SOS-Water deliverable report

D2.1- Assessment of dynamic interactions of water use in the case studies and architecture of improved water use modules

Lead beneficiary	4 - POLIMI	Due Date	30 Sep 2023
WP no	2	New due date (if delay)	
Task no	2.1	Actual Delivery Date	29 Sep 2023
Dissemination level	PU – Public	Status	FINAL

Authors

Authors	Partner no	Partner organisation	Name of author
Main author	4	POLIMI	Xia Wei, Castelletti Andrea, Giuliani Matteo, Carlino Angelo
Contributing author(s)			

Review

Authors	Partner no	Partner organisation	Name of author
WP Leader review	2	UU	
Technical review	3	UPV	
Language review – *if applicable*			

Document history

Date	Version	Chapters affected	Description of change	Author	Document status
19/8/2023	0.1	outline	Outline and Summary	Xia Wei, Castelletti Andrea, Giuliani Matteo	DRAFT
2/9/2023	0.2	all	Deliverable Draft for internal review	Xia Wei, Castelletti Andrea, Giuliani Matteo, Carlino Angelo	DRAFT
20/9.2023	1		Language review	Xia Wei, Castelletti Andrea, Giuliani Matteo, Carlino Angelo	DRAFT
26/09/2023	1	All	Finalisation (formatting, small last changes)	Silvia Artuso	FINAL
01/07/2024	2	All	Correction of small formatting changes	Silvia Artuso	FINAL

Publishable Executive Summary

This deliverable describes the assessment of water use in the case studies and the construction and performance of machine learning models for predicting water use at the country level across the globe. The result in this report is produced by the activities undertaken in task T2.1 (Improved water use modelling). The broader goal of WP2 is to provide modelling tools and simulations needed to establish current and future water availability under different water use scenarios and management pathways.

Specifically, D2.1 explores machine learning methods to build the relationship between anthropogenic water use and socioeconomic, land use change, and climatic indicators. The report considers four types of water use as prediction targets: agricultural, industrial, municipal, and livestock water demand. Training the machine learning model involved using yearly data samples spanning 50 years (1965 to 2019) for agricultural, industrial, and municipal water demand, and 20 years (2000 to 2019) of annual data samples were used for livestock water demand. Seven input features were utilized to predict water demand: population, GDP, average annual precipitation, GDP per capita, population density, aridity, and irrigation area. Overall, the results indicate that the model generally performs well in terms of cross-validation, and the performance in time-basis validation is also promising (except for the prediction of livestock water demand). This suggests that the model could be employed to make predictions for future scenarios as it is robust in time and shows good predictive power. However, it is essential to acknowledge that the model has certain limitations when predicting water demand across diverse geographic regions.

Table of Content

Description of deliverable

1. Introduction

This document is Deliverable D2.1 (Assessment of dynamic interactions of water use in the case studies and architecture of improved water use modules), which describes the results of the activities undertaken in Task T2.1 (Improved water use modelling) of SOS-WATER project (SOS-WATER, 2023). The deliverable reports the assessment of water use in the case studies and the construction and results of machine learning models for predicting water use at the country level across the globe.

The broader goal of WP2 is to provide modelling tools and simulations needed to establish current and future water availability under different water use scenarios and management pathways. Specifically, D2.1 explores machine learning methods to build the relationship between anthropogenic water use and socioeconomic, land use change, and climatic indicators. The model and insights developed in D2.1 will be used to project future water uses under various future scenarios, which will be used as input to the water system models (WSMs) in T2.2 (Advancing state-of-the-art WSMs and linking with regional impact models (IMs)). The model could also contribute to high-quality scenario-building in T1.2 (Co-creation of future water scenarios).

The WSMs (e.g., PCR-GLOBWB (van Beek, 2011; Wada, 2014), CWATM (Burek P. S., 2020), WaterGAP2.2 (Flörke, 2013; Müller Schmied, 2014)) work within a process-based framework, computing water supply and human water use at the grid level. These models require priory information about potential water use at any given location. They rely on input data on the agricultural, domestic, industrial and livestock use for water, and then calculate which fraction of this use can be fulfilled by locally available resources. Typically, water use input can be generated by integrated assess models like IMAGE (Stehfest, 2014) or GCAM (Calvin, 2019). Domestic and industrial water use is determined based on variables like GDP and population, while irrigation water use is calculated using meteorological conditions, crop intensity, irrigated area, and irrigation efficiency assumptions. However, the projection of this information is not always available at the spatial and temporal resolutions required for reliable water use simulations. Consequently, obtaining this information from integrated assessment models at spatial and temporally relevant scales for multiple future scenarios is challenging. In addition, these integrated assessment models consider detailed and complex processes and hence are usually computationally expensive, making them less suitable for predictions under a large number of future scenarios.

It could therefore be beneficial to develop a simplified model that uses a reduced set of variables with available future projections for water use prediction while maintaining computational efficiency. This approach allows for exploring a wide range of future scenarios. In addition, by focusing on a subset of essential variables, we can gain a clearer understanding of the key factors influencing water use. This knowledge can aid in identifying actionable strategies for use management, infrastructure planning, and resource allocation. Machine learning models such as Random Forest and Decision Trees, known for their computational efficiency and interpretability, are promising for this purpose, although the literature lacks examples of machine learning models being employed for country-level water use prediction.

In the context of our research project, we have undertaken an in-depth analysis focusing on four distinct categories of water use predictions that are commonly required by WSMs, namely: agricultural, industrial, municipal, and livestock water use. To achieve this, we incorporated a range of relevant input features, including socioeconomic variables such as GDP and population, climate variables like precipitation and aridity, as well as land use variables exemplified by irrigated areas. These data were sourced from publicly available databases such as AQUSASTAT, ERA5-Land (Muñoz Sabater, 2019), CHELSA-W5E5 dataset (Dirk N. Karger, 2022). Throughout our study, we assessed the model's performance and evaluated the importance of each feature in our predictive framework.

The resulting models will have significant practical implications for water resources management. The simplified model will enable efficient predictions of water use under various projected future scenarios, allowing stakeholders to assess and plan for future scenarios without incurring excessive computational costs. The reduced complexity and streamlined approach will also facilitate the integration of the model into decision support systems, making it more accessible to water resource managers and policymakers.

The report is structured as follows: Section 2 describes the data and machine learning algorithms; Section 3 illustrates the model performance on water use prediction; Appendix provides more detailed information on model construction.

2. Methodology

2.1 Data

2.1.1 Overview

We focus on global country-level water use prediction. We work at country-level instead of grid level that is required in WSMs, because we only have available data of the water use at the country level. The country level results from the model developed in this study will be further downscaled into grid level as inputs to WSMs model using the standard routines adopted for the downscaling of historical data. We conducted a comprehensive analysis concentrating on four specific water use categories: agricultural, industrial, municipal, and livestock water use. To accomplish this, we integrated various input features, including socioeconomic indicators like GDP and population, climate variables such as precipitation and aridity, and land use data, exemplified by irrigated areas.

Table 1 summarises the water use prediction target variables as well as the input variables considered in this project. These datasets were obtained from publicly available sources and existing scientific literature, such as AQUASASTAT dataset (FAO, 2023); Static global irrigated map (Meier, 2018); CWatM input land fraction (Burek, 2020); precipitation data from ERA5-Land (Muñoz Sabater, 2019) and potential evapotranspiration from the CHELSA-W5E5 dataset (Dirk N. Karger, 2022).

Table 1 Summary of prediction target and input variables

Variable Name (Unit)	count of data samples	mean	std	min	max	Source*
Agricultural water use (10^9 m^3/year)	5369	17	65	0	688	1
Industrial water use (10^9 m^3/year)	5445	4	23	0	305	1

Municipal water use (10^9 m^3/year)	5652	2	8	0	79	1
Water use for livestock (watering and cleaning) (10^9 m^3/year)	640	0	1	0	10	1
GDP per capita (current US$/inhabitants)	9291	8361	16829	34	189507	1
Gross Domestic Product (current 10^9 US$)	10141	170	921	0	21433	1
Long-term average annual precipitation in volume (10^9 m^3/year)	9976	603	1596	0	15175	1
Population density (inhabitants/km^2)	10500	233	1187	1	19360	1
Total population (1000 inhabitants)	10891	27779	113166	1	1465634	1
Irrigated area (Static) (km^2)	10907	19175	88778	0	875858	2
Irrigated area (Dynamic) (km^2)	10787	12169	54512	0	631099	3
Aridity (-)	6846	1	1	0	7	4

* 1: AQUASTAT dataset (FAO, 2023); 2: Static global irrigated map (Meier, 2018); 3: CwatM input land fraction (Burek P. S., 2020); 4: Calculated using precipitation data from ERA5-Land (Muñoz Sabater, 2019) and potential evapotranspiration from the CHELSA-W5E5 dataset (Dirk N. Karger, 2022)

2.1.2 Water use data

Figure 1 to Figure 4 (a) show the global map of country-level average annual water use for agricultural (Figure 1), industrial (Figure 2), municipal (Figure 3), and livestock use (Figure 4). The agricultural and industrial water use datasets encompass data from 178 countries over a 55-year span from 1965 - 2019. The municipal water use dataset includes data from 181 countries for the same 55-year period. The data for water uses related to livestock (for watering and cleaning purposes) are available for 70 countries from 2000 to 2019.

The distribution of the water use datasets is illustrated in the second panel of Figure 1, Figure 2, Figure 3, and Figure 4. The water use distribution exhibits a significant skew, indicating a concentration of data values towards low water usage. As the volume of water usage increases, the number of data values systematically decreases. It is important to note that this skewed distribution has an impact on the model's performance, as the presence of a few extremely high use values affects the prediction of lower water use amounts. To address this issue, we have modified the model structure to mitigate the impact of the skewed distribution on prediction performance, which will be further discussed in Section 2.2.2 and Appendix A2.

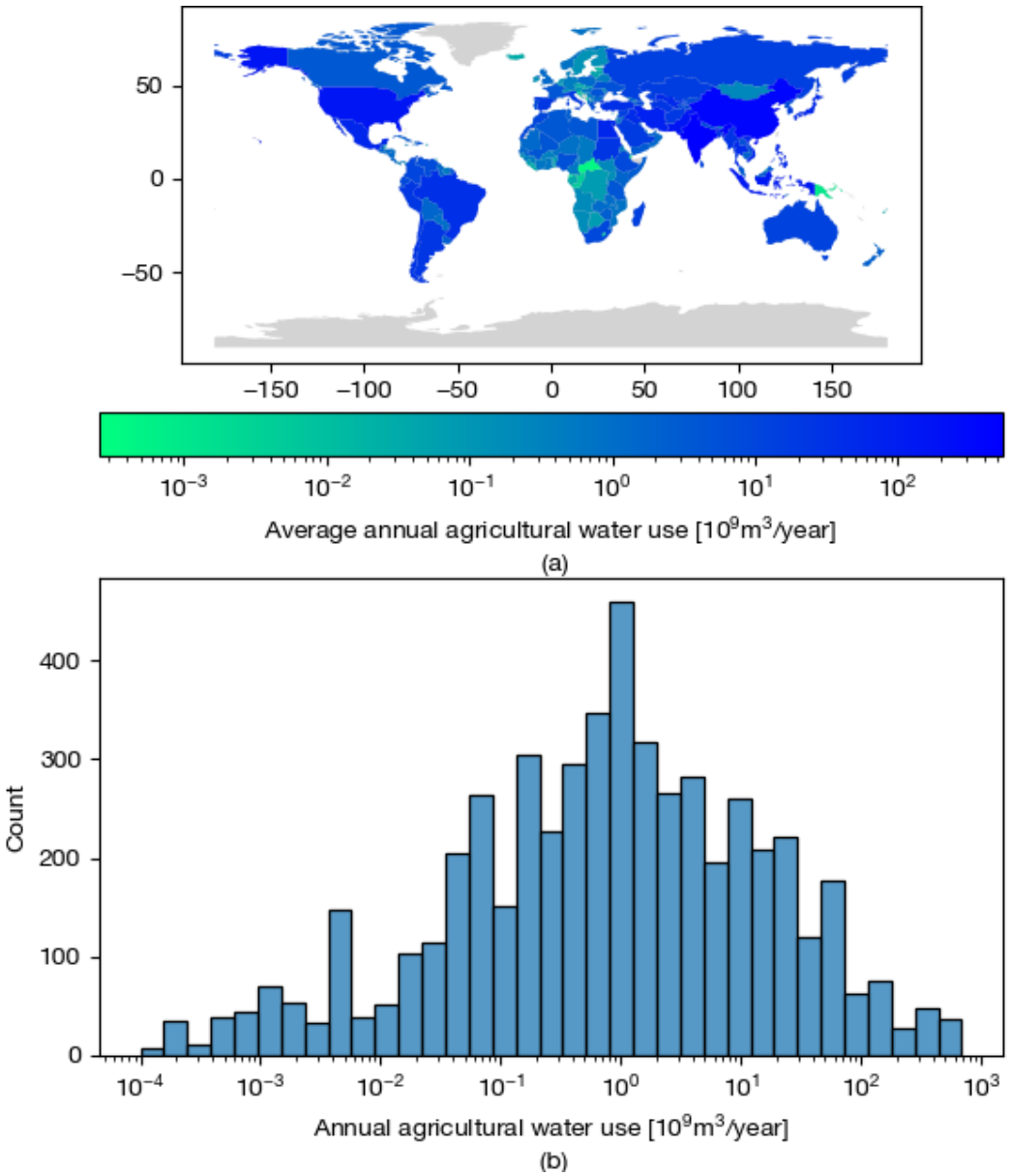

Figure 1 Country-level average annual agricultural water use and histogram of agricultural water use of samples from 178 countries covering the years 1965-2019. Countries in grey colour in panel (a) do not have water use samples. The x-axis of panel (b) is in log scale.

