4	0.26	0.26
crit_len_p	150	150	150	150	150	150
load_n	20	1.4	4	0	6.3	6.3
eff_n	0.25	0.4	0.35	0.02	0.05	0.05
crit_len_n	25	70	40	5	5	5
proportion_ subsurface_n	0	0	0	0	0	0

Table S5. Carbon density coefficients for different land use types in the CS module.

Land Cover	Carbon density (t/hm^2)			
	C_{above}	C_{below}	C_{soil}	C_{dead}
Cultivated land	4.98	23.44	31.47	2.83
Forest	12.33	33.68	46.12	4.08
Grassland	10.16	25.12	29.04	2.18
Waters	0.09	0	0	0
Construction land	0.72	7.97	5.1	0
Unused land	0.39	0.1	6.29	0

S4. Sensitive analysis of the inverse-distance weighting

To examine the sensitivity of the DLTCCD model to the distance decay parameter p in the inverse distance weight matrix, we selected five values: $p = 0.5,1.0,1.5,2.0$ (baseline), and 3.0. For each p, we recalculated the DLTCCD for all counties and evaluated four indicators: the mean DLTCCD, the standard deviation of DLTCCD, the global Moran's I (based on the inverse distance weight matrix with $p = 2$ to maintain a consistent spatial neighborhood definition), and the root mean square error (RMSE) relative to the baseline case of $p = 2$. By comparing these indicators across the five p values, we assessed the stability of the model's overall level, dispersion, spatial clustering pattern, and consistency with the baseline results.

Table S6. Sensitivity analysis results of DLTCCD under different distance.

p value	Mean of DLTCCD	Std. dev. of DLTCCD	Global Moran's	RMSE (relative to $p = 2$)
0.5	0.612	0.087	0.523	0.024
1.0	0.608	0.084	0.531	0.011
1.5	0.605	0.082	0.536	0.004
2.0	0.604	0.081	0.538	0.000
3.0	0.602	0.079	0.540	0.009
0.5	0.612	0.087	0.523	0.024

Note: Values in the table are illustrative; please replace them with actual results. The baseline case is $p = 2$. Global Moran's I is calculated based on the inverse distance weight matrix ($p = 2$), and all I values are significant at the $p < 0.01$ level. RMSE is computed for the DLTCCD values of all counties.

S5. Robustness analysis of the XGBoost model

To test whether the constructed XGBoost model has overfitting issues and to comprehensively evaluate its predictive robustness, we conducted a systematic verification from multiple perspectives. All validations were independently carried out for the dataset at the watershed scale and the county scale, as follows:

1. Five-fold cross-validation: During the hyperparameter optimization stage, the training set was randomly divided into 5 subsets, and the 4 subsets were used for training and the 1 subset for validation in turn. The R^2 values of each fold were recorded and the mean and standard deviation were calculated. This method can assess the stability of the model on different data subsets.
2. Independent test set verification (random 80/20 split): The full sample was randomly divided at a 80% training and 20% testing ratio. The test set did not participate in any training or hyperparameter adjustment during the entire modeling process. The model was judged for severe overfitting (large performance gap between the training set and the test set) by the R^2 and RMSE of the test set.
3. Random split stability test (repeated 20 times): The above random 80/20 division was repeated 20 times (using different random seeds each time), and the mean and standard deviation of the test set R^2 for each test were calculated. A low standard deviation indicates that the excellent performance of the model is not caused by a specific data partition by chance.
4. Noise injection test: Gaussian noise with an amplitude of 5% of the feature standard deviation was added to the features of the training set, and the model was retrained. The R^2 was evaluated on the original noise-free test set. By comparing the decrease in test set R^2 (threshold set at <5%), it was determined whether the model overly relied on individual features, thereby verifying its robustness.

The results of the above-mentioned verifications are summarized in Table S7. The specific values and analyses are as follows.

Table S7. The robustness validation results of the XGBoost model.

Verification type	Evaluation Metric	Result (watershed/county)	Description
Five-fold cross-validation	Mean R^2	0.9211/0.9361	Average goodness-of-fit from 5-fold CV during hyperparameter selection.
Independent test set (random 80/20 split)	R^2	0.9373/0.9640	Test set R^2 close to CV mean, no severe overfitting observed.
	RMSE	0.2630/0.2001	
Random split stability (20 repetitions)	Mean test R^2	0.9154/0.9176	Stable performance across different random seeds
	Standard deviation of test R^2	0.0184/0.0221	Very low variability, results are not coincidental.
Noise injection test (5%)	Test R^2	0.8965/0.9234	Approximately 4.6%/4.4% drop compared to original test R^2 (0.9373/0.9640)
	Performance drop	< 5%	The model does not overly rely on specific features and has good robustness.

