

## ARTICLE OPEN



# Enhancing the explanation of household water consumption through the water-energy nexus concept

Zonghan Li<sup>1</sup>, Chunyan Wang<sup>1</sup>, Yi Liu<sup>1</sup> and Jiangshan Wang<sup>1</sup>

Estimating household water consumption can facilitate infrastructure management and municipal planning. The relatively low explanatory power of household water consumption, although it has been extensively explored based on various techniques and assumptions regarding influencing features, has the potential to be enhanced based on the water-energy nexus concept. This study attempts to explain household water consumption by establishing estimation models, incorporating energy-related features as inputs and providing strong evidence of the need to consider the water-energy nexus to explain water consumption. Traditional statistical (OLS) and machine learning techniques (random forest and XGBoost) are employed using a sample of 1320 households in Beijing, China. The results demonstrate that the inclusion of energy-related features increases the coefficient of determination ( $R^2$ ) by 34.0% on average. XGBoost performs the best among the three techniques. Energy-related features exhibit higher explanatory power and importance than water-related features. These findings provide a feasible modelling basis and can help better understand the household water-energy nexus.

npj Clean Water (2024)7:8; <https://doi.org/10.1038/s41545-024-00298-6>

## INTRODUCTION

The rapid increase in residential water consumption has garnered significant attention<sup>1</sup>. Due to the rapid growth of urban settlements, household water consumption is expected to escalate substantially by 2050, with a potential 300% increase in Asia and Africa<sup>2</sup>. Addressing the challenges posed by this surge necessitates a comprehensive understanding of household water consumption. Understanding factors that influence household water consumption can yield profound insights into household water use patterns and consumption, thereby facilitating improvements in infrastructure, facilities, and municipal planning and thus guaranteeing a stable supply capacity<sup>3,4</sup>. Furthermore, differentiated demand management and saving measures for different households could be implemented to reduce water usage based on these explanations<sup>5</sup>. However, the estimation and explanation of household water consumption remains one of the most formidable challenges in cities to date<sup>3</sup>.

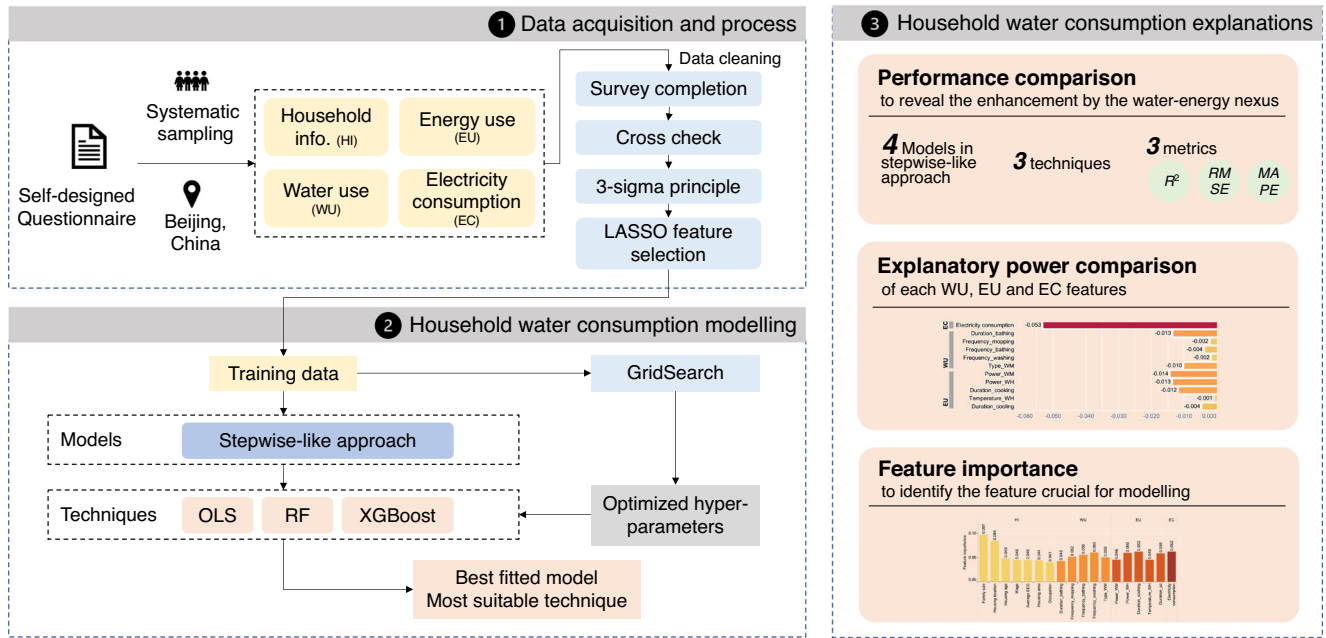
To enhance the explanation of household water consumption in a specific period (e.g., a year or a month), researchers have experimented with various assumptions regarding the possible influencing factors<sup>6</sup> (defined as features in this study). Commonly tested features include (i) water use-related features<sup>7–9</sup>, (ii) household demographic and economic information, such as population<sup>10,11</sup>, income<sup>12,13</sup> and education level<sup>14,15</sup>, and (iii) housing information, such as housing area<sup>13,16</sup> and housing type<sup>9,17</sup>. With these assumptions, various techniques, including traditional statistical techniques such as multiple regression based on the ordinary least squares (OLS) approach<sup>8,14</sup> and the autoregressive integrated moving average (ARIMA) method<sup>18,19</sup>, machine learning techniques such as artificial neural networks (ANNs)<sup>19,20</sup> and tree-based models<sup>21,22</sup>, etc., have been widely employed to model and explain household water consumption. Despite the extensive use of such a wide range of features and techniques, more than half of the variability in household water consumption remains unexplained, as indicated by an average coefficient of determination ( $R^2$ ) of less than 0.50 for most

models<sup>16</sup>. This unexplained variability is observed as residuals in the models. Identifying the appropriate features to characterize the unexplained variability has become a challenge in current household water consumption models.

Water and energy are frequently consumed simultaneously in households, particularly for behaviours such as laundry, bathing and culinary<sup>23</sup>. For instance, previous studies have revealed that up to 65.6% of household water consumption in Beijing, China is associated with energy use<sup>24</sup>, and 54.5% of total electricity consumption is associated with water use<sup>25</sup>. This interdependence between water and energy consumption could be described as the water-energy nexus, which is defined as the interconnection or cause-effect relationship between water and energy<sup>26–28</sup> in previous studies. When explaining household water consumption, the water-energy nexus concept can be applied by considering energy-related features in feature assumptions. However, most studies have either ignored the nexus or limited the concept to a few household appliances and behaviours (e.g., modelling the use behaviours of water heaters<sup>9,29</sup> and hot water consumption<sup>30,31</sup>) rather than considering it as a whole. The unexplored features, which cover major energy-consuming household appliances (e.g., washing machines, cooking devices and water heaters), are also significant water consumers in households<sup>32</sup>. Due to the challenges of data collection at the household level, few studies have considered these features in the form of energy use or energy consumption in household water consumption explanation models. Within this context, the feasibility of using energy use and consumption as suitable features for water consumption estimation models should be explored, and the results can be used to determine whether the water-energy nexus concept should be considered when explaining household water consumption.