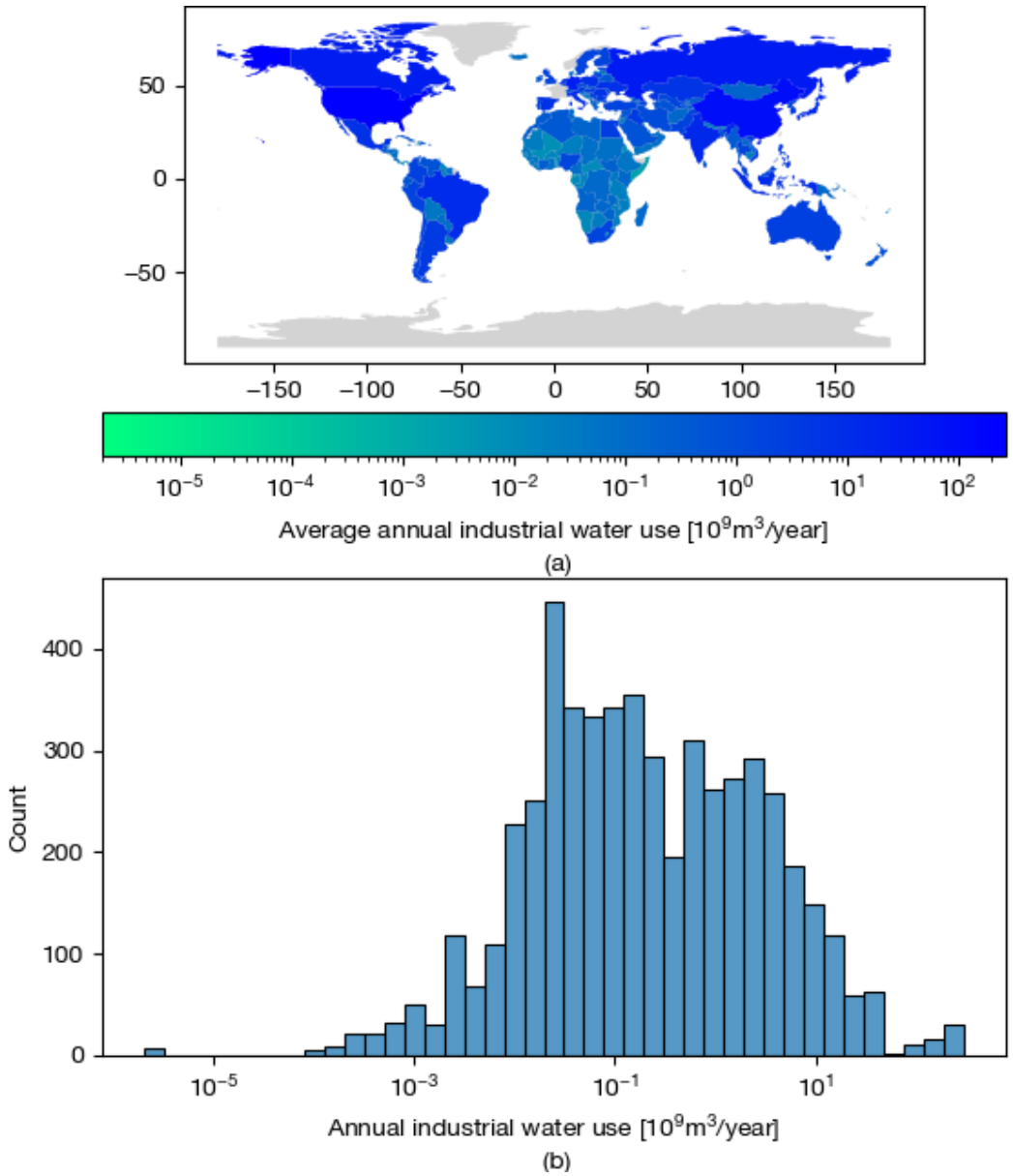

Figure 2 Country-level average annual industrial water use and histogram of industrial water use of samples from 178 countries covering 1965-2019. Countries in grey colour in panel (a) do not have water use samples. The x-axis of panel (b) is in log scale.

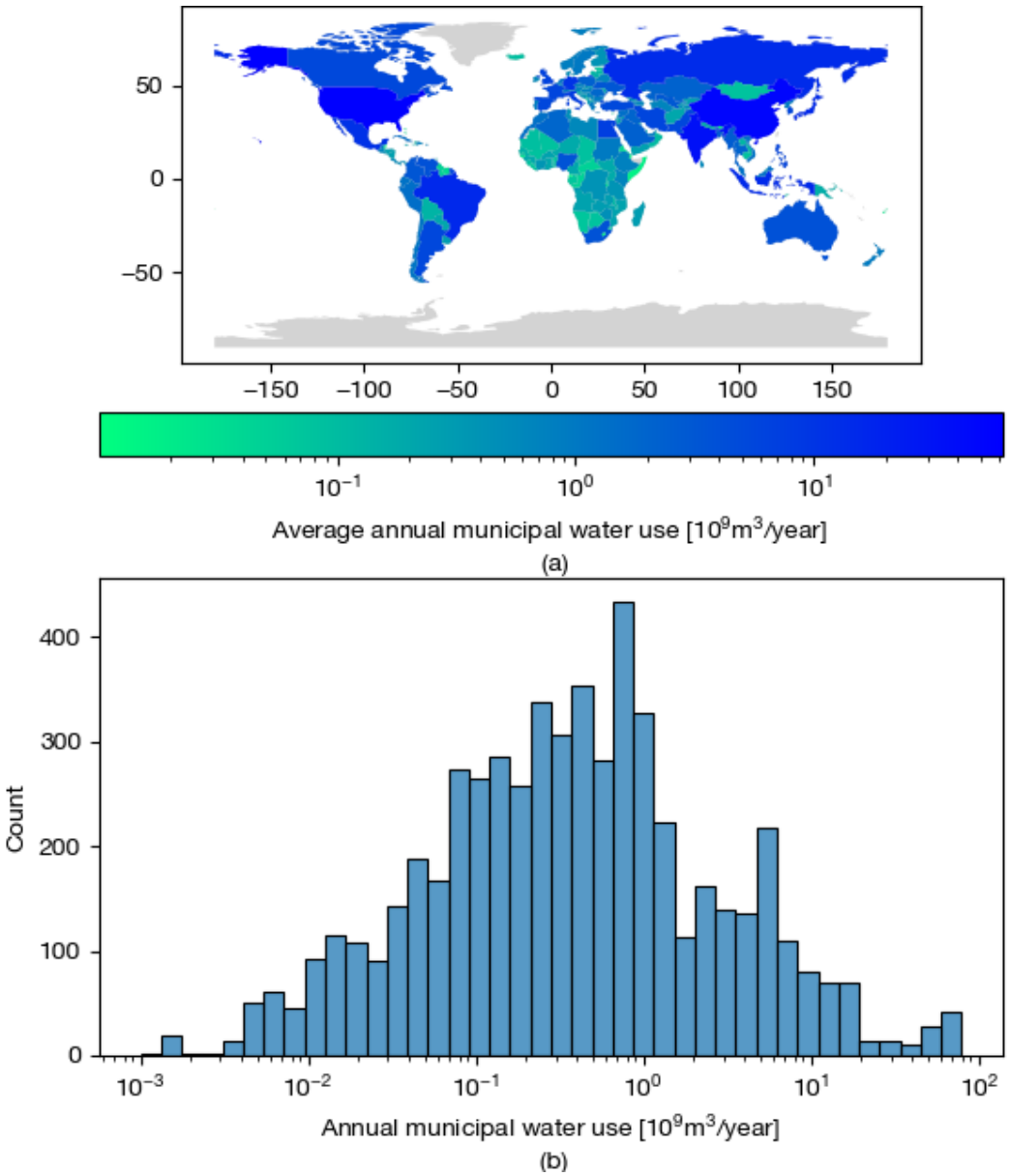

Figure 3 Country-level average annual municipal water use and histogram of municipal water use of samples from 181 countries covering the years 1965-2019. Countries in grey colour in panel (a) do not have water use samples. The x-axis of panel (b) is in log scale.

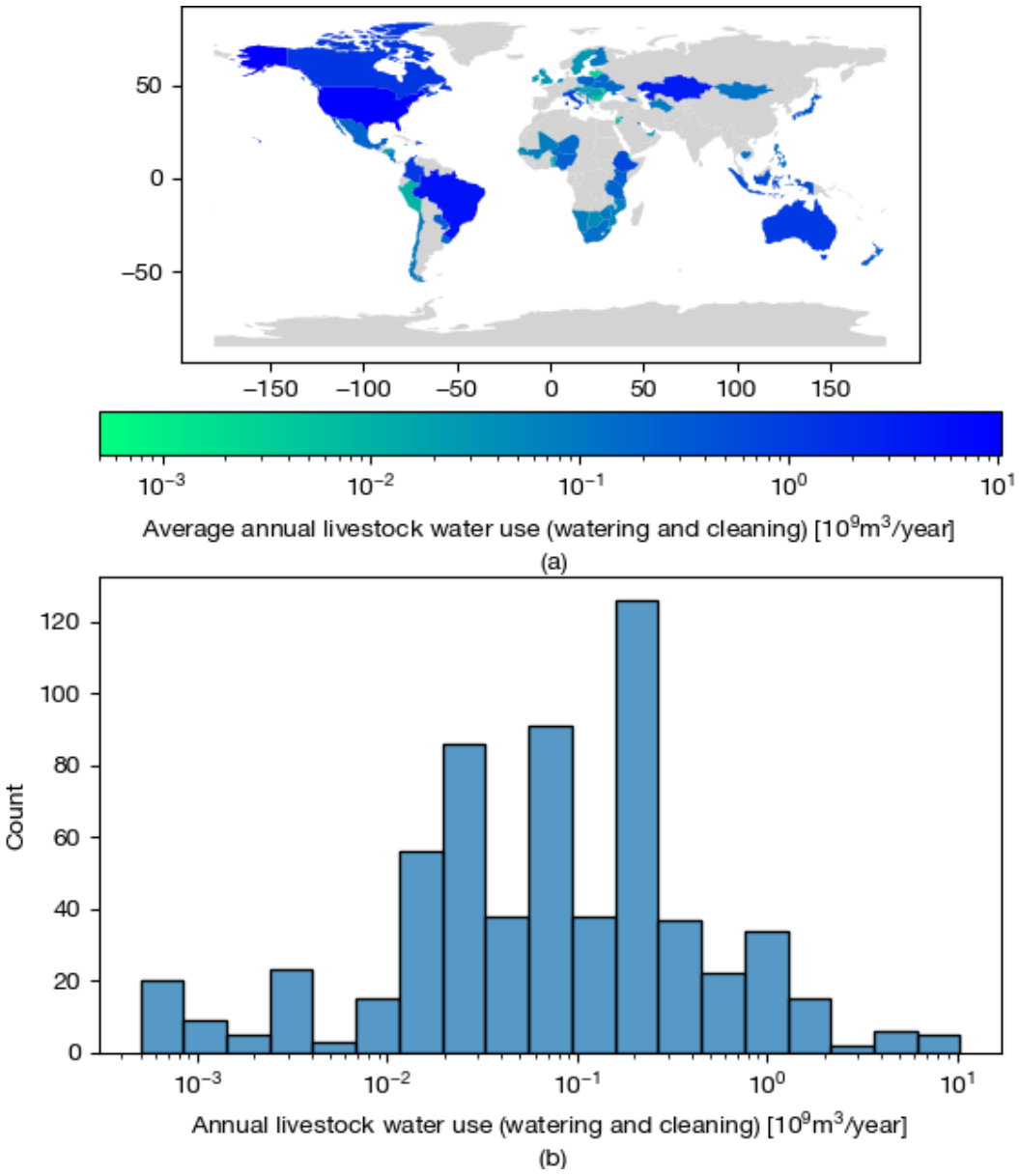

Figure 4 Country-level average annual livestock (watering and cleaning) water use and histogram of livestock (watering and cleaning) water use of samples from 70 countries covering the years 2000-2019. Countries in grey colour in panel (a) do not have water use samples. The x-axis of panel (b) is in log scale.

2.1.3 Input features data

In this report, we have examined seven types of feature data input for forecasting water use values. These features include GDP, population, average annual precipitation, GDP per capita, population density, aridity, and irrigation area. The first five input features (GDP, population, average annual precipitation, GDP per capita, and population density) are sourced from the AQUASTAT dataset (FAO, 2023). Aridity data is derived by calculating the total precipitation from the reanalysis dataset of ERA5-Land (Muñoz Sabater, 2019) and potential evaporation from the CHELSA-W5E5 dataset (Dirk N. Karger, 2022). Reanalysis datasets are created by assimilating ("inputting") climate observations using the same climate model throughout the entire reanalysis period in order to reduce the effects of modelling

changes on climate statistics. We have explored two types of irrigation area datasets: one is a static global irrigated map with a resolution of 1km (Meier, 2018) and the other is a dynamic irrigation map with a resolution of 5 arcminutes, serving as input land fraction for the CWatM model (Burek P. S., 2020).

