SUPPLEMENTARY MATERIAL

Revealing Spatiotemporal Evolution and Driving Mechanisms of Grey Water Footprint in Land Consolidation Areas Using Explainable Machine Learning Models: Evidence from Yan'an Region, Shaanxi Province

Qiaoyang Yang[abc], Hui Qian[abc*], Qi Long[abc], Yicheng Duan[abc], Zhiming Cao[abc]

a. School of Water and Environment, Chang'an University, Xi'an, 710054, Shaanxi, PR China

b. Key Laboratory of Subsurface Hydrology and Ecological Effect in Arid Region of the Ministry of Education, Chang'an University, Xi'an, 710054, Shaanxi, PR China

c. Key Laboratory of Eco-hydrology and Water Security in Arid and Semi-arid Regions of Ministry of Water Resources, Chang'an University, Xi'an, 710054, Shaanxi, PR China

*Corresponding Author: Hui Qian. E-mail(s): qianhui@chd.edu.cn

List of Texts

Text S1. Definition of greywater footprint.

Text S2. Methodology for GWF calculation.

Text S3. R^2, mean absolute error (MAE), and root mean square error (RMSE).

Text S4. Multiple Linear Regression

List of Figures

Figure S1. Schematic diagram of regional Gini coefficient.

List of Tables

Table S1. Values of parameters for livestock and poultry breeding grey water footprint.

Table S2 Fitting results of the multiple linear regression model.

Text S1. Definition of greywater footprint

In conventional hydrology, the term "grey water" typically refers to relatively clean wastewater originating from domestic sources such as bathrooms, sinks, and washing machines. This definition is primarily based on the classification of wastewater according to its source and cleanliness. However, in this study, the term "grey water" deviates from the traditional hydrological definition. It is used in strict accordance with the theoretical framework of the grey water footprint (GWF), as first proposed by Hoekstra et al. (2011). In this context, "grey water" refers to "the volume of freshwater required to dilute pollutants in water bodies to a predetermined water quality standard." The focus here is on quantifying the degree of water pollution by measuring the volume of freshwater that must be contaminated for dilution to meet water quality standards. It should be noted that the GWF serves as an indicator for assessing the degree of water pollution, rather than a method for treating it.

Text S2. Methodology for GWF calculation

Agricultural GWF

① GWF derived from planting. This study accounted solely for nitrogen fertilizer in the calculation of the agricultural GWF. This is due to nitrogen fertilizer is used in much larger quantities than other types, and the water required to dilute its pollution is typically sufficient to cover that of other fertilizers. Moreover, phosphate fertilizers readily react with soil minerals to form insoluble compounds, while potassium ions from potassium fertilizers are easily adsorbed by soil particles. As a result, their contributions are not considered in the calculation. Considering that chemical fertilizers do not contribute to COD pollution and that chemical pesticides exhibit considerable variation in both type and permissible concentration limits, nitrogen was chosen as the key indicator in the GWF assessment of agricultural cultivation in this study. The calculation formula is expressed as follows:

$$GWF_{pla} = \frac{\alpha \times Appl}{C_{max} - C_{nat}} \tag{S1}$$

Where α denotes the nitrogen leaching rate, defined as the mass percentage of nitrogen

that enters the water body, and is set at 10% based on previous studies; *Appl* represents the amount of nitrogen fertilizer applied in the current year that contributes to nitrogen pollution in wastewater; C_{max} is the maximum acceptable concentration of pollutants according to a given water quality standard, whereas C_{nat} represents the background concentration of pollutants in the natural water environment.

② The GWF from livestock and poultry breeding is mainly attributed to COD and TN in the waste. Representative livestock, including pigs, cattle, sheep, and poultry were selected in this study. The feces and urine produced during their breeding were regarded as the main pollution sources in the agricultural GWF. To prevent double counting, livestock and poultry are categorized and quantified according to their respective annual feeding cycles: (i) the number of cattle and sheep with a cycle of one year or more was counted as year-end stock; (ii) the number of poultry and pigs with a feeding cycle of less than one year was selected as the year-end slaughtering quantity. Given the high levels of total nitrogen and COD present in livestock and poultry manure and urine, and considering that both pollutants can be diluted concurrently in receiving water bodies, the GWF of them is calculated based on the maximum value between the two. The detailed calculation formula is presented as follows:

$$GWF_{bre} = max\left\{GWF_{bre}(COD),\ GWF_{bre}(TN)\right\} \tag{S2}$$

$$GWF_{bre(i)} = \frac{L_{bre(i)}}{C_{max} - C_{nat}} \tag{S3}$$

$$L_{bre(i)} = \sum_{h=1}^{4} N_h \times D_h \times \left(f_h \times P_{hf} \times \beta_{hf} + u_h \times P_{hu} \times \beta_{hu} \right) \tag{S4}$$

where $GWF_{bre(i)}$ refers to the *GWF* of *i* (COD or TN) pollutant in the livestock and poultry breeding in agriculture; $L_{bre(i)}$ stands for the pollution emissions of *i* pollutant in the livestock and poultry breeding; *h* represents livestock and poultry (cattle, sheep, pigs and poultry); N_h is the number of feeding *h*; D_h signifies the feeding cycle; f_h and u_h denote the daily excretion of feces and urine, respectively; p_{hf} and p_{hu} indicate the pollutant content per unit of feces and urine; β_{hf} and β_{hu} represent the daily fecal output and urine output, respectively.