In essence, the water-energy nexus can potentially be used to explain more variability in household water consumption, as it may serve as a “proxy” for the residuals in models. Incorporating such features may be the key to further improving the explanatory

<sup>1</sup>School of Environment, Tsinghua University, Tsinghua, China. ✉email: wangchunyan@tsinghua.edu.cn



**Fig. 1** The workflow of this study.

power of water consumption estimation models, but they have yet to be explored in detail. Thus, this study aims to assess the importance of considering the water-energy nexus concept when explaining household water consumption. The nexus was considered based on energy use (EU, use patterns such as duration, frequency and energy-related household appliance parameters) features and electricity consumption (EC, total electricity usage, measured in kWh) features in the models. Following the workflow in Fig. 1, a stepwise-like approach modelling scheme was utilized to compare models with and without EU and EC features. Household-level data collected from 1320 surveyed households in Haidian and Tongzhou Districts in Beijing, China, in 2020 were utilized as a case study. Four annual water consumption models were established using the stepwise-like approach. The modelling process involved employing a traditional statistical technique, OLS, as well as machine learning techniques including random forest (RF) and extreme gradient boosting (XGBoost). The key influential features were then identified and discussed. Improvements in model evaluation metrics upon incorporating EU and EC features, differences in performance among different modelling techniques, and variations in the explanatory power of different features were anticipated to be observed. Additionally, key features that influence household water consumption were identified.

## RESULTS

### Descriptive statistics

To examine the benefits of considering the water-energy nexus for explaining household water consumption, a questionnaire survey was conducted in 2020 in Haidian and Tongzhou Districts in Beijing, China. The questionnaire was distributed to subdistricts in the study area and had 1320 responses (1257 valid responses were retained after data cleaning). The questionnaire contained 78 questions regarding household information (HI), EU, water use (WU) and EC, and the results were integrated into 24 features for modelling. The questionnaire in this study differed from previous surveys of household water and/or energy consumption<sup>15,17,33–35</sup> by offering a broader perspective that encompassed various factors, including appliances, behaviours, water consumption and energy consumption. It also specifically focused on the concurrent

use of water and energy. Details of the questionnaire and the sampling scheme are provided in the Data section in the Methods section.

For HI, the average family size of the sample households was 2.9 (SD = 1.1), with over 40% of the households having 3 family members. All sampled houses and apartments were located within the 6<sup>th</sup> Ring Road, with approximately 60% of all samples located outside the 5<sup>th</sup> Ring Road (*housing location*). The average *housing area* of the samples was 75.4 m<sup>2</sup>. Approximately 43% of the sampled households were located in buildings constructed after 2000 (*housing age*). The income of the respondents of the questionnaire (*income*) exhibited a pyramidal pattern, and approximately 94% of respondents were employed or retired (*occupation*).

For WU features, the daily mean frequencies of mopping, bathing and laundryclothes washing were 0.78, 0.64 and 0.35 (*frequency\_mopping*, *frequency\_bathing*, and *frequency\_laundry*) for the sample households, respectively. On average, baths lasted 0.23 hours (SD = 0.11, range [0, 0.75], *duration\_bathing*). The majority of sampled households used impeller washing machines (approximately 92%, *type\_WM*).

In terms of EU, the average powers of washing machines and water heaters were 0.54 kW (SD = 0.53 kW) and 7.27 kW (SD = 9.29 kW), respectively (*power\_WM* and *power\_WH*). The average durations of cooking and air conditioning (specifically in summer) were 0.36 hours and 1.71 hours per day, respectively (*duration\_culinary* and *duration\_ac*). For households with storage water heaters (approximately 53%), the average temperature setting for heating was 50.6 °C (*temperature\_WH*).

Household water and electricity exhibited strong heterogeneity. *Water consumption* was the label (or explained variable in traditional statistical techniques) of the models; it averaged 120.1 m<sup>3</sup>/year, with a standard deviation of 47.0 m<sup>3</sup>/year. For comparison, the statistical value of the city-wide average household water consumption in Beijing was 130.5 m<sup>3</sup>/year in 2020<sup>36,37</sup> (based on calculations). Similarly, *electricity consumption* displayed relatively high dispersion, as its mean value and standard deviation were 2351.5 kWh/year and 941.0 kWh/year, respectively. More comprehensive descriptive statistics of the main contents of the questionnaire are shown in Supplementary Tables 1, 2 and Supplementary Fig. 1.

To comprehensively assess the effect of considering the water-energy nexus concept in terms of explaining household water consumption, a stepwise-like approach was designed to compare the contributions of different groups of features, and four models were established by using different feature combinations. In addition to HI, which was considered in all four models, Model (1) used WU, Model (2) contained EU, Model (3) was based on WU and EU, and Model (4) included WU, EU and EC. The following evaluations were conducted: (1) performance comparisons of the four models, (2) comparisons of the explanatory power of each EU, WU, and EC feature individually, and (3) identification of key features crucial for modelling household water consumption.

### Model performance comparisons

The performance of Models (1) to (4) established based on the OLS, RF and XGBoost techniques is summarized in Table 1.

The rows in Table 1 compare the performance of different techniques. Generally, machine learning techniques outperform the OLS technique. Specifically, the XGBoost technique exhibits a higher ability to explain variance in household water consumption than RF, indicating better performance. For instance, in Model (4), compared with those of the OLS and RF techniques, the average  $R^2$  value of the XGBoost technique is 62.5% (increasing from 0.32 to 0.52) and 4.0% (increasing from 0.50 to 0.52) higher, respectively; the average  $RMSE$  value decreased by 16.4% and 1.6%; and the average  $MAPE$  value was reduced by 19.4% and 10.7%.

The columns in Table 1 are different models based on the stepwise-like approach, and the enhancement achieved by considering the water-energy nexus concept is assessed. With the four models established based on the XGBoost technique as an example, all three metrics for Model (2) are better than those for Model (1), indicating that the selected EU features provide a better explanation of household water consumption than does WU. After the EU features were added to Model (1) to create Model (3), the average  $R^2$  value increased by 12.2% (from 0.41 to 0.46), the average  $RMSE$  value decreased by 4.8%, and the average  $MAPE$  was reduced by 7.1%. Building upon Model (3), Model (4) incorporated the EC feature. In this case, the average  $R^2$  value increased by 13.0% (from 0.46 to 0.52), and the average  $RMSE$  and  $MAPE$  values decreased by 5.1% and 3.8%, respectively.

Notably, compared to Model (1), Model (4) exhibited an increase in  $R^2$  from 0.33 to 0.45 (by 30.4%; average of the three techniques), highlighting the significant explanatory power of energy-related

features in household water consumption. In other words, incorporating the water-energy nexus concept considerably enhanced the explanatory power of the models.