Figure 5 illustrates the correlation between water use and the various input features. Higher correlations are in red, while lower correlations are in blue. In general, GDP, population, annual precipitation, and irrigated area (both dynamic and static) exhibit stronger correlations with water use compared to population density, GDP per capita, and aridity. Notably, population and irrigated areas demonstrate correlations exceeding 0.9 (all dark red), specifically with agricultural water use. In contrast, the correlations between water use variables and population density, GDP per capita, and aridity are typically below 0.35 (all light blue). However, it is essential to recognize that correlation solely measures linear relationships, and a lower linear correlation between input variables does not necessarily imply their lack of usefulness for water use prediction.

We initially focused on three fundamental and relevant features to evaluate the effectiveness of different input features: GDP, population, and average annual precipitation. Subsequently, we introduced four additional input features to assess their impact on water use prediction. Due to the unavailability of projected irrigation area data in the future, we conducted model tests both with and without including irrigation area as an input feature.

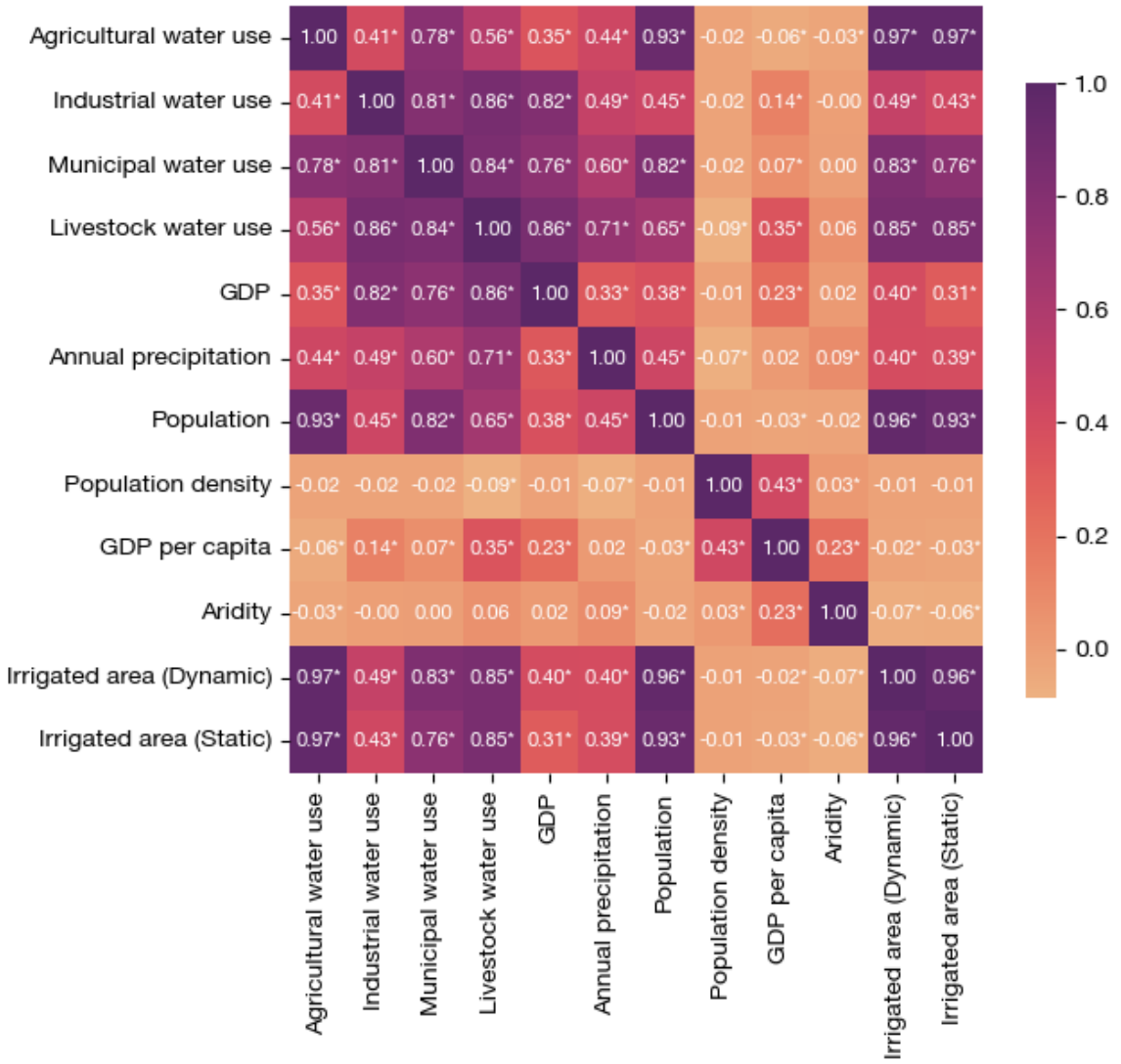

*Figure 5 Correlation heatmap between prediction target (i.e., 4 different water uses) and different input features. Correlation coefficient that are statistically significant (with P-value less than 0.05) are marked with *.*

2.1.4 Water use and input feature data in case studies

SOS-Water's focus involves exploring water challenges across five distinct case studies that encompass varying environmental and socio-economic conditions. These case studies include: (i) the Jucar River basin; (ii) the Upper Danube region; (iii) the Danube delta; (iv) the Rhine and Rhine-Meuse Delta; and (v) the Mekong River basin. In this section, our examination delves into the dynamics, specifically the annual variations, of water use and input feature data within these particular case studies. As the water use and input feature data pertain to the country level, we present comprehensive information covering all countries within each of the five designated case studies. The countries encompassed by each case study are detailed in Table 2. Please note that data may not be accessible for all countries included in the list concerning water usage and input features. Furthermore, our analysis exclusively focuses on the variations within six input features: GDP, population, aridity, dynamic irrigated area, GDP per capita, and

population density. Other input features, such as long-term annual precipitation and Static Irrigated Area, retain constant values across all years, thus they are excluded from the subsequent analysis.

Table 2 List of countries in each of the five case studies in SOS-WATER project

Case Studies	Countries in each case study
Jucar River basin	Spain
Upper Danube	Germany, Austria, Switzerland, and the Czech Republic
Danube Delta	Romania, and Ukraine
Rhine and Rhine-Meuse Delta	Switzerland, Austria, France, Germany, Belgium, Luxembourg and the Netherlands, the Principality of Liechtenstein
Mekong basin	China, Myanmar, Thailand, Lao PDR, Cambodia and Viet Nam

2.1.4.1 Jucar River basin

The Júcar River is a significant river in eastern Spain, flowing through the regions of Castilla-La Mancha and Valencia. Figure 6 illustrates the time series depicting water use across three sectors in Spain: agriculture, industry, and municipal. Livestock water use data for Spain is not presented due to its unavailability. The water usage within the agricultural, municipal, and industrial sectors generally exhibits a declining trend from the 1980s to the 2010s (as illustrated in Figure 6). This trend stands in contrast to the data concerning Spain's GDP, population, and irrigated areas, which have shown consistent growth over the past six decades, spanning from the 1960s to the 2010s (as depicted in Figure 7). This contrasting pattern could suggest an enhancement in the efficient utilization of water resources in Spain, a fact that warrants further investigation.

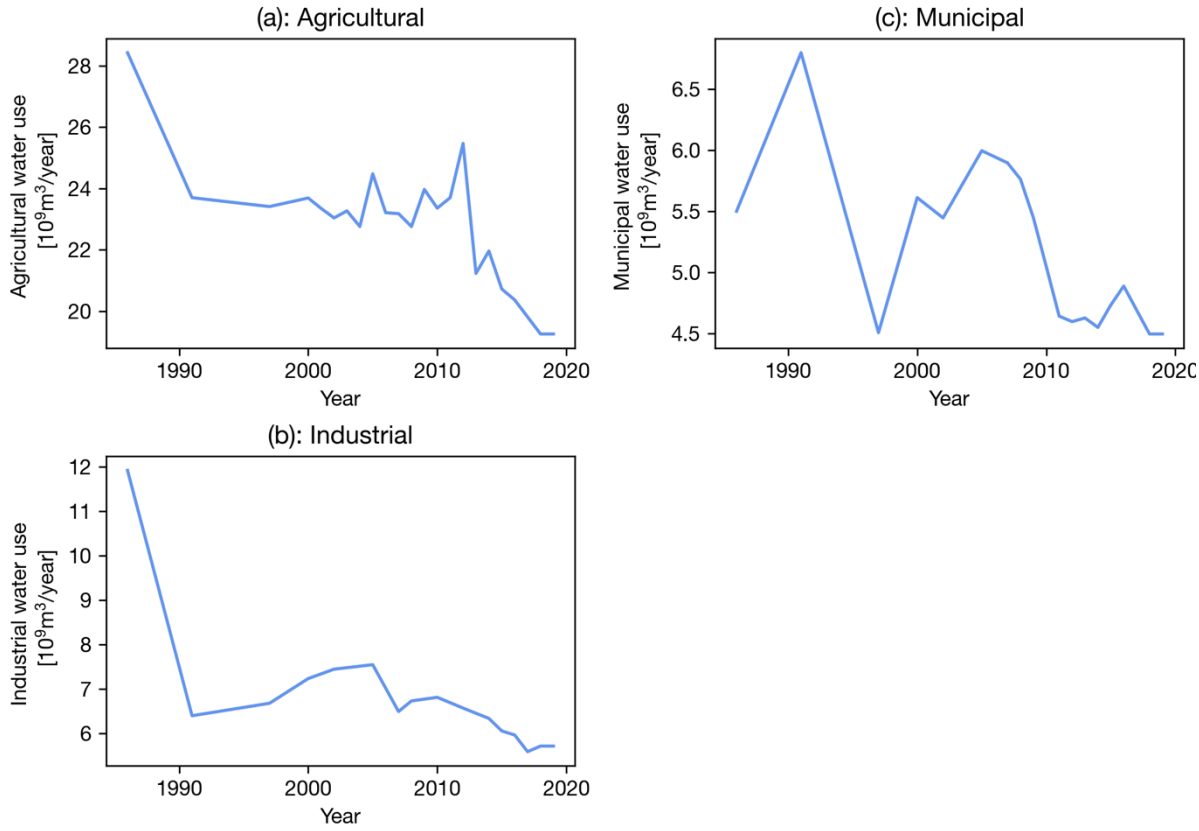

Figure 6 Time series plot of water use in different sectors: (a) agricultural, (b) industrial, and (c) municipal of country in Jucar River basin (i.e., Spain).

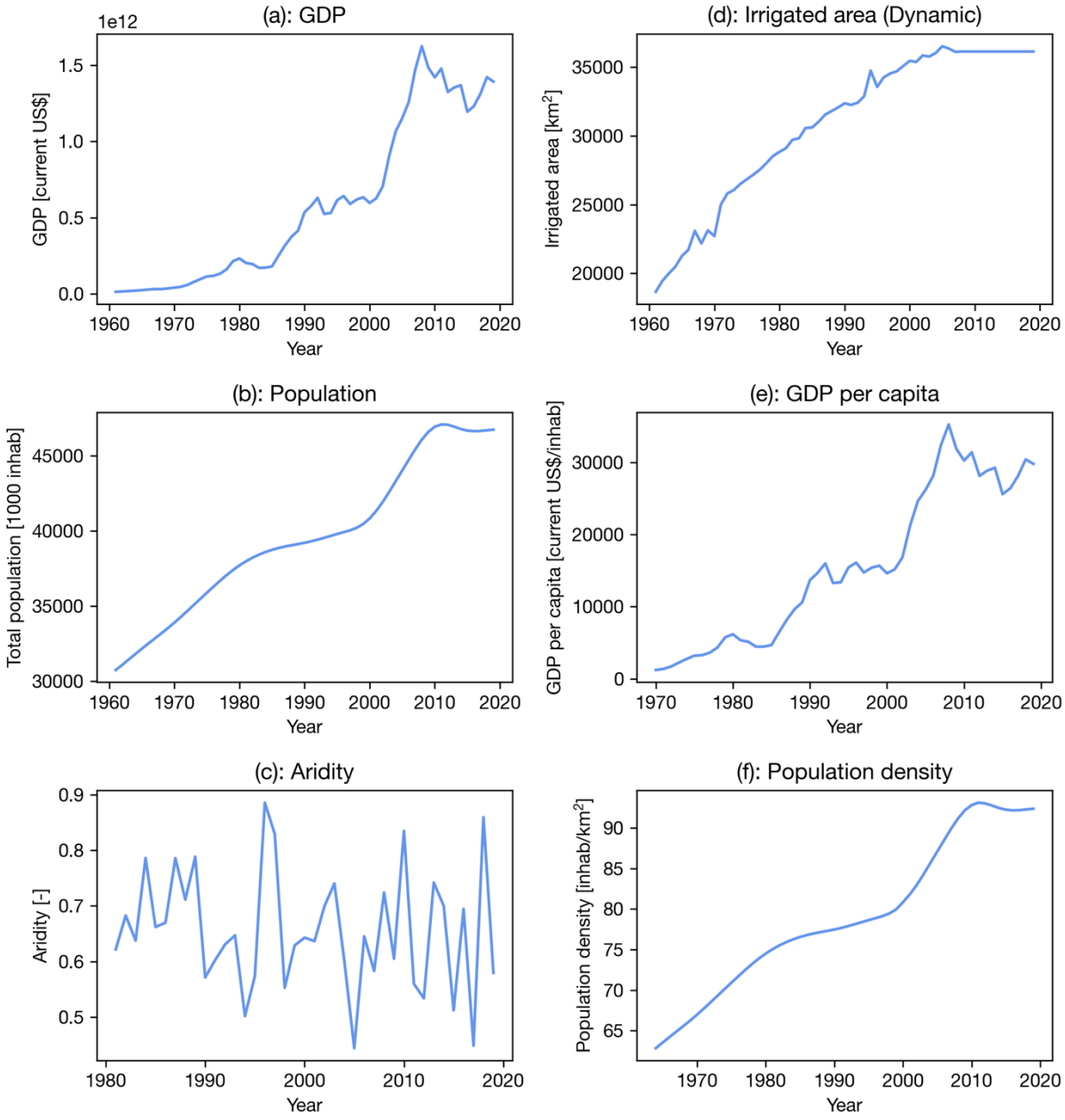

Figure 7 Time series plot of different input features: (a) GDP, (b) population, (c) aridity, (d) irrigated area (Dynamic), (e) GDP per capita, and (f) population density of country in Jucar River basin (i.e., Spain).