③ Total agricultural GWF. Because the water required to assimilate one pollutant can also assimilate others concurrently, summing the GWFs of different pollutants may lead to double counting. Therefore, the total agricultural GWF is calculated by first summing the GWF for the same type of pollutant, and then taking the maximum value among different pollutant types. The total agricultural GWF is calculated as shown in Eq. (S5):

$$GWF_{agr} = max\left\{GWF_{bre}(COD),\ GWF_{bre}(TN) + GWF_{pla}\right\} \tag{S5}$$

Industrial GWF

Industrial pollution mainly stems from designated wastewater discharge outlets and is considered point source pollution. According to previous studies, the main pollutants associated with industrial wastewater are COD and NH_3–N. Consequently, this study adopts these two parameters as the core indicators in the calculation of the industrial GWF:

$$GWF_{ind} = max\left\{GWF_{ind}(COD),\ GWF_{ind}(NH_3 - N)\right\} \tag{S6}$$

$$GWF_{ind(i)} = \frac{L_{ind(i)}}{C_{max} - C_{nat}} \tag{S7}$$

Domestic GWF

This study assessed the domestic GWF by selecting COD or NH_3–N as representative pollutants. The calculation method adopted was consistent with that used for the industrial GWF.

$$GWF_{dom} = max\left\{GWF_{dom}(COD),\ GWF_{dom}(NH3 - N)\right\} \tag{S8}$$

$$GWF_{dom(i)} = \frac{L_{dom(i)}}{C_{max} - C_{nat}} \tag{S9}$$

Regional total GWF

The total regional GWF comprises three factors: agricultural, industrial, and domestic GWF. It is represented by Eq. (S10):

$$GWF_{total} = GWF_{agr} + GWF_{ind} + GWF_{dom} \tag{S10}$$

Text S3. Model validation

Model performance was rigorously assessed using three statistical metrics: Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) (Huang et al., 2025). The higher the R^2 value, the stronger the model's explanatory power over the total variance of the dependent variable, which indirectly reflects its higher prediction accuracy. The lower the RMSE value, the higher the model's prediction accuracy, which also indicates its stronger effectiveness in capturing the distribution characteristics of the real data. Similarly, lower MAE values indicate more accurate predictions with smaller deviations from the observed values. The calculation formulas are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{S11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2} \tag{S12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right| \tag{S13}$$

where y_i and $\widehat{y_i}$ refer to actual and predicted values of the i-th sample, respectively; $\overline{y}$ is the mean of the actual values of all samples, and n is the total number of samples.

Text S4. Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical technique used to examine the linear relationship between a dependent variable and multiple independent variables. In this model, the dependent variable Y is assumed to be a linear combination X1, X2, X3, ⋯, Xn, expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \varepsilon$$

Where β_0 is the intercept term, β1, β2, …, βn are the regression coefficients,

representing the degree of influence of each independent variable on the dependent variable, and ε is the error term. Multiple linear regression is commonly used to predict target variables and analyze the relationship between independent and dependent variables. Its assumptions include a linear relationship between the independent and dependent variables, no multicollinearity among the independent variables, and the error term following a normal distribution. The results of the model are presented in Table S2.

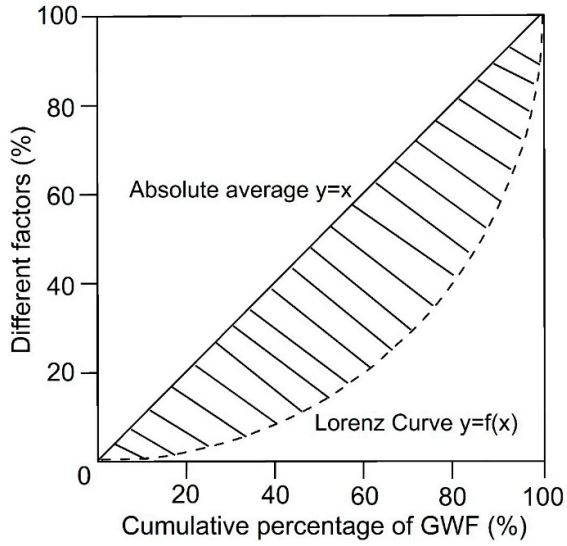


Figure S1. Schematic diagram of regional Gini coefficient.

Table S1 Values of parameters for livestock and poultry breeding grey water footprint.

Species of livestock		Cattle	Pig	Sheep	Poultry
Excrement	Raising cycle	365	199	365	55
	Daily excretion coefficient ($kg \cdot d^{-1}$ $head^{-1}$)	30	2.2	2.6	0.15
	The average amount of COD ($kg \cdot t^{-1}$)	31	52	4.63	45
	The average amount of TN ($kg \cdot t^{-1}$)	4.37	5.88	7.5	10.4
	Runoff fraction of COD (%)	6.16	5.58	5.5	8.59
	Runoff fraction of TN (%)	5.68	5.34	5.3	8.47
Urine	Daily excretion coefficient ($kg \cdot d^{-1}$ $head^{-1}$)	18	3.3	1	-
	The average amount of COD ($kg \cdot t^{-1}$)	6.0	9	4.63	-
	The average amount of TN ($kg \cdot t^{-1}$)	8	3.3	14	-
	Runoff fraction of COD (%)	50	50	50	-
	Runoff fraction of TN (%)	50	50	50	-

Table S2 Fitting results of the multiple linear regression model.

Model	R	R^2	Adjusted R^2	RMSE
	0.875	0.766	0.762	0.906

References

Huang, X.B., Liu, X.S., Jin, Y.H., Gao, X., Chen, Y.L. Identification and attribution analysis of integrated ecological zones based on the XGBoost-SHAP model: A case study of Chengdu, China. Ecol. Indic. **2025,** 177, 113787.

Hoekstra A.Y., Chapagain A.K., Aldaya M.M., Mekonnen M.M. The Water Footprint Assessment Manual: Setting the Global Standard. London, UK. **2011.**