### Explanatory power of each WU, EU and EC feature

To further assess the contributions of EU and EC features to the models' performance, the explanatory power of each WU, EU and EC feature was examined. The fitting of Model (4) using the XGBoost technique was repeated with one WU, EU or EC feature removed in each case. By quantifying the changes in  $R^2$ ,  $RMSE$  and  $MAPE$  when each feature was removed, the explanatory power of EC and EU in the full set of features was compared. The results are illustrated in Fig. 2.

Removing EC from the model had a negative effect on  $R^2$ ,  $RMSE$  and  $MAPE$  (−0.053, 1.786 and 0.015, respectively), with magnitudes greater than those observed for other features. Specifically,  $R^2$  value was reduced by 10.2%. This reveals the strong explanatory power of EC and suggests the indispensability of including EC in household water consumption modelling.

The average reduction in  $R^2$  when removing one EU feature was 0.008, which was higher than the reduction observed when removing one WU feature (0.006). This indicates that EU features may possess stronger explanatory power than WU features and are more crucial in explaining water consumption. This observation aligns with the finding that the performance metric values of Model (1) were worse than those of Model (2) across all techniques.

### Key feature identification

For machine learning techniques, the normalized importance of each feature can be calculated to reflect the level of importance in explaining household water consumption. The feature importance, which ranges from 0 to 1, is the node impurity value for  $t$  mean-squared error in the model. In this study, the feature importance of the best-fitting Model (4) established based on the XGBoost technique was calculated. The importance of each feature was compared, and the features with the highest levels of importance (Fig. 3), which are referred to as the "key features", were identified. According to the results, *family size* was the most important feature in the model, contributing 0.10 to the total feature importance (total importance = 1). *Housing location* was also a vital HI feature, accounting for 0.08 of the total importance.

Among all 10 WU and EU features, behavioural features (*frequency* and *duration*) displayed slightly greater feature importance (0.06 on average) in relation to household water consumption than did household appliance features (*power* and *sort*, 0.05). Consistent with the findings in previous subsections, the cumulative importance of EU features (0.27) exceeded that of WU features (0.26). EU features with importance greater than 0.05 included *power\_WH*, *duration\_culinary* and *duration\_ac*, and WU features with importance exceeding 0.05 included *frequency\_landlury*, *frequency\_bathing*, and *frequency\_mopping*. In addition to the conclusion that considering EC in the model significantly improved the model fit, EC was also an important feature in the XGBoost Model (4), contributing 0.06 to the total importance.

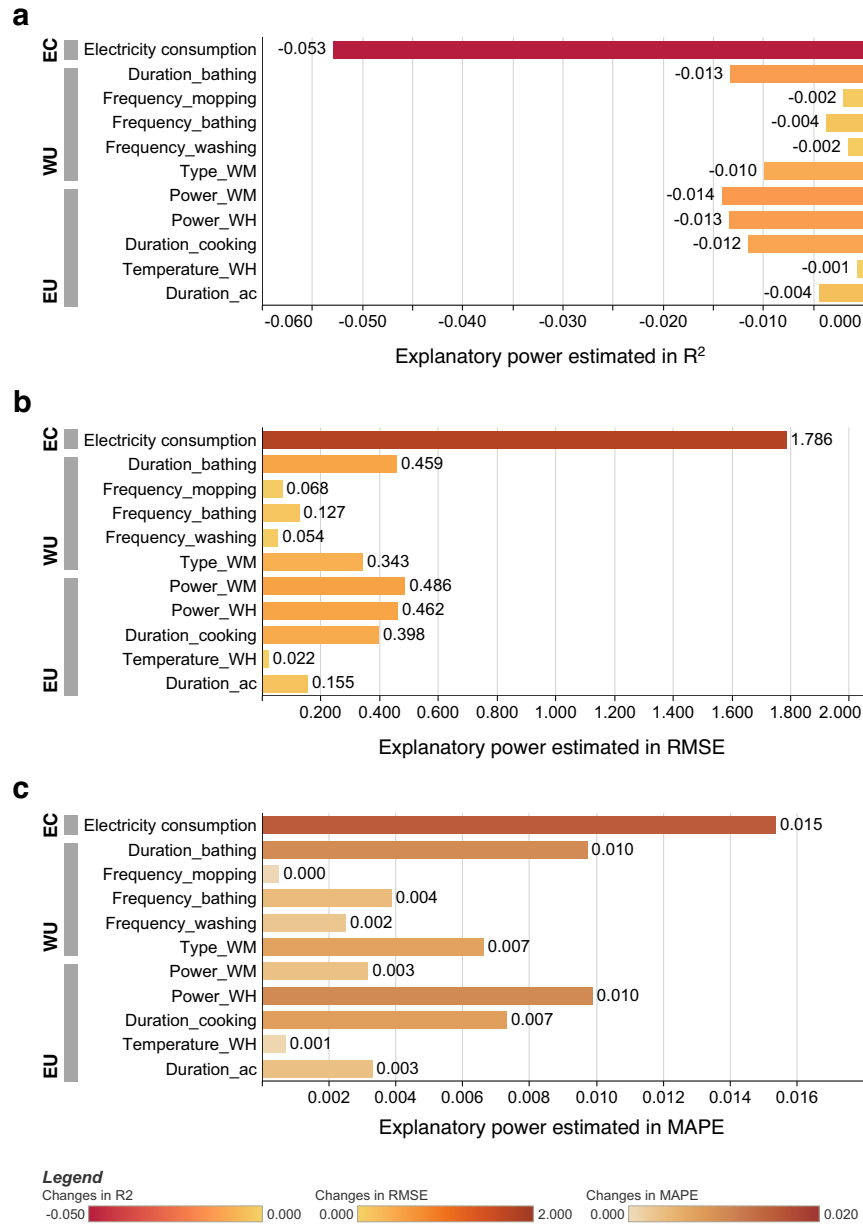
### DISCUSSION

In this study, the water-energy nexus concept was introduced to household water consumption models. A case study was conducted using a dataset of 1320 samples (1257 valid) collected in Beijing, China, in 2020, which included four groups of features: HI, WU, EU, and EC. Compared to existing models with similar sample sizes, the XGBoost model improved the explained variance by at least 23.8%. The explanation of household water consumption was enhanced significantly, as supported by the findings through various evaluation approaches. In conclusion, the

**Table 1.** A comparison of model performance using different techniques.

		Model (1) Inputs: HI & WU	Model (2) Inputs: HI & EU	Model (3) Inputs: HI & WU & EU	Model (4) Inputs: HI & EU & WU & EC
OLS	$R^2$	0.25	0.26	0.29	0.32
	$RMSE$	41.78	41.56	40.87	39.98
	$MAPE$	0.33	0.32	0.32	0.31
RF	$R^2$	0.36 (0.34)	0.46 (0.44)	0.47 (0.45)	0.52 (0.50)
	$RMSE$	38.58 (39.00)	35.50 (35.99)	35.08 (35.47)	33.53 (33.97)
	$MAPE$	0.31 (0.31)	0.29 (0.30)	0.29 (0.29)	0.27 (0.28)
XGBoost	$R^2$	0.44 (0.41)	0.49 (0.46)	0.50 (0.46)	0.55 (0.52)
	$RMSE$	36.02 (37.01)	34.44 (35.46)	33.90 (35.22)	32.48 (33.44)
	$MAPE$	0.27 (0.28)	0.25 (0.26)	0.25 (0.26)	0.24 (0.25)