2.1.4.2 Upper Danube

The Upper Danube River basin spans across multiple countries in Central and Southeastern Europe, including Germany, Austria, Switzerland, the Czech Republic. This basin is characterized by its diverse landscapes, from the Alpine regions to the rolling hills, and serves as a vital source of water and habitat for both human populations and diverse ecosystems.

In Figure 8, the time series showcases the water use across four sectors (i.e., agricultural, industrial, municipal, and livestock) for 4 countries within the Upper Danube River basin. Notably, Germany emerges as the leader in water use across agricultural, industrial and municipal sector. Within Germany, water use in agricultural and industrial sectors shows a general downward trend, while water use in municipal remains relatively steady (depicted in Figure 8 (a), (b), and (c)). Additionally, Germany also has the highest GDP, population, and irrigated area (as illustrated in Figure 9).

Conversely, the other three countries (i.e., Austria, Switzerland, and the Czech Republic) demonstrate comparable levels of agricultural, industrial, and municipal water use that have remained relatively stable over the past four decades. These three countries also have comparable population size (as shown in Figure 9 (b)). Austria stands alone in possessing livestock water use data, albeit with only four years of available data from 2016 to 2019, consistently hovering around $0.056´10^9 m^3$/year (as shown in Figure 9 (d)). It's worth noting that the livestock water use data for Austria from 2017 to 2019 may have been extrapolated from the 2016 water use data.

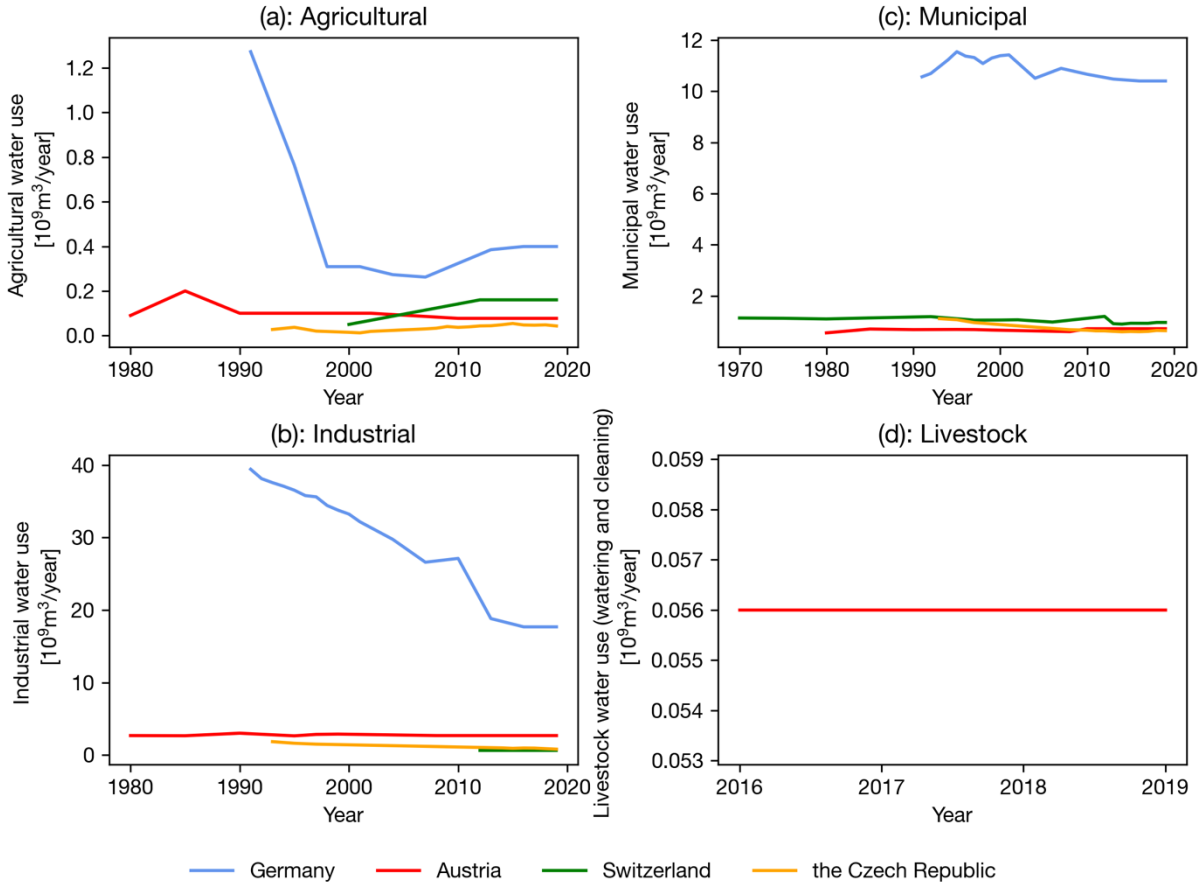

Figure 8 Time series plot of water use in different sectors: (a) agricultural, (b) industrial, (c) municipal, (d) livestock of countries in Upper Danube basin (i.e., Germany, Austria, Switzerland, the Czech Republic).

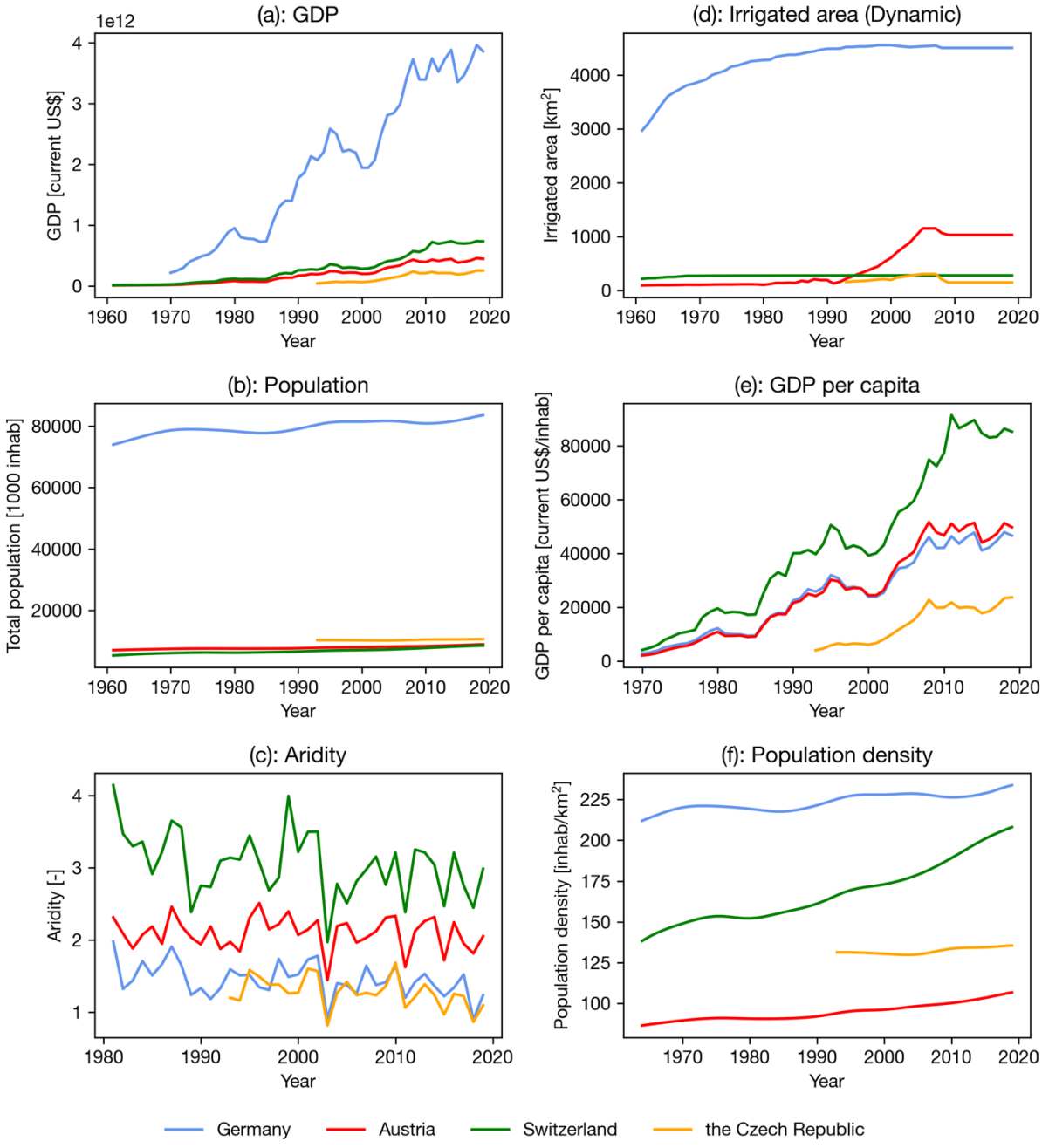

Figure 9 Time series plot of different input features: (a) GDP, (b) population, (c) aridity, (d) irrigated area (Dynamic), (e) GDP per capita, and (f) population density of countries in Upper Danube basin (i.e., Germany, Austria, Switzerland, the Czech Republic).

2.1.4.3 Danube Delta

The Danube Delta, located in Eastern Europe, is a unique and biodiverse wetland formed at the confluence of the Danube River and the Black Sea. It primarily encompasses Romania and Ukraine, with the delta's intricate network of channels, lakes, and marshes providing a crucial habitat for numerous species of plants and wildlife.

Figure 10 presents the temporal evolution of water use within the Danube Delta for Romania and Ukraine in four sectors: agricultural, industrial, municipal, and livestock. Romania has water use data accessible for the agricultural, industrial, and municipal sectors spanning from 1970 to 2019. The agricultural water usage of Romania peaked at around $9´10^9m^3$/year in 1990, subsequently witnessing a notable decline to approximately $0.94´10^9m^3$/year by 2000. Notably, the population and irrigated area also exhibited their highest values around 1990 (as depicted in Figure 11). Post-2000, Romania's agricultural water use demonstrated remarkable stability. Meanwhile, Romania's industrial water use reached a zenith at $9.81´10^9m^3$/year in 1980 before gradually tapering to roughly $3.9´10^9m^3$/year in 2019. The peak municipal water use occurred in 1986, amounting to $2.58´10^9m^3$/year, and then exhibited a consistent decline to approximately $1.06´10^9m^3$/year by 2006. The municipal water use of Romina has remained steady at around $1.0´10^9m^3$/year after 2006, excluding the water use in 2009.

Ukraine's water use data is available only from 1992 onward for the agricultural, industrial, and municipal sectors. Across the past three decades, there has been a general downward trajectory in agricultural, industrial, and municipal water use. This trend is potentially attributed to the decreasing population and irrigated areas, as indicated in Figure 11 (b) and (d). Both Romania and Ukraine only possess livestock water use data only from the year 2014 onwards, during this period, the livestock water use for both countries has exhibited notable stability.