S6. Simulation accuracy of the XGBoost model

Figures S2 and S3 showed the simulation accuracy of the XGBoost model at two different scales.

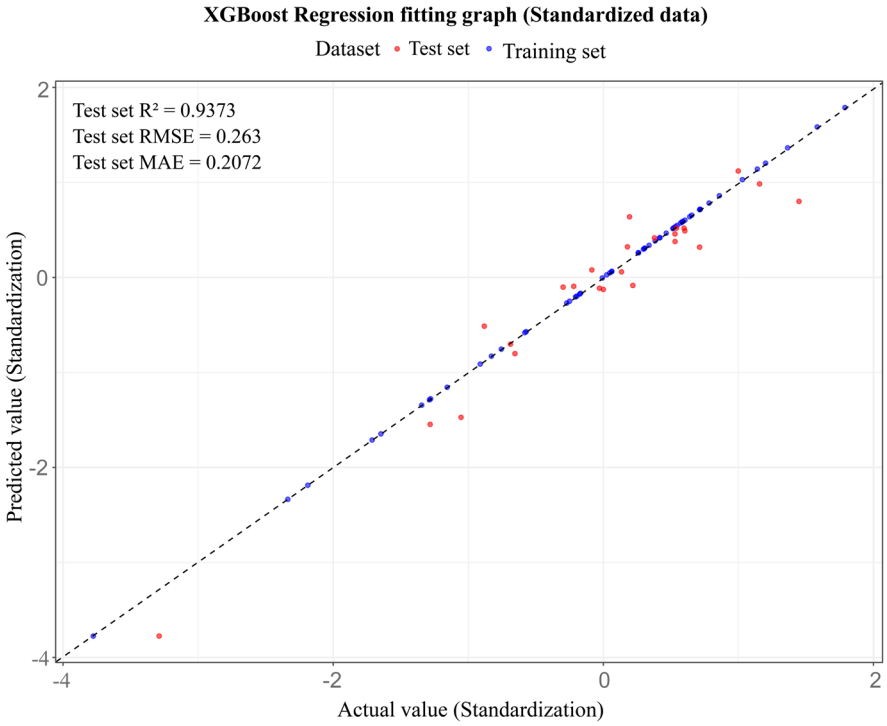

Figure S2. The simulation accuracy of the XGBoost model at the watershed scale.

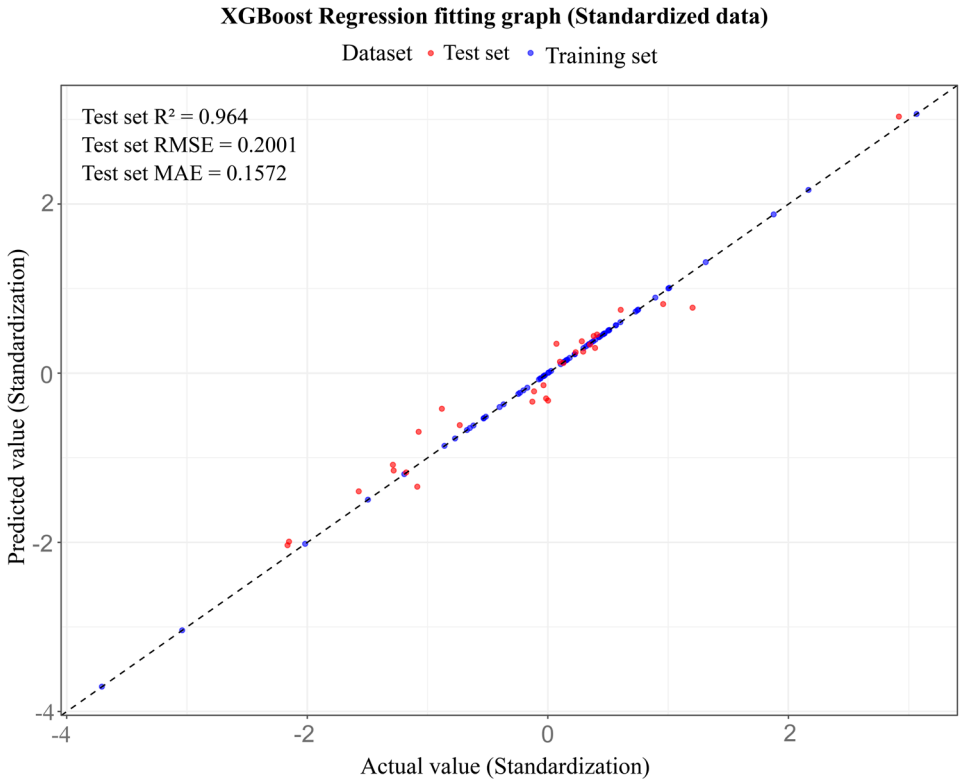

Figure S3. The simulation accuracy of the XGBoost model at the county scale.