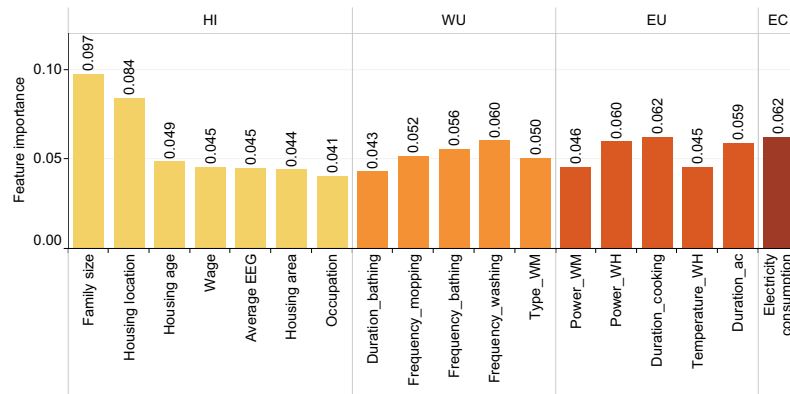
Values in the cells indicates the optimized value of each indicator. Values in the brackets denotes the average value of each indicator over 500 repeated runs of the RF and XGBoost models. The coefficients and their statistical significance for the OLS technique are reported in Supplementary Table 4.



**Fig. 2** The evaluation of explanatory power for each WU, EU and EC feature. **a** Explanatory power estimated based on  $R^2$ . The numbers indicate the decrease in  $R^2$  when removing a WU, EU or EC feature from the model with replacement. **b** Explanatory power estimated based on  $RMSE$ . The numbers indicate the increase in  $RMSE$  for each removal case. **c** Explanatory power estimated based on  $MAPE$ . The numbers indicate the increase in  $MAPE$  for each removal case. The results were estimated by deploying the XGBoost technique. The learning rate was constant in each case.

consideration of EU and EC features significantly increased the explanatory power of household water consumption, i.e., a 34.0% increase in  $R^2$ , an 8.8% decrease in  $RMSE$ , and an 8.7% decrease in  $MAPE$ . Furthermore, among all WU and energy-related features, EC feature exhibited the highest explanatory power (0.05 in  $R^2$ ) and feature importance (0.06). Compared to WU features, EU features demonstrated larger explanatory power (0.04 vs. 0.03 in  $R^2$ ) and feature importance (0.27 vs. 0.26). These findings offer a feasible modelling basis to investigate variance in household water consumption, thereby improving the modelling accuracy. This also provides a better understanding of the water-energy nexus at a household scale and thus facilitate a desired improvement of sustainable water supply and consumption.

Considering the modelling techniques, the emerging data-driven techniques significantly improve the performance of the household water consumption explanation models and display potential for broad application, as there are less stringent hypotheses regarding the statistical distributions of features than those in the OLS approach. Regression is widely applied in modelling to explain the consumption of various resources, such as electricity<sup>38</sup>, natural gas<sup>39</sup> and water<sup>14,40</sup>. To better understand the contribution of machine learning techniques to explain household water consumption, in this study, both traditional statistical (OLS) and machine learning techniques (RF and XGBoost) were used, and their goodness of fit, model performance results and explanatory power were compared. XGBoost was found to be the most suitable technique for modelling and



**Fig. 3 Feature importance of Model (4).** The importance was derived from the XGBoost technique.

explaining household water consumption with the selected features and the cross-sectional data. A possible reason may be the nonlinear impacts of HI, WU, EU and EC features on household water consumption<sup>41</sup> and the excellent ability to capture the nonlinear interactions between features and labels in the XGBoost technique<sup>42</sup>.

Despite the limited interpretability to some degree when using XGBoost, some insights can still be obtained from the feature importance results. With the two most important HI features as examples, an increase in *family size* directly leads to an increase in household water consumption. *Housing location* reflects information such as the construction age of the housing unit, which can have a significant impact on household water consumption. Moreover, household appliance features explain more of the variance in household water consumption than behavioural features. This suggests that future research can focus on residents' water and energy use behaviours and corresponding intervention measures to control household water consumption.

The explanatory power of household water consumption was improved by at least 0.10 (23.8%, from 0.42 to 0.52) compared with the values reported in previous studies that utilized cross-sectional data and similar sample sizes ( $1320 \pm 660$ ), as shown in Table 2. The Model (4) established using the XGBoost technique achieved an optimized  $R^2$  value of 0.55, with an average value of 0.52. In previous studies with similar sample sizes<sup>8,9,14,16,21</sup>, the highest  $R^2$  value only reached 0.42<sup>16</sup>, with an average value of 0.33. However, in studies with relatively small sample sizes,  $R^2$  values may exceed 0.50<sup>11,13,15</sup>. This study mainly focuses on studies with similar sample sizes in this comparison because a larger sample size generally leads to more accurate parameter estimates and model results that are closer to reality but lower  $R^2$  values<sup>43</sup>. Additionally, various measures were taken to ensure the robustness of the conclusions. First, as described in the Methods section, this study used a rigorous systematic sampling method to ensure the representativeness of the sample. Second, when designing the models, the core factors influencing household water consumption in previous studies<sup>44</sup> were referred. Third, the localized characteristics of the study area, such as using the Ring Road to distinguish household locations, were considered.

The inclusion of EU and EC features increased the  $R^2$  value by 0.11 (34.0% from 0.33 to 0.45), providing strong evidence supporting the necessity of considering the water-energy nexus when modelling household water consumption. The water-energy nexus serves as a proxy as model residuals and contributes to capturing a significant portion of the unexplained variability in traditional water consumption models (such as Model (1)). The evidence supporting the importance of the nexus can also be observed in the models. The EU features display larger explanatory power and feature importance on average than the WU features.

Based on the relationship between water and energy consumption, EC and EU features have strong explanatory power and can considerably improve model performance. The models and research roadmap have strong potential to be widely applied to other regions and at other temporal scales. Previous research and statistics have revealed a strong relationship between household water and energy consumption in various regions and at different temporal scales<sup>25,45–47</sup>. However, due to differences in the household population structure, water use behaviour, water consumption, and strength of the water-energy relationship, the explanatory power of EU and EC features and their impacts on model performance in different regions may vary. It should also be noted that for studies in other regions, the model needs to be retrained, and hyperparameters need to be reoptimized.