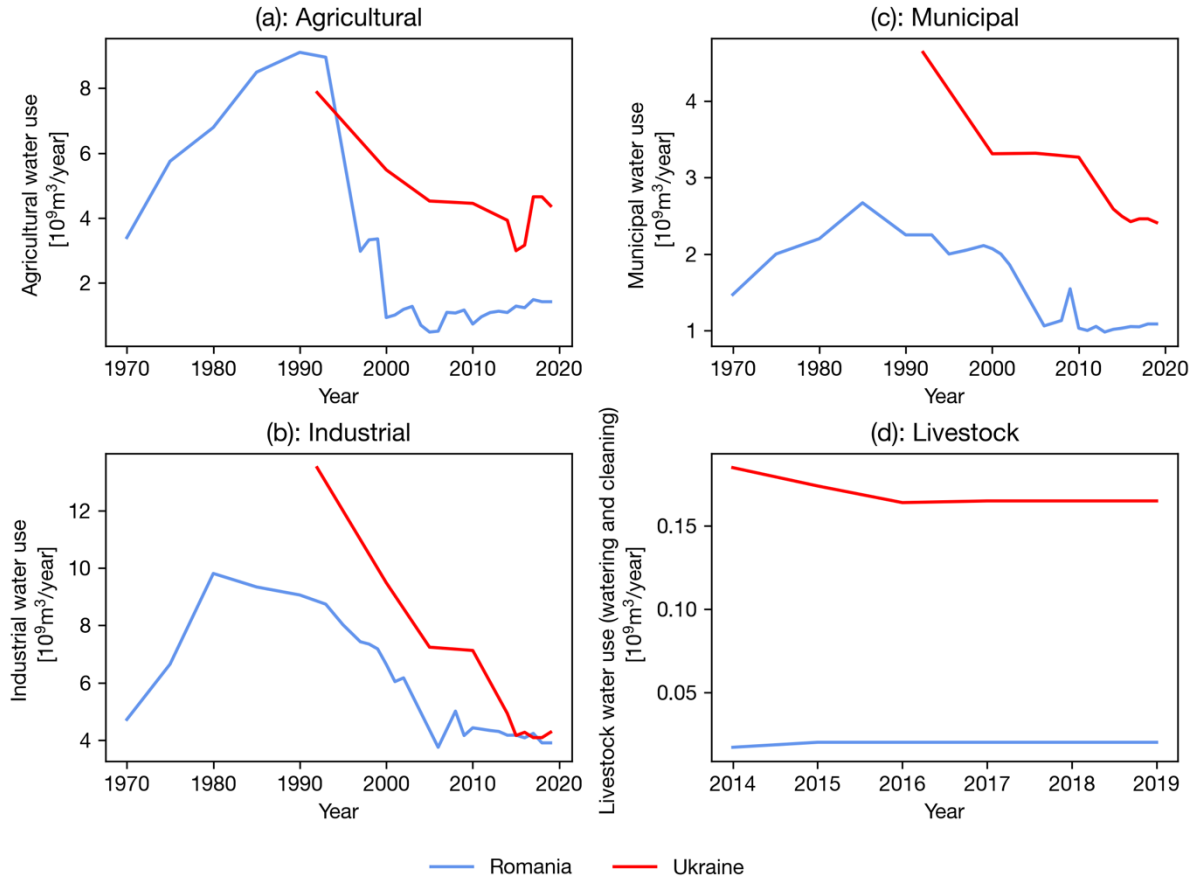

Figure 10 Time series plot of water use in different sectors: (a) agricultural, (b) industrial, (c) municipal, (d) livestock of countries in Danube Delta (i.e., Romania and Ukraine).

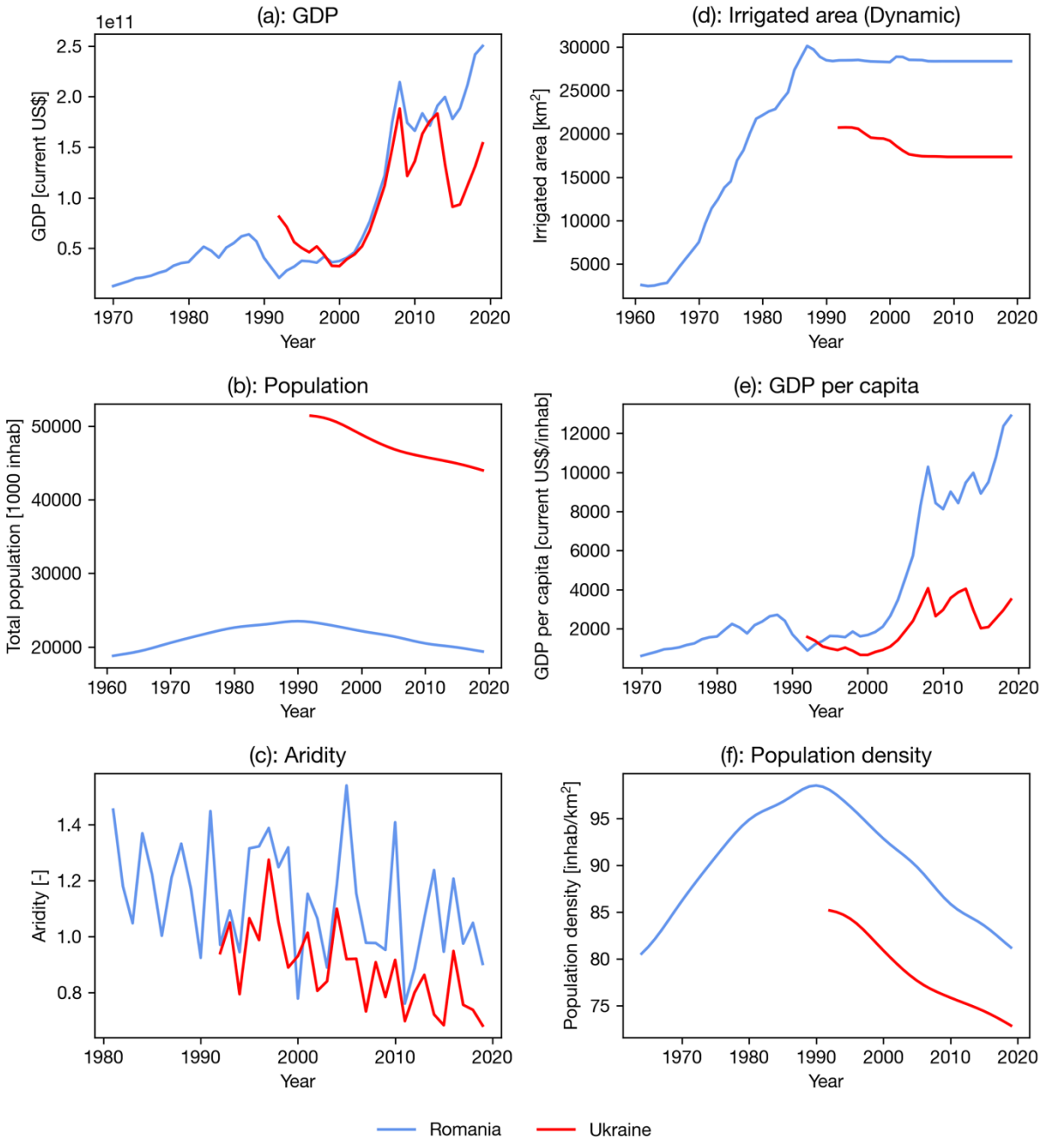

Figure 11 Time series plot of different input features: (a) GDP, (b) population, (c) aridity, (d) irrigated area (Dynamic), (e) GDP per capita, and (f) population density of countries in Danube delta (i.e., Romania and Ukraine).

2.1.4.4 Rhine and Rhine-Meuse Delta

The Rhine is one of Europe's major rivers, flowing through multiple countries including Switzerland, Germany, France, the Netherlands, and others. It eventually reaches the Rhine-Meuse Delta, a complex network of waterways, tidal rivers, and wetlands in the Netherlands and Belgium, where the Rhine and Meuse rivers merge and create a diverse and ecologically important region.

Figure 12 depicts the temporal progression of water usage for eight countries within the Rhine and Rhine-Meuse Delta, encompassing four sectors: agriculture, industry, municipality, and livestock. Over the past three decades, France has exhibited the highest agricultural water consumption. The trends in agricultural water use have experienced fluctuations but generally showcased a decline from $5.5´10^9m^3$/year in 2006 to around $3´10^9m^3$/year in 2019. On the other hand, Germany has demonstrated the greatest utilization of water in the industrial and municipal sectors compared to other nations. For a detailed exploration of water use in Germany, please refer to Section 2.4.1.2.

Switzerland and Austria stand as the only countries with available data on livestock water use. Switzerland's livestock water use has maintained stability, hovering around $0.055´10^9m^3$/year from 2015 to 2019, while Austria's figures have similarly remained constant at around $0.036´10^9m^3$/year from 2016 to 2019.

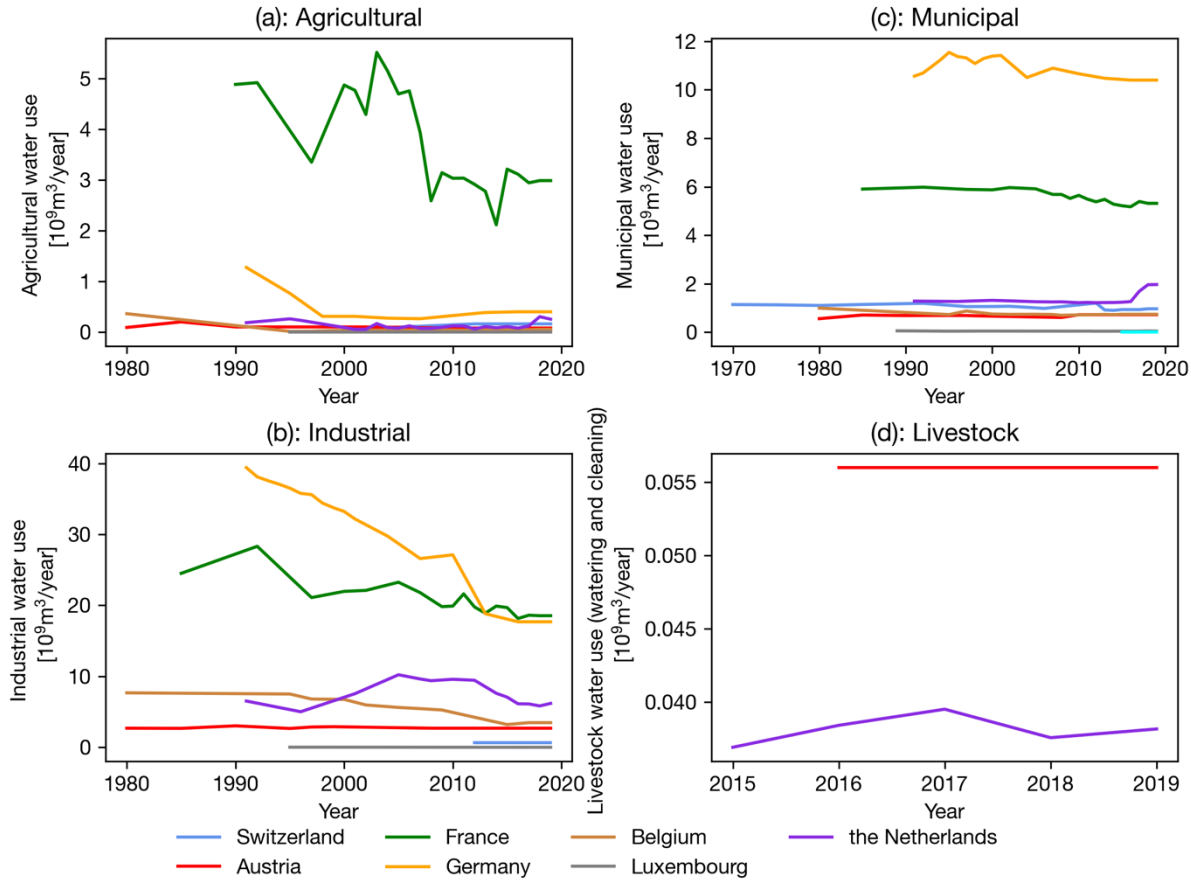

Figure 12 Time series plot of water use in different sectors: (a) agricultural, (b) industrial, (c) municipal, (d) livestock of countries in Rhine and Rhine-Meuse delta (i.e., Switzerland, France, Belgium, the Netherlands, Austria, Germany, Luxembourg).