Some limitations remain in this study. First, the water consumption data used in this study are based on cross-sectional data. Future studies may include long time series of annual-scale water and electricity consumption data to verify the conclusions of this study based on the trend of water consumption over time. Second, the COVID-19 pandemic has altered the using behaviours of water and electricity<sup>48,49</sup>, potentially impacting the water-energy nexus and the explanatory power on household water consumption. Analysis on this impact is expected and required in future research. Third, this study aims to predict household water consumption. Further studies could explore the causal relationship between various features and water consumption based on different approaches, such as using interventions (for example, reshaping water consumption behaviours by providing feedback) and social experiments at the household scale.

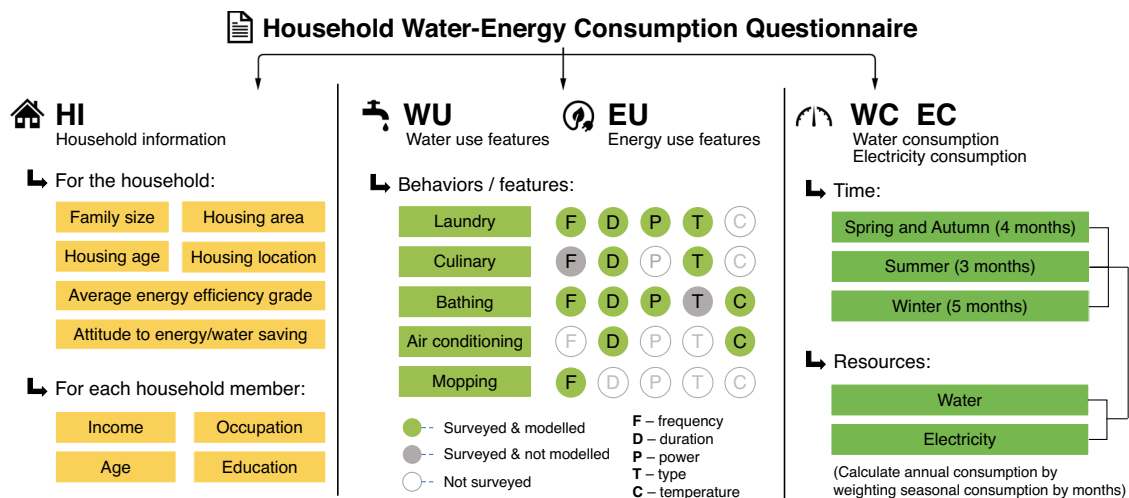
## METHODS

In this section, the data collection process and the methodology used to quantify the impacts of considering the water-energy nexus concept on explaining household water consumption are described. Broadly, this study consisted of three main steps. First, the HI, WU, EU, WC and EC data were collected via a self-designed questionnaire. The collected data underwent cleaning and cross-checking (i.e., determined if different responses to the same question matched, as detailed in the "Data collection and cleaning" section) processes to ensure data quality. Second, a stepwise-like approach was introduced to establish four models and explain household water consumption with different combinations of feature groups. Third, the enhancement associated with incorporating the water-energy nexus concept into models were verified. The enhancements were examined for OLS, RF and XGBoost techniques. The verification steps included model performance evaluation using cross-validation (i.e., evaluating

**Table 2.** Existing household water consumption models with different sample sizes.

Authors	Temporal scale	Technique(s)	Features	Studied period	Studied location	Sample size	R <sup>2</sup>
Bennett et al. <sup>20</sup>	Annually	ANN	HI, GEO, WU	2010	Queensland, Australia	205	0.30–0.41
Jeandorn et al. <sup>11</sup>	Daily	Logistic regression	HI, GEO, INF	2017	Uvira, Democratic Republic of the Congo	416	0.61
Gregory and Leo <sup>66</sup>	Annually	SEM	HI, WU	1996.7–1997.6	New South Wales, Australia	471	0.33
Singha et al. <sup>15</sup>	(not reported)	SEM	HI, ATT	2021	Fukuoka Prefecture, Japan	514	0.55
Lee and Derrible <sup>13</sup>	Daily	GBM, OLS, RF, SVM	HI, GEO, WB	2016	USA & Canada	531	0.33–0.69
Mostafavi et al. <sup>9</sup>	Daily	Stepwise regression	HI, WEA, EU, WU	2009	USA	771	0.12–0.24
Ito et al. <sup>14</sup>	Daily	OLS	HI, WEA	2015.1–2015.4 2015.12–2016.2 2016.8–2016.9	Kathmandu Valley, Nepal	992	0.26–0.35
Duerr et al. <sup>21</sup>	Monthly	ARIMA, BART, GBM, RF	GEO, WEA	1998–2010	3 counties in Florida, USA	973	–
Jayarathna et al. <sup>8</sup>	Quarterly	OLS	HI, WEA, CAL	2009–2011	Queensland, Australia	1214	0.29
THIS STUDY	Annually	OLS, RF, XGBoost	HI, WU, EU, EC	2019	Beijing, China	1257	0.32–0.52 (Model (4))
Bich-Ngoc et al. <sup>16</sup>	Annually	OLS	HI, WU	2014	Wallonia, Belgium	~2000	0.40–0.42
Hoşgör and Fischbeck <sup>10</sup>	Annually & Daily	OLS	HI, WEA, CAL	2009–2011	Gainesville, USA	7022	0.08–0.14

HI household information, INF infrastructure, ATT attitude, GEO geographic, WEA weather, CAL calendar feature, WU water use, EU energy use.

**Fig. 4** The structure of the household water-energy consumption questionnaire.

model performance by randomly splitting data into subsets for training and testing<sup>50–52</sup>), explanatory power clarification of each feature by comparing the changes of removing one feature from the model, and identification of key features by calculating feature importance.

### Data

The data used in this study were obtained from the “household water-energy consumption” questionnaire survey developed by the authors in 2020 using a sampling approach similar to that used by Li et al.<sup>53</sup> The questionnaire consisted of 78 items covering HI, EU, WU and EC features. These items were consolidated into 24 features for modelling (Fig. 4). The related items in the questionnaire are provided in Supplementary Note 4.

The HI part comprised 18 items aimed at collecting data on both households and individuals. Regarding households, information on living features, such as the location, area and age of the house, was collected. Family size was included as a demographic

feature. The average energy efficiency grade of household appliances and the attitude towards water and energy savings were surveyed to assess the water and energy savings level of each household. For individuals, data on age, occupation, income level, and education level were collected for all family members within the households. However, considering the information of all family members in the models may introduce many features that provide similar information, thereby increasing multicollinearity in the models. Consequently, only the information of the respondent of the questionnaire, as a representative of all family members, was included in the models.

The WC and EC parts were implemented using a form that captured the monthly water and electricity consumption in different seasons: summer (June to August), winter (November to March) and spring and autumn (April, May, September, and October). Respondents had the option to provide either consumption or cost information based on their bills. In cases in which cost was provided, the cost was converted to consumption using

local unit prices after the survey. The annual WC and EC were calculated by weighting the seasonal consumption by month. For the division of seasons, please refer to Supplementary Note 3.