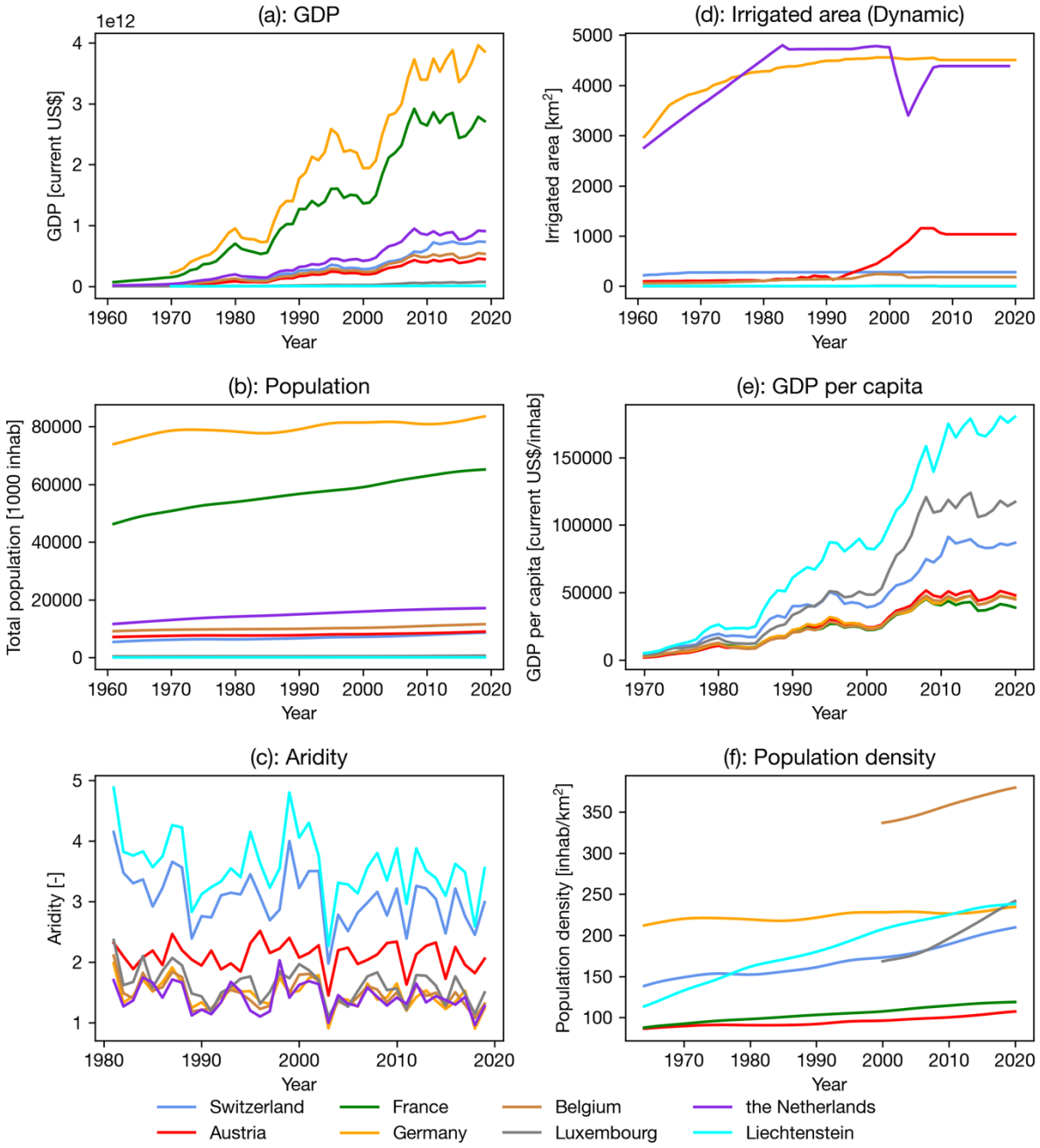

Figure 13 Time series plot of different input features: (a) GDP, (b) population, (c) aridity, (d) irrigated area (Dynamic), (e) GDP per capita, and (f) population density of countries in Rhine and Rhine-Meuse delta (i.e., Switzerland, France, Belgium, the Netherlands, Austria, Germany, Luxembourg, Liechtenstein).

2.1.4.5 Mekong basin

The Mekong River basin is a vast and crucial region in Southeast Asia, encompassing countries such as China, Myanmar, Laos, Thailand, Cambodia, and Vietnam. The basin is defined by the Mekong River's flow, supporting millions of people through its fertile plains, diverse ecosystems, and providing vital resources for agriculture, transportation, and livelihoods.

Figure 14 reports the chronological progression of water use across four sectors—agricultural, industrial, municipal, and livestock—for six countries within the Mekong River basin. China emerges as the leading consumer of water in the agricultural, industrial, and municipal domains. Over the past four decades, China's agricultural water use has demonstrated remarkable stability, while its industrial and municipal water use have displayed an upward trajectory. Specifically, China's agricultural water use has averaged around 400 ´10^9m^3/year in the past 4 decades, while that of other countries remains below 80 ´10^9m^3/year. Noteworthy is the fact that China's irrigated area expanded by nearly 1.5 times from 1980 to 2019, indicating potential enhancements in agricultural water use efficiency.

China's industrial water use increased almost threefold, rising from around 45.73´10^9m^3/year in 1980 to around 130´10^9m^3/year in 2019. Similarly, municipal water use surged from around 6.8´10^9m^3/year in 1980 to 79.4´10^9m^3/year in 2019, marking an increase of over elevenfold. This aligns with the upward trends in China's GDP and population (as shown in Figure 15). Note that only Cambodia has livestock water use data available. Over the period from 2008 to 2019, Cambodia's livestock water use experienced nearly a threefold rise, escalating from about 5.2 ´10^9m^3/year to 15.2 ´10^9m^3/year.

Figure 14 Time series plot of water use in different sectors: (a) agricultural, (b) industrial, (c) municipal, (d) livestock of countries in Mekong basin (i.e., China, Myanmar, Thailand, Laos, Cambodia, and Vietnam).

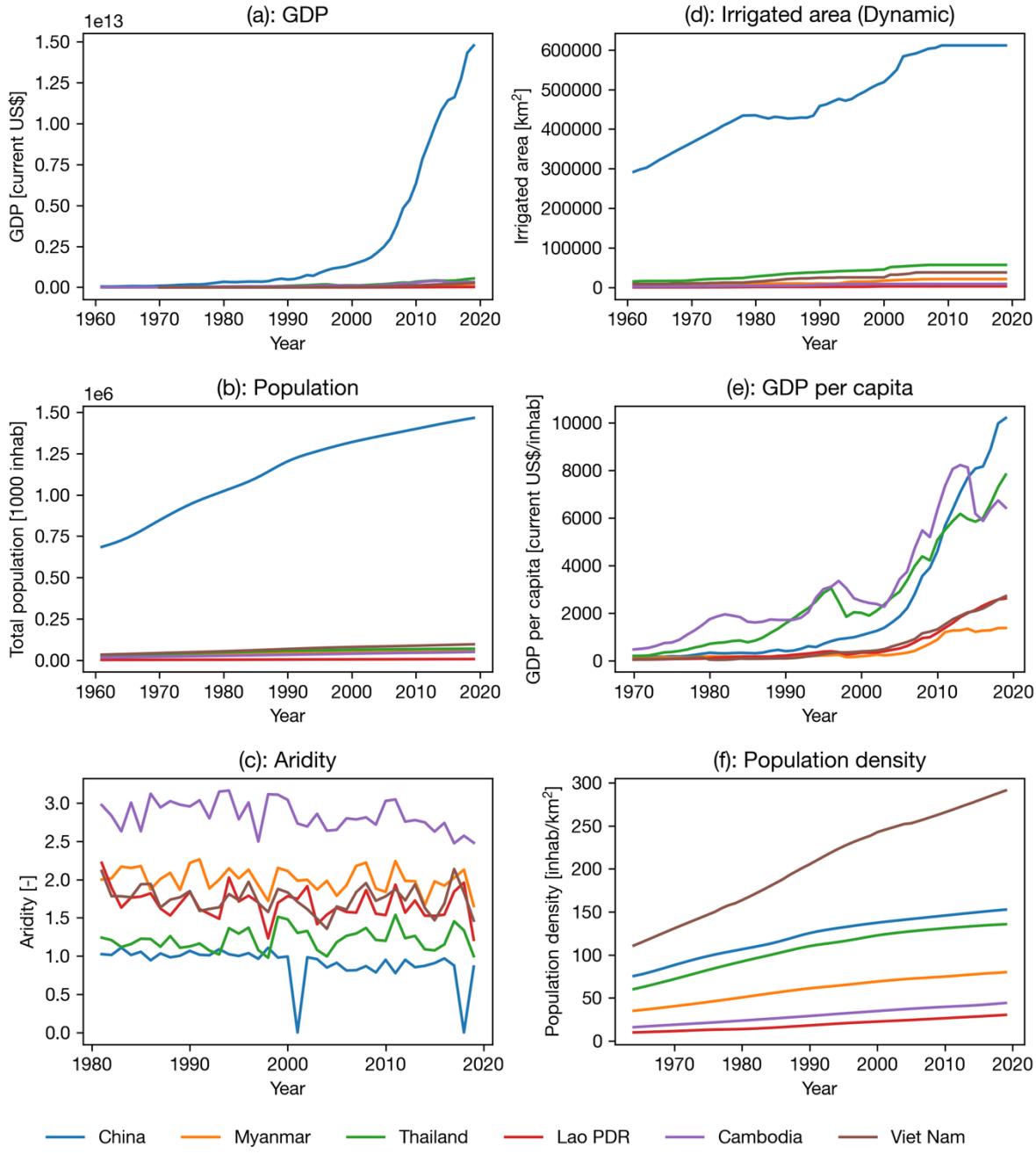

Figure 15 Time series plot of different input features: (a) GDP, (b) population, (c) aridity, (d) irrigated area (Dynamic), (e) GDP per capita, and (f) population density of countries in Mekong basin (i.e., China, Myanmar, Thailand, Laos, Cambodia, and Vietnam).

2.2 Machine Learning algorithms

2.2.1 Random Forest

The Random Forest algorithm (Breiman, 2001) is an ensemble learning method that combines multiple decision trees (Quinlan, 1986) to create a robust and accurate predictive model. Each decision tree within the Random Forest is constructed independently by randomly sampling both the training data and the features. This randomness helps to reduce the correlation between the trees and improves the

model's generalization performance, by reducing the risk of overfitting. The prediction process in Random Forest involves aggregating the predictions of individual decision trees. In the case of regression tasks, the predictions from all the trees are averaged to obtain the final predicted value.

Let us denote the training dataset as D, consisting of N samples with their corresponding input features X and target variable y. For water use prediction, the input features include variables such as population, GDP, and climate data (such as precipitation and evaporation).

The Random Forest algorithm constructs a collection of decision trees $T = \{T_1, T_2, ..., T_n\}$, where each tree T_i is built using a subset of the training data D and a random subset of features. Each decision tree T_i is trained using a modified version of the classical decision tree algorithm.

The Random Forest algorithm also provides a measure of variable importance which indicates the relative importance of each input feature in predicting water use. This importance is calculated based on the decrease in prediction performance when a particular feature is randomly permuted in the training data. The higher the decrease in performance, the more influential the feature is considered to be. Moreover, the interpretability of Random Forest models is derived from the transparency of decision trees. Each decision tree in the ensemble represents a series of if-else conditions based on the selected features, making it easy to interpret and understand the prediction process.

Based on the above advantages of Random Forest, including the ability to capture complex relationships, robustness against overfitting, variable importance assessment, and interpretability, we select it as the machine learning algorithm for water use prediction. We use the Random Forest from the *sklearn* Python package (Pedregosa, 2011).

In Appendix A1, we also compare the random forest and the classical decision tree to determine which methodology is better suited for our analysis. These results indicate that the random forest is consistently better than the decision tree for predicting all four water use targets. We instead decided to exclude the development of deep learning models (Schmidhuber, 2015) because, despite their growing use in different fields, they require large datasets for the model identification, exceeding the dimension of the water use and associated feature data introduced in the previous section.

2.2.2 Three-phase target transform model

As mentioned in Section 2.1 the distribution of the predicted target is highly skewed towards low water use countries, therefore if we train the model directly from these original datasets, it could lead to biased solutions. Here, we used two techniques to alleviate the impact of the skewed distribution on the model performance: dataset division and target transform.

In the dataset division, we divided the data set into different groups based on high or low population. This is important because there is a remarkable difference between water use in large countries (such as India and China) and small countries (such as Singapore). The difference in water use can be hundreds of times larger. Therefore, it makes sense to divide the samples into different groups so that each model could better represent the large and low water use samples. For simplicity, we used the total population per country to divide the samples into three groups: low, middle, and high. The number of samples and the specific criteria are listed in Table 3 below. Since the number of samples for the fourth target

variable (i.e., water use for livestock) is relatively limited (n = 635), we do not divide the model into three groups.

Table 3 The number of samples in each divided group. The samples are divided based on the population size of the sample into low, middle, and high three different groups

No. Samples	Low (Population < 20M)	Middle (20M<=Population<600M)	High (Population >= 600M)
Agricultural	3802	1428	85
Industrial	3854	1462	85
Municipal	4016	1502	85

To remedy the skewed data, we also transformed the target variables to make the skewed distribution more symmetric. To do this, we took the natural logarithm of one plus the target variables before training the model. When it comes to the prediction phase, we have reversed the transformation by taking the natural exponentiation of the predicted output minus 1.

The combination of the dataset division and target transform yields the three-phase target transform model (Figure 16). During the training phase, we partition the data into segments based on the population feature. Subsequently, we apply a logarithmic transformation to the target variable y, which becomes log(1+y). This modified target variable, denoted as y' = log(1+y), in combination with the input feature vector X, becomes the foundation for training the random forest model. For each distinct group created during the division process, a distinct model is trained. When transitioning to the prediction phase, the samples are initially segregated into clusters following the population feature criteria, as observed in the training phase. Subsequently, the model established during training is employed to forecast the transformed target variable y', given an input feature vector X. To obtain the projected target variable y, the predicted transformed value y' undergoes reversion through an inverse transformation process—specifically, y = exp(y') - 1.

We also tested the effectiveness of other solutions, such as 1) partitioning the dataset into two groups instead of three groups, 2) dataset division without target transform, and 3) target transform only. The model comparisons leading to the final model selection are included in Appendix A2. From our model comparison, the three-phase target transform model structure gives the best result in general by considering the model performance for both high-water use samples and low-water use samples. Hence, we adopt the three-phase target transform model structure in our project.