The WU and EU parts consisted of 59 items that focused on household behaviours and appliances related to water-energy consumption. Behaviours such as doing laundry, cooking, bathing (including showers and baths), air conditioning and mopping were mainly considered in this part. Data on behaviour frequency, behaviour duration, appliance power, appliance type and (or) temperature were collected based on the specific nature of each behaviour (detailed in Fig. 4). However, not all collected features were utilized in the models. For instance, only households with electrical water heaters were used to provide temperature setting information, and households with gas water heaters did not have to answer the question (with a zero value). The type of water heater was inferred from the temperature of the bathing appliance. To avoid multicollinearity, the “type of water heater” feature was not included in the models.

Importantly, due to the utilization of cross-sectional data in this study, certain commonly used features, such as price<sup>54,55</sup>, calendar features<sup>10,56</sup>, temperature and rainfall<sup>57,58</sup>, were excluded from both the survey and the models, as they did not exhibit significant variance or did not change during the study period, following the practice of Bich-Nog<sup>16</sup>.

Prior to the survey, as a pilot study, the questionnaire was first distributed online to 15 researchers using convenience sampling. Based on their feedback, the wording of the items was improved, and the option settings for the family information items were adjusted (e.g., changing from collecting precise figures of family income to intervals of 50,000). The questionnaire was distributed in 21 subdistricts in Haidian District (“Jiedao” in Chinese, 22 in total), 4 subdistricts in Tongzhou District (4 in total) and 4 towns in Tongzhou District (8 in total) in 2020 (detailed in Supplementary Table 2). A systematic sampling approach was employed in the chosen subdistricts with a random starting point. A sample size of 660 was used in each of the Haidian and Tongzhou districts, resulting in a total sample size of 1320. To ensure the authenticity of the participants’ responses, interviewers with professional knowledge and extensive practical experience were invited to collect data through face-to-face interviews. Participants were informed about the objectives, purposes and procedures of the survey. They signed an informed consent form and were assured that the anonymity and confidentiality of the answers would be guaranteed. Participants were given the option to skip questions or discontinue the questionnaire if desired.

After conducting the survey, the collected data were processed with a 3-step validation process for cleaning, as follows: (i) verifying the validity of the questionnaire based on the completion level for each question, (ii) calculating WC based on WU and comparing the value with the answer to the question “total household water use”, and (iii) applying the 3-sigma principle to identify and remove samples with outlier water consumption values. After data cleaning, a total of 1257 samples out of the initial 1320 were used for model establishment. To avoid possible multicollinearity, the least absolute shrinkage and selection operator (LASSO) regression was employed to exclude repetitive or unnecessary features in the data cleaning process. Details of the deployment of LASSO can be found in Supplementary Note 1 and Supplementary Table 3. The correlations among features are provided in Supplementary Fig. 2. The Results section mainly discusses the retained features.

As most of the considered behaviours lead to the simultaneous consumption of water and energy, in this study, WU and EU were classified according to the degree of influence each feature has on water and energy consumption for different household behaviours. For laundry, as washing machines are primarily used for washing clothes, electricity only provides

kinetic energy and is not a primary desired end-use. Therefore, only the power of washing machines (*power\_WM*) was considered an EU feature. The type of washing machine (*type\_WM*) and frequency of using washing machines (*frequency\_laundry*) were listed as WU features. For bathing, the frequency and duration of bathing are related to water consumption, and the features *frequency\_bathing* and *duration\_bathing* were associated with WU. The temperature setting when using a water heater (*temperature\_WH*) has a considerable impact on electricity consumption and was included as an EU feature. For culinary, water consumption is mainly determined by frequency, and duration has a greater influence on energy consumption. Therefore, the feature “*duration\_culinary*” was categorized as an EU feature. For air conditioning (and mopping), there may be little water (and electricity) consumption. Consequently, *duration\_ac* and *frequency\_mopping* were classified as EU and WU features, respectively.

### Modelling strategy: A stepwise-like approach

To verify whether the explanatory power (i.e., the model performance) could be improved with the consideration of the water-energy nexus concept, a stepwise-like approach was adopted for modelling in this study. Four models were built using different combinations of features (detailed in Supplementary Fig. 3). Models (1) and (2) compared the explanatory powers of WU and EU in relation to household water consumption. These models utilized the same number of features to control the effect of the number of features on the results. Models (1), (2) and (3) were used to assess the improvement in explanatory power by adding EU to the model. Models (3) and (4) were employed to observe the improvement in explanatory power with the inclusion of EC in the model.

Notably, including more features in the models could naturally improve their goodness of fit. To further validate the explanatory power of each WU, EU and EC feature, the most effective technique among OLS, RF, and XGBoost was selected based on their performance. The selected technique was utilized to fit model (4) repeatedly, with one WU, EU or EC feature removed in each fitting case. By quantifying the changes in  $R^2$ ,  $RMSE$  and  $MAPE$  when each feature was removed, the explanatory powers of EC and EU for the full set of features were assessed.

### Modelling techniques

(1) *OLS multiple regression.* In this study, the OLS technique was firstly used to model and explain household water consumption. The OLS technique is the most broadly applied regression technique<sup>59</sup> but can only be used to investigate linear relationships<sup>60</sup>. Here, the most complex Model (4) is described in Eq. (1); for other models, please refer to Supplementary Note 2.

$$WaterCons_i = \alpha_1 H_i + \alpha_2 WU_i + \alpha_3 EU_i + \alpha_4 EC_i + \alpha_0 + \varepsilon \quad (1)$$

Here,  $H_i$ ,  $WU_i$ ,  $EU_i$  and  $EC_i$  are the matrices of features,  $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$  and  $\alpha_4$  are the matrices of estimated coefficients,  $\alpha_0$  is the intercept term, and  $\varepsilon$  is the residual.

(2) *Random forest.* The RF technique is a machine learning algorithm that was proposed by Breiman<sup>61</sup> in 2001. It can handle high-dimensional datasets and displays good reliability and low time complexity<sup>61</sup>. As an ensemble learner, it can efficiently prevent overfitting<sup>62</sup>. The basic principle of the RF is to introduce the bagging algorithm to CART decision trees multiple times with put-back random sampling and then to perform training to obtain a single decision tree classifier to complete the construction of an integrated model. That is, when the RF receives an ( $x$ ) input vector, made up of the values of the different evidential features analysed for a given training area, a number  $K$  of regression trees is built, and the results are averaged. RF regression with the  $K$  tree

regressor is shown in Eq. (2):

$$\hat{f}_{rf(x)}^K = \frac{1}{K} \sum_{k=1}^K T(x) \quad (2)$$

(3) **XGBoost**. Introduced by Chen and Guestrin<sup>63</sup>, XGBoost is a novel tree-based machine learning algorithm. Compared with other tree-based machine learning algorithms, XGBoost can help reduce overfitting<sup>64</sup> and increase computing speed<sup>65</sup>. Moreover, it displays excellent ability to capture the nonlinear interactions between features and labels<sup>42</sup>. XGBoost can construct new trees by continuously performing feature splits to fit the residuals of the last modelled label values and the observed values, and the results of all trees are summed as the final model results. The algorithm can be expressed as shown in Eq. (3):

$$\hat{y}_i = \sum_{k=1}^k f_k(x_i), f_k \in F \quad (3)$$

Here,  $f_k(x_i)$  is the penalty function for the  $k$ -th independent decision tree,  $x_i$  is the feature vector,  $\hat{y}_i$  is the predicted value of the label, and  $F$  is the space of the regression trees.

As hyperparameters may have a large impact on the modelling performance, they were optimized by deploying an exhaustive grid search in the models established based on the RF and XGBoost techniques to increase explanatory power. The fitting and training processes of the four models using three techniques were implemented in Python 3.10.

### Performance evaluation methods

In this study, the three most widely used model performance metrics, namely,  $R^2$ ,  $RMSE$  and  $MAPE$ , was used to evaluate the performance of the models. Equation (4) to Eq. (6) give the formulas for the 3 metrics.

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (\bar{Y} - Y_i)^2} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n}} \quad (5)$$

$$MAPE = \frac{\sum_{i=1}^n \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right|}{n} \quad (6)$$

where  $Y_i$  is the observed value of household water consumption,  $\hat{Y}_i$  is the predicted value of household water consumption,  $\bar{Y}$  is the mean observed value of water consumption, and  $n$  is the number of samples.

### DATA AVAILABILITY

The data will be made available upon reasonable request.

### CODE AVAILABILITY

The Python code will be made available upon reasonable request.

Received: 14 August 2023; Accepted: 17 January 2024;

Published online: 12 February 2024

### REFERENCES

- Mazzoni, F. et al. Investigating the characteristics of residential end uses of water: A worldwide review. *Water Res.* **230**, 119500 (2023).

- Boretti, A. & Rosa, L. Reassessing the projections of the World Water Development Report. *npj Clean. Water* **2**, 15 (2019).
- Surendra, H. J. & Deka, P. C. Municipal residential water consumption estimation techniques using traditional and soft computing approach: a review. *Water Conserv. Sci. Eng.* **7**, 77–85 (2022).
- Kim, J. et al. Development of a deep learning-based prediction model for water consumption at the household level. *Water* **14**, 1512 (2022).
- Alharsha, I., Memon, F. A., Farmani, R. & Hussien, W. E. A. An investigation of domestic water consumption in Sirte, Libya. *Urban Water J.* **19**, 922–944 (2022).
- Donkor, E. A., Mazzuchi, T. A., Soyer, R. & Roberson, J. A. Urban water demand forecasting: review of methods and models. *J. Water Resour. Plan. Manag.* **140**, 146–159 (2014).
- Grespan, A., Garcia, J., Brikalski, M. P., Henning, E. & Kalbusch, A. Assessment of water consumption in households using statistical analysis and regression trees. *Sust. Cities Soc.* **87**, 104186 (2022).
- Jayarathna, L. et al. A GIS based spatial decision support system for analysing residential water demand: A case study in Australia. *Sust. Cities Soc.* **32**, 67–77 (2017).
- Mostafavi, N., Gándara, F. & Hoque, S. Predicting water consumption from energy data: Modeling the residential energy and water nexus in the integrated urban metabolism analysis tool (IUMAT). *Energy Build.* **158**, 1683–1693 (2018).
- Hoşgör, E. & Fischbeck, P. S. Predicting residential energy and water demand using publicly available data. *Energy Conv. Manag.* **101**, 106–117 (2015).
- Jeandron, A., Cumming, O., Kapepula, L. & Cousens, S. Predicting quality and quantity of water used by urban households based on tap water service. *npj Clean. Water* **2**, 23 (2019).
- Hussien, W. E. A., Memon, F. A. & Savic, D. A. Assessing and modelling the influence of household characteristics on per capita water consumption. *Water Resour. Manag.* **30**, 2931–2955 (2016).
- Lee, D. & Derrible, S. Predicting residential water demand with machine-based statistical learning. *J. Water Resour. Plan. Manag.* **146**, 04019067 (2020).
- Ito, Y. et al. Physical and non-physical factors associated with water consumption at the household level in a region using multiple water sources. *J. Hydrol. -Reg. Stud.* **37**, 100928 (2021).
- Singha, B., Karmaker, S. C. & Eljamal, O. Quantifying the direct and indirect effect of socio-psychological and behavioral factors on residential water conservation behavior and consumption in Japan. *Resour. Conserv. Recycl.* **190**, 106816 (2023).
- Bich-Ngoc, N., Prevedello, C., Cools, M. & Teller, J. Factors influencing residential water consumption in Wallonia, Belgium. *Util. Policy* **74**, 101281 (2022).
- Abu-Bakar, H., Williams, L. & Hallett, S. H. Contextualising household water consumption patterns in England: A socio-economic and socio-demographic narrative. *Clean. Respons. Consum.* **8**, 100104 (2023).
- Gelažanskas, L. & Gamage, K. A. A. Forecasting hot water consumption in residential houses. *Energies* **8**, 12702–12717 (2015).
- Al-Zahrani, M. A. & Abo-Monassar, A. Urban residential water demand prediction based on artificial neural networks and time series models. *Water Resour. Manag.* **29**, 3651–3662 (2015).
- Bennett, C., Stewart, R. A. & Beal, C. D. ANN-based residential water end-use demand forecasting model. *Expert Syst. Appl.* **40**, 1014–1023 (2013).
- Duerr, I. et al. Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study. *Environ. Modell. Softw.* **102**, 29–38 (2018).
- Carvalho, T. M. N. & de Assis de Souza Filho, F. Variational mode decomposition hybridized with gradient boost regression for seasonal forecast of residential water demand. *Water Resour. Manag.* **35**, 3431–3445 (2021).
- Jiang, S. et al. Residential water and energy nexus for conservation and management: A case study of Tianjin. *Int. J. Hydrog. Energy* **41**, 15919–15929 (2016).
- Wang, C., Zhou, Y., You, K. & Liu, Y. Analysis of carbon emissions accounting and influencing factors of water-energy consumption behaviors in Beijing residents. *China Environ. Manag.* **13**, 56–65 (2021).
- Yu, M., Wang, C., Liu, Y., Olsson, G. & Bai, H. Water and related electrical energy use in urban households—Influence of individual attributes in Beijing, China. *Resour. Conserv. Recycl.* **130**, 190–199 (2018).
- Kenway, S. J., Lant, P. A., Priestley, A. & Daniels, P. The connection between water and energy in cities: a review. *Water Sci. Technol.* **63**, 1983–1990 (2011).
- Vahabzadeh, M., Afshar, A. & Molajou, A. Energy simulation modeling for water-energy-food nexus system: a systematic review. *Environ. Sci. Pollut. Res.* **30**, 5487–5501 (2023).
- Maftouh, A. et al. The application of water–energy nexus in the Middle East and North Africa (MENA) region: a structured review. *Appl. Water Sci.* **12**, 83 (2022).
- Song, D., Yue, D., Chen, C. & Wang, Y. Water-using prediction method for water heater, involves obtaining image of target user, providing prediction result with water usage time and water usage in first preset time period, and using water usage amount and water temperature. CN113803888-A. <https://patents.google.com/patent/CN113803888A/zh?qoq=CN113803888-A>.