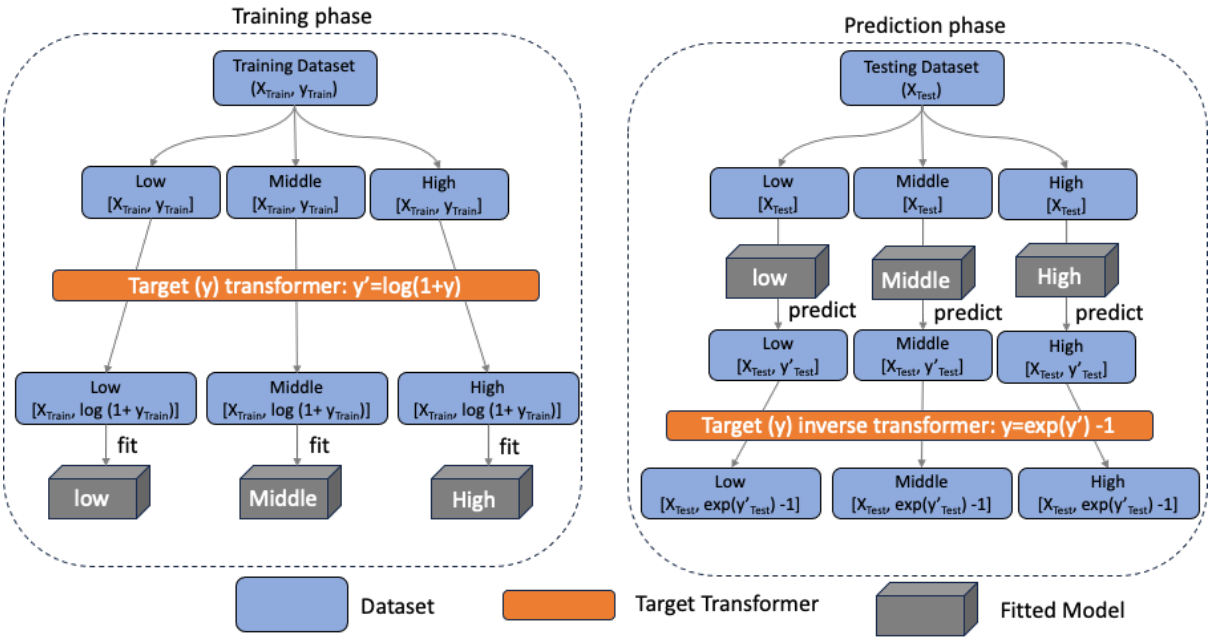

Figure 16 The workflow diagram of the three-phase target transform model. The dataset is divided into three groups: low, middle, and high, based on the total population input feature (as listed in Table 2). X *denotes the model input feature vector, and y denote the target variable.*

2.3 Model performance validation

We evaluate the model's performance through cross-validation, time-based validation, and country-based validation. Cross-validation provides an overall assessment of the model's predictive performance, taking into account the entire dataset and mitigating the impact of data partitioning. It helps estimate the model's accuracy and detect potential overfitting or underfitting. Specifically, cross-validation involves partitioning the available data into multiple subsets or "folds." The model is trained on a portion of the data and then tested on the remaining fold. This process is repeated multiple times, with each fold serving as the test set at least once. The results are then averaged to obtain an overall performance metric. Cross-validation helps assess how well the model generalizes to unseen data and provides an estimate of its predictive accuracy. In our cases, we use 5-fold cross-validation (i.e., the dataset is divided into 5-folds).

Time-based validation evaluates the model's ability to handle temporal dynamics and make accurate predictions on future data. The model is trained using data from earlier time periods and tested on more recent data. This validation technique is particularly useful when the data exhibits temporal patterns or trends and it helps in evaluating the model's ability to make predictions in real-world scenarios where future data is unavailable during the model training. Country-based validation assesses the model's performance across different countries or regions, accounting for variations in water use factors. It helps identify if the model is sensitive to geographic differences and if it can provide reliable predictions in diverse settings.

3. Results

In this section, we present an analysis of the model's performance in predicting country-level water use. To maintain simplicity, we only report the results of the final model selected. For the prediction of agricultural, industrial, and municipal water uses, our final model is the tree-phase target transform model with seven input features: GDP, population, average annual precipitation, GDP per capita, population density, aridity, and static irrigation area. We investigate the model performance through cross-validation, time-based validation and country-based validation as described in Section 2.3.

3.1 Cross-validation

Figure 17 compares water use predictions generated by the Random Forest model. Note the 3-phase target transform model is used for agricultural, industrial, and municipal water use prediction and the classical random forecast model is used for livestock water use prediction. The predictions from each of the 5-fold are compared against the actual observed water use values. The results shown in Figure 17 demonstrate that the final model performs very well in cross-validation, achieving a coefficient of determination above 0.9 for all four prediction problems.

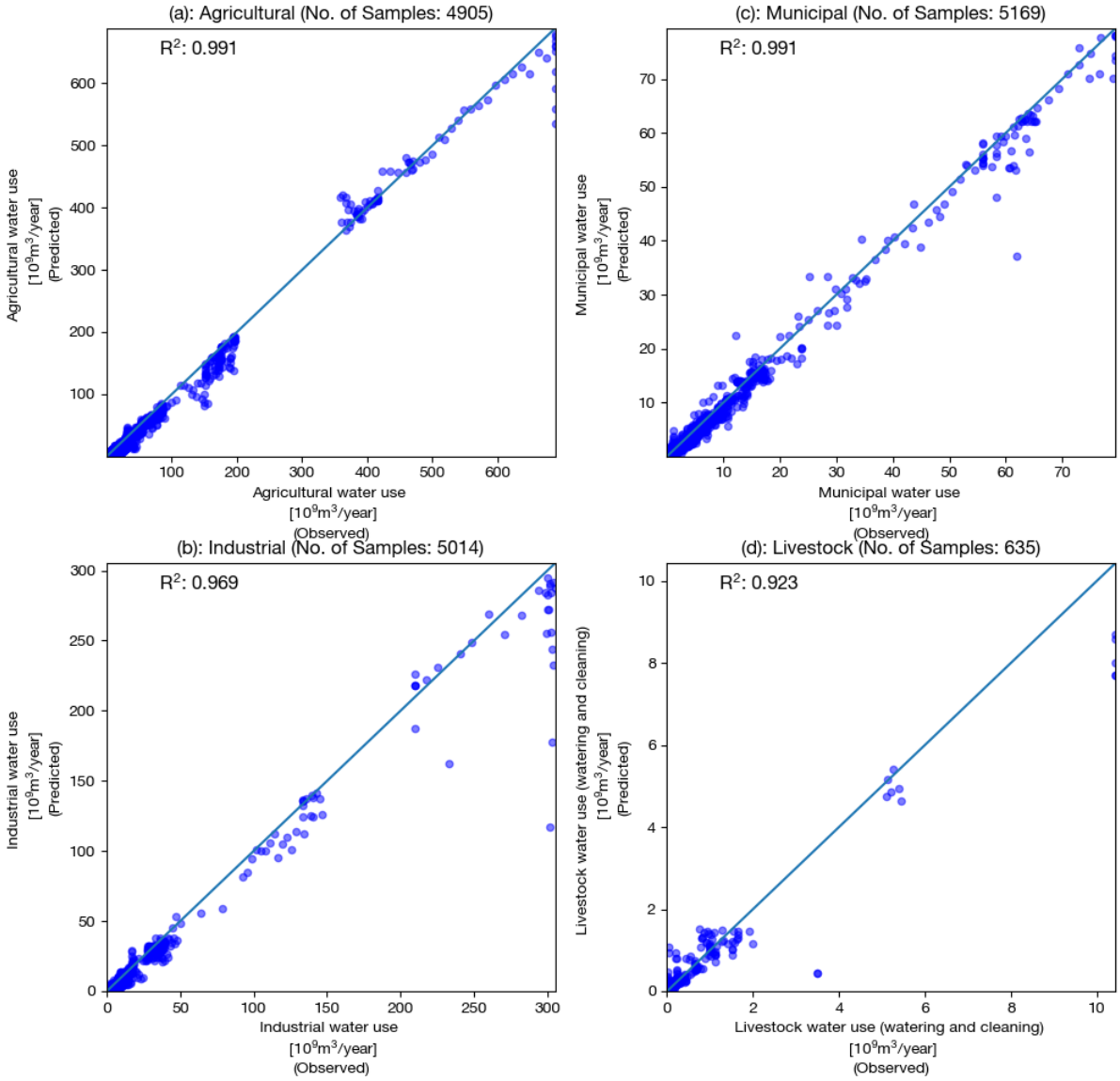

Figure 17 Predictions versus observations from 5-fold cross-validation. The line in each subplot is a 1:1 line representing perfect agreement or equality between predicted and observed values. The value of the coefficient of determination between predictions and observations is reported in each subplot

3.2 Feature importance

Feature importance analysis is essential for understanding which features significantly influence the model's predictions. This allows us to identify key drivers in the prediction of water use. It aids in building transparent and accurate models, leading to better decision-making and improved model performance.

To get the feature importance of each input variable, we use all the datasets to train the model and extract the feature importance on the trained model (Figure 18, Figure 19, Figure 20 and Figure 21). Since we used the 3-phase target transform model for agricultural, industrial, and municipal water use prediction, we present the feature importance of 3 separate models trained in low, middle and high water use samples (as shown in Figure 18, Figure 19, Figure 20). For the prediction of livestock water

use, one Random Forest model is trained with all samples. Hence, we only have features importance from one model (as shown in Figure 21).

For agricultural water use prediction, we found that precipitation, population and irrigated area are the three most important features when predicting the low and middle water use samples (as shown in the Figure 18 (a) and (b)). When predicting the high agricultural water use samples, the feature of population density becomes the most important before the precipitation, population and irrigated area (as shown in Figure 18 (c)).

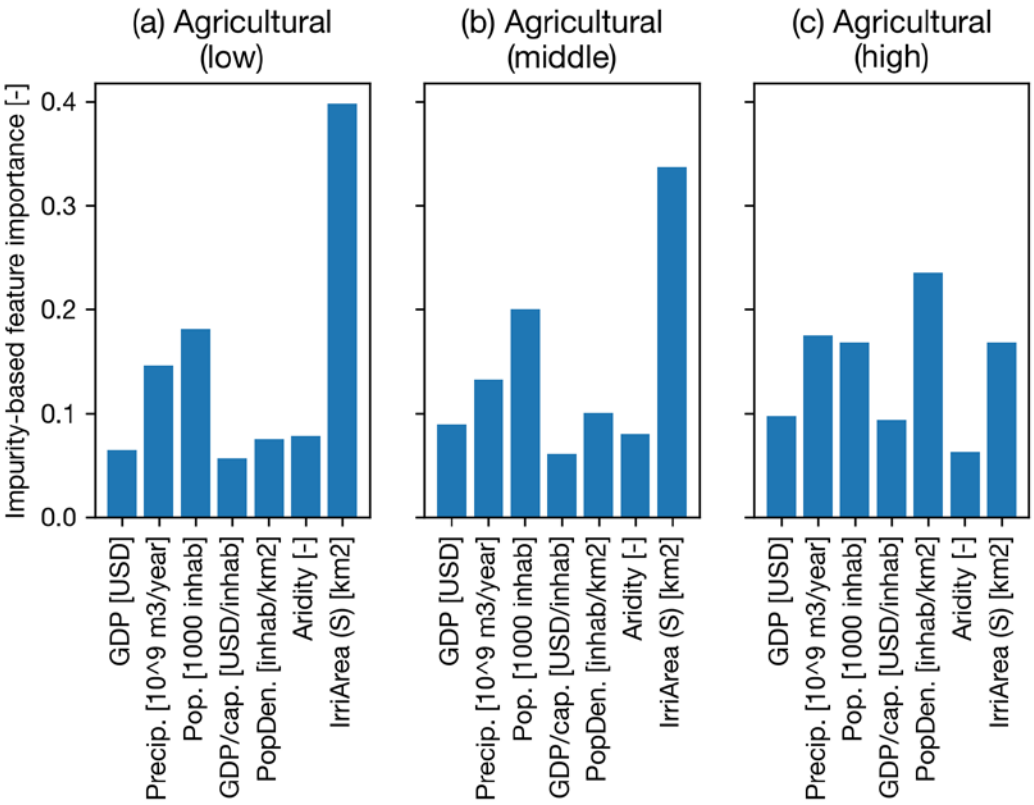

Figure 18 Feature importance of the trained random forest model on each divided sample (low (a), middle (b), and high(c), as shown in Table 3) for the agricultural water use prediction

When predicting industrial water use, we found that GDP is much more important than it was in the model for agricultural water use. As shown Figure 19 (a) and (b), the GDP is among the four most important features, together with precipitation, population and irrigated area. The feature importance of the trained model with high water use samples is quite different from the trained model with lower and middle water use samples (similar to the features importance of the model for prediction of agricultural water use). This said, precipitation, population and population density are still the most important features.

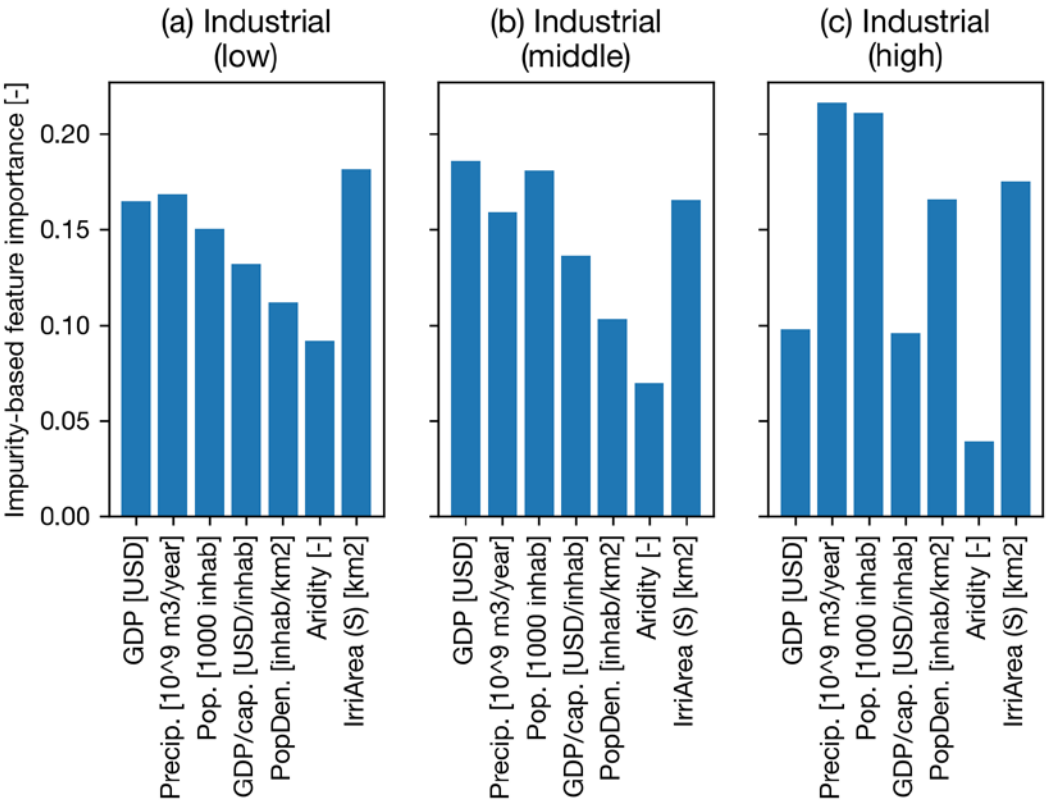

Figure 19 Feature importance of the trained random forest model on each divided sample (low (a), middle (b), and high (c), as shown in Table 3) for the industrial water use prediction

When predicting municipal water use, GDP, population, and irrigated area are the three most important features for the low water use subset (Figure 20 (a)). The importance of precipitation and GDP per capita increased when predicting the middle water use samples. Interestingly, precipitation becomes the least important feature when predicting the high-water use samples, where population and GDP per capita become the most important.

The prediction of livestock uses only one model instead of three separate models. Precipitation, population, and irrigated area are the three most important features (Figure 21). Population density and aridity are the two least important features.

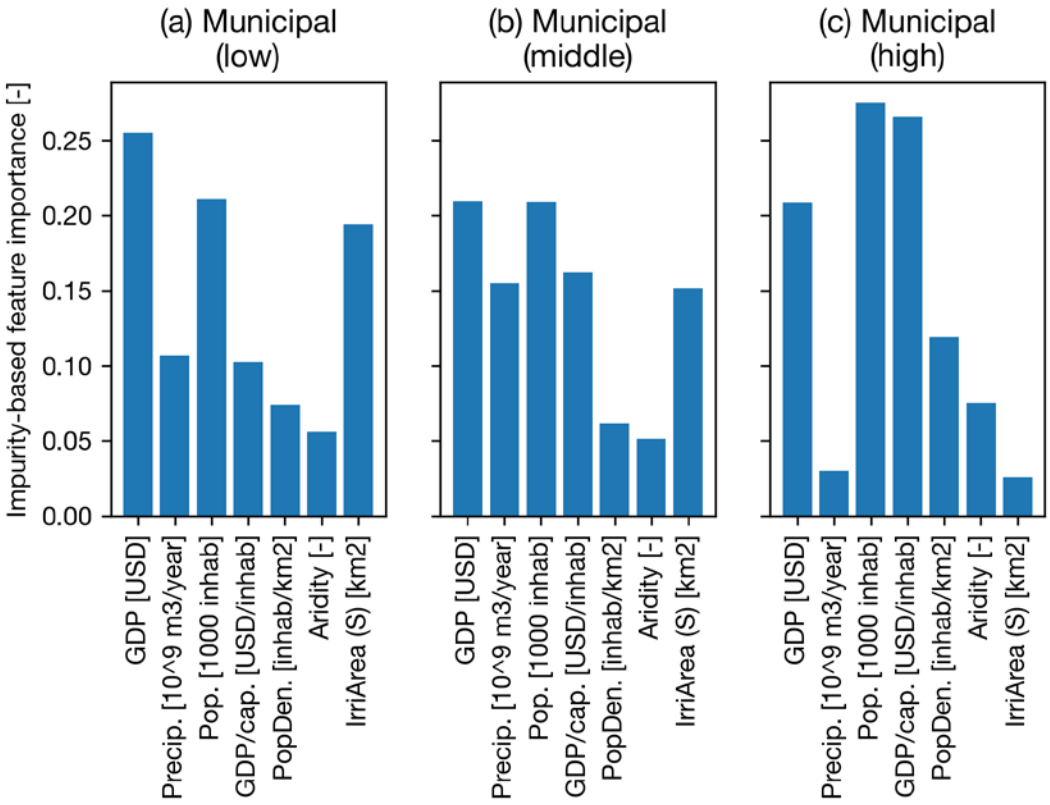

Figure 20 Feature importance of the trained random forest model on each divided sample (low (a), middle(b), and high(c), as shown in Table 3) for the municipal water use prediction

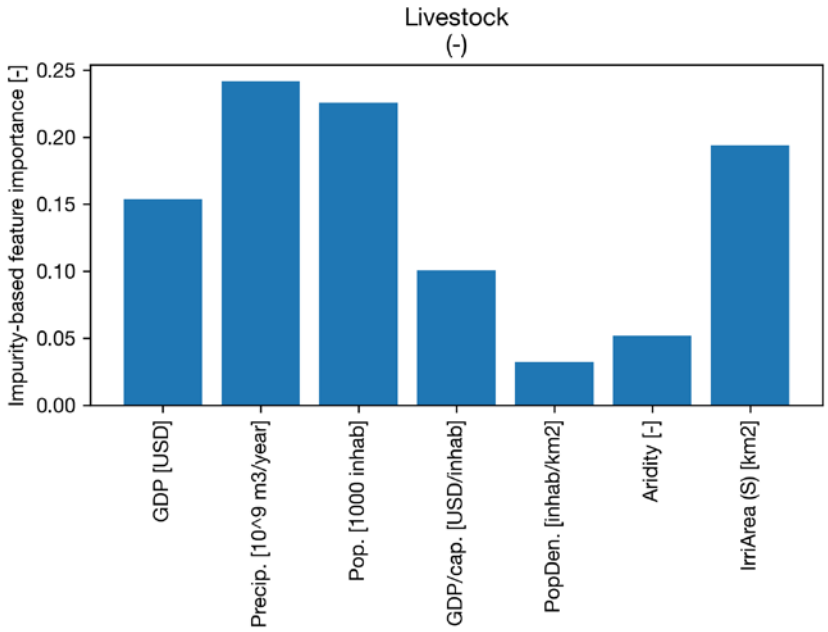

Figure 21 Feature importance of the trained random forest model for the livestock water use prediction

3.3 Time-based validation

Time-based validation involves dividing the data based on time, typically into training and testing sets. In our case, we divide the data samples into training and validation sets based on the criterion in Table 4. Note that livestock water use has data from the year 2000 to the year 2019 only. Hence, we divide the dataset into a more recent year (2015) threshold.

Table 4 The partition criterion of model train and validation dataset for time-basis validation

Prediction Target	Train	Validation
Agricultural water use	Year < 2010	Year >=2010
Municipal water use	Year < 2010	Year >=2010
Industrial water use	Year < 2010	Year >=2010
Livestock water use	Year < 2015	Year >= 2015

In Figure 22, the model's predictions are plotted against the observations using the validation dataset, which comprises samples from more recent years. The model demonstrates strong performance in time-based validation for predicting agricultural, industrial, and municipal water use, consistently achieving a coefficient of determination above 0.9 on the validation dataset for agriculture, municipal, and industrial water use. The model also tends to underestimate the high-use regions, resulting in a lower coefficient of determination of 0.428 when evaluated on this subset of data only (e.g., time-based validation dataset). This discrepancy is attributed to the skewed distribution of the target variable (Figure 1 (b), Figure 2 (b), Figure 3 (b), and Figure 4 (b)). This makes it difficult for the model to learn the patterns of extremely high-water use predictions.

However, when it comes to predicting livestock water use, the model faces challenges due to limited data availability. The number of samples for livestock water use is significantly smaller (more than five times less) compared to the other three types of water use variables. This scarcity of data hinders the model's ability to predict livestock water use accurately and consequently reduces its overall performance to 0.43.

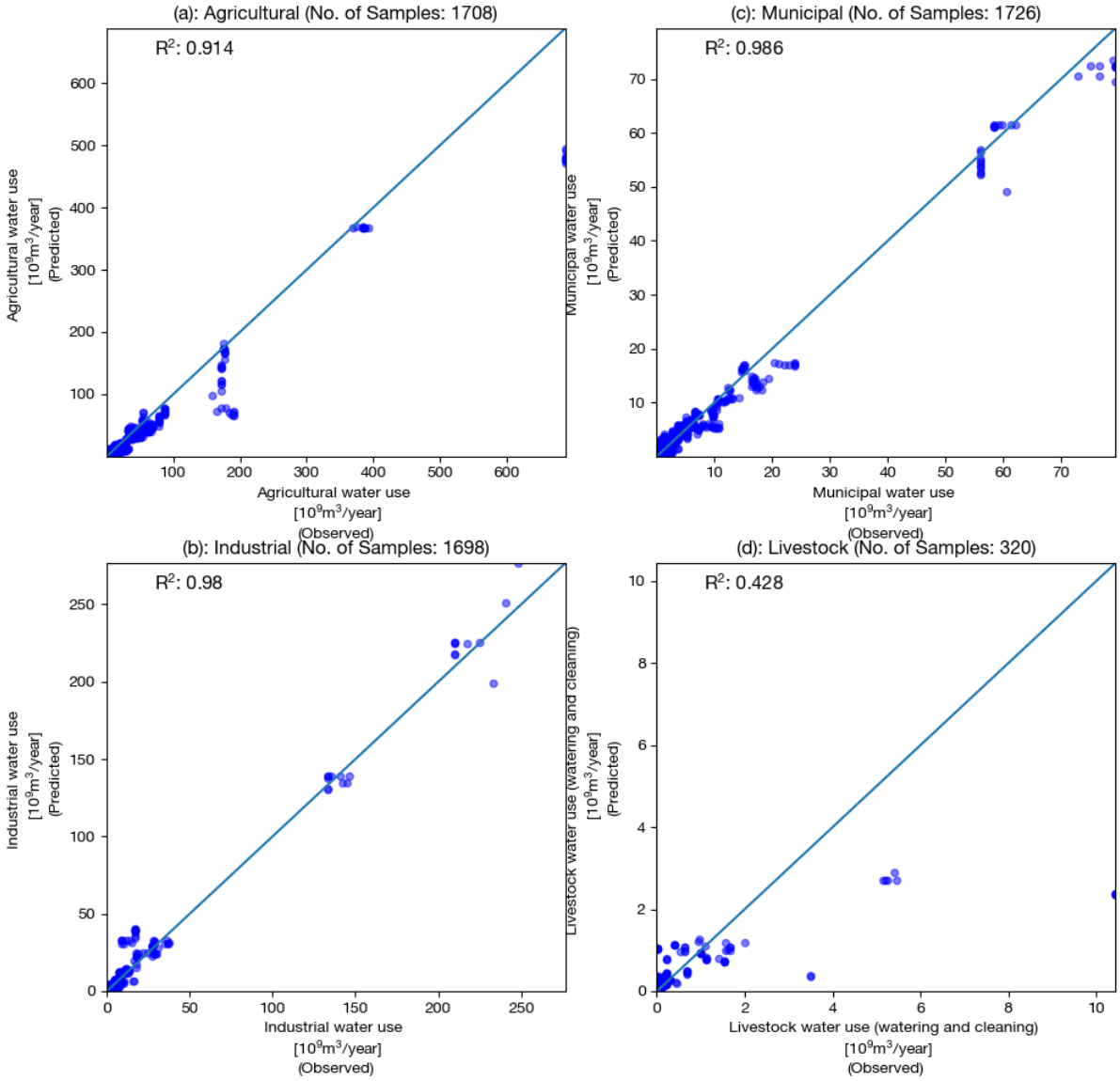

Figure 22 Predictions versus ground truth on validation dataset from time-basis validation. The line in each subplot is a 1:1 line representing perfect agreement or equality between predicted and observed values. The value of the coefficient of determination between predictions and ground truth on the validation dataset is listed in each subplot

3.4 Country-based validation

Country-based validation involved splitting the data based on the countries or regions under consideration. The model is trained on data from certain countries and tested on data from different countries. This validation technique is applicable when there are variations in water use patterns, factors influencing use, or data availability across different countries. Country-based validation helps assess the model's performance across diverse geographic regions and provides insights into its generalizability across different countries.

We randomly selected approximately 30% of countries from each of the three groups (low, middle, and high, as shown in Table 3) to be used as validation samples. The remaining countries were used for training the model. It is important to note that, although we did not train the model specifically for predicting livestock water use on separate low, middle, and high water use samples, we still maintained the same proportion (i.e., 30%) of countries from each group during the country-basis validation. The reason behind this approach is that the target variables in the low, middle, and high groups exhibit a significant difference in magnitude. Therefore, training the model exclusively on samples from the low group would not lead to accurate predictions of the target variables in the high group. By including a representative selection of countries from each group during validation, we aim to assess the model's performance across diverse water use levels and ensure a more comprehensive evaluation of its predictive capabilities.