30. Pérez-Fargallo, A., Bienvenido-Huertas, D., Contreras-Espinoza, S. & Marín-Restrepo, L. Domestic hot water consumption prediction models suited for dwellings in central-southern parts of Chile. *J. Build. Eng.* **49**, 104024 (2022).
31. Meireles, I., Sousa, V., Bleys, B. & Poncelet, B. Domestic hot water consumption pattern: Relation with total water consumption and air temperature. *Renew. Sust. Energ. Rev.* **157**, 112035 (2022).
32. Zheng, X. & Wei, C. Household energy consumption in China: 2016 report. (Springer), (2019).
33. Le, V. T. & Pitts, A. A survey on electrical appliance use and energy consumption in Vietnamese households: Case study of Tuy Hoa city. *Energy Build.* **197**, 229–241 (2019).
34. Zheng, X. et al. Characteristics of residential energy consumption in China: Findings from a household survey. *Energy Policy* **75**, 126–135 (2014).
35. Wee, S. Y., Aris, A. Z., Yusoff, F. M., Praveena, S. M. & Harun, R. Drinking water consumption and association between actual and perceived risks of endocrine disrupting compounds. *npj Clean. Water* **5**, 25 (2022).
36. Department of Urban Socio-Economic Survey, National Bureau of Statistics. China City Statistical Yearbook 2021. (China Statistics Press), (2022).
37. National Bureau of Statistics. China Statistical Yearbook 2021. (China Statistics Press), (2022).
38. Nsangou, J. C. et al. Explaining household electricity consumption using quantile regression, decision tree and artificial neural network. *Energy* **250**, 123856 (2022).
39. Lu, H., Ma, X. & Azimi, M. US natural gas consumption prediction using an improved kernel-based nonlinear extension of the Arps decline model. *Energy* **194**, 116905 (2020).
40. Wang, C. et al. Residential water and energy consumption prediction at hourly resolution based on a hybrid machine learning approach. *Water Res.* **246**, 120733 (2023).
41. Suárez-Varela, M. Modeling residential water demand: An approach based on household demand systems. *J. Environ. Manage* **261**, 109921 (2020).
42. Liu, J., Zhang, S. & Fan, H. A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. *Expert Syst. Appl.* **195**, 116624 (2022).
43. Reisinger, H. The impact of research designs on R2 in linear regression models: an exploratory meta-analysis. *J. Empir. Gen. Mark. Sci.* **2**, 78 (1997).
44. Cominola, A. et al. The determinants of household water consumption: A review and assessment framework for research and practice. *npj Clean. Water* **6**, 11 (2023).
45. Escrivá-Bou, A., Lund, J. R. & Pulido-Velázquez, M. Modeling residential water and related energy, carbon footprint and costs in California. *Environ. Sci. Policy* **50**, 270–281 (2015).
46. Makonin, S., Ellert, B., Bajić, I. V. & Popowich, F. Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. *Sci. Data* **3**, 160037 (2016).
47. Shen, T., Chen, Y. & Yang, Q. Energy consumption in urban household water use and influencing factors (IN CHINESE). *Resources. Science* **37**, 744–753 (2015).
48. Abu-Bakar, H., Williams, L. & Hallett, S. H. Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England. *npj Clean. Water* **4**, 13 (2021).
49. Zapata-Webb, E. et al. The impact of COVID-19 on household energy consumption in England and Wales from April 2020 to March 2022. *Energy Build.* **297**, 113428 (2023).
50. Hartonen, T. et al. Nationwide health, socio-economic and genetic predictors of COVID-19 vaccination status in Finland. *Nat. Hum. Behav.* **7**, 1069–1083 (2023).
51. Chen, J. et al. City- and county-level spatio-temporal energy consumption and efficiency datasets for China from 1997 to 2017. *Sci. Data* **9**, 101 (2022).
52. Barton, N. A., Hallett, S. H., Jude, S. R. & Tran, T. H. Predicting the risk of pipe failure using gradient boosted decision trees and weighted risk analysis. *npj Clean. Water* **5**, 22 (2022).
53. Li, Z., Wang, C. & Liu, Y. A dataset on energy efficiency grade of white goods in mainland China at regional and household levels. *Sci. Data* **10**, 445 (2023).
54. Marzano, R. et al. Determinants of the price response to residential water tariffs: Meta-analysis and beyond. *Environ. Modell. Softw.* **101**, 236–248 (2018).
55. Wichman, C. J. Perceived price in residential water demand: Evidence from a natural experiment. *J. Econ. Behav. Organ.* **107**, 308–323 (2014).
56. Romano, M. & Kapelan, Z. Adaptive water demand forecasting for near real-time management of smart water distribution systems. *Environ. Modell. Softw.* **60**, 265–276 (2014).
57. Gato, S., Jayasuriya, N. & Roberts, P. Temperature and rainfall thresholds for base use urban water demand modelling. *J. Hydrol.* **337**, 364–376 (2007).
58. Kavya, M., Mathew, A., Shekar, P. R. & P. S. Short term water demand forecast modelling using artificial intelligence for smart water management. *Sust. Cities Soc.* **95**, 104610 (2023).
59. Searle, S. R. Linear models. Vol. 65 (John Wiley & Sons), (1997).
60. Huang, S. In International Encyclopedia of Education (eds Robert J. Fazal Rizvi T., & Erican K.) 548–557 (Elsevier), (2023).
61. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
62. Byeon, H. A prediction model for mild cognitive impairment using random forests. *Int. J. Adv. Comput. Sci. Appl.* **6**, 8 (2015).
63. Chen, T. & Guestrin, C. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794 (2016).
64. Patnaik, B., Mishra, M., Bansal, R. C. & Jena, R. K. MODWT-XGBoost based smart energy solution for fault detection and classification in a smart microgrid. *Appl. Energy* **285**, 116457 (2021).
65. Luo, J., Zhang, Z., Fu, Y. & Rao, F. Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. *Results Phys.* **27**, 104462 (2021).
66. Gregory, G. D. & Leo, M. D. Repeated behavior and environmental psychology: the role of personal involvement and habit formation in explaining water consumption. *J. Appl. Soc. Psychol.* **33**, 1261–1296 (2003).

## ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 52091544, No. 71974110, and No. 72004115).

## AUTHOR CONTRIBUTIONS

Zonghan Li: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, and Writing – Original draft preparation. Chunyan Wang: Conceptualization, Methodology, Supervision, Writing – Reviewing and editing, Funding acquisition. Yi Liu: Conceptualization, Resources, Supervision, Writing – Reviewing and editing, Funding acquisition. Jiangshan Wang: Data curation.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41545-024-00298-6>.

**Correspondence** and requests for materials should be addressed to Chunyan Wang.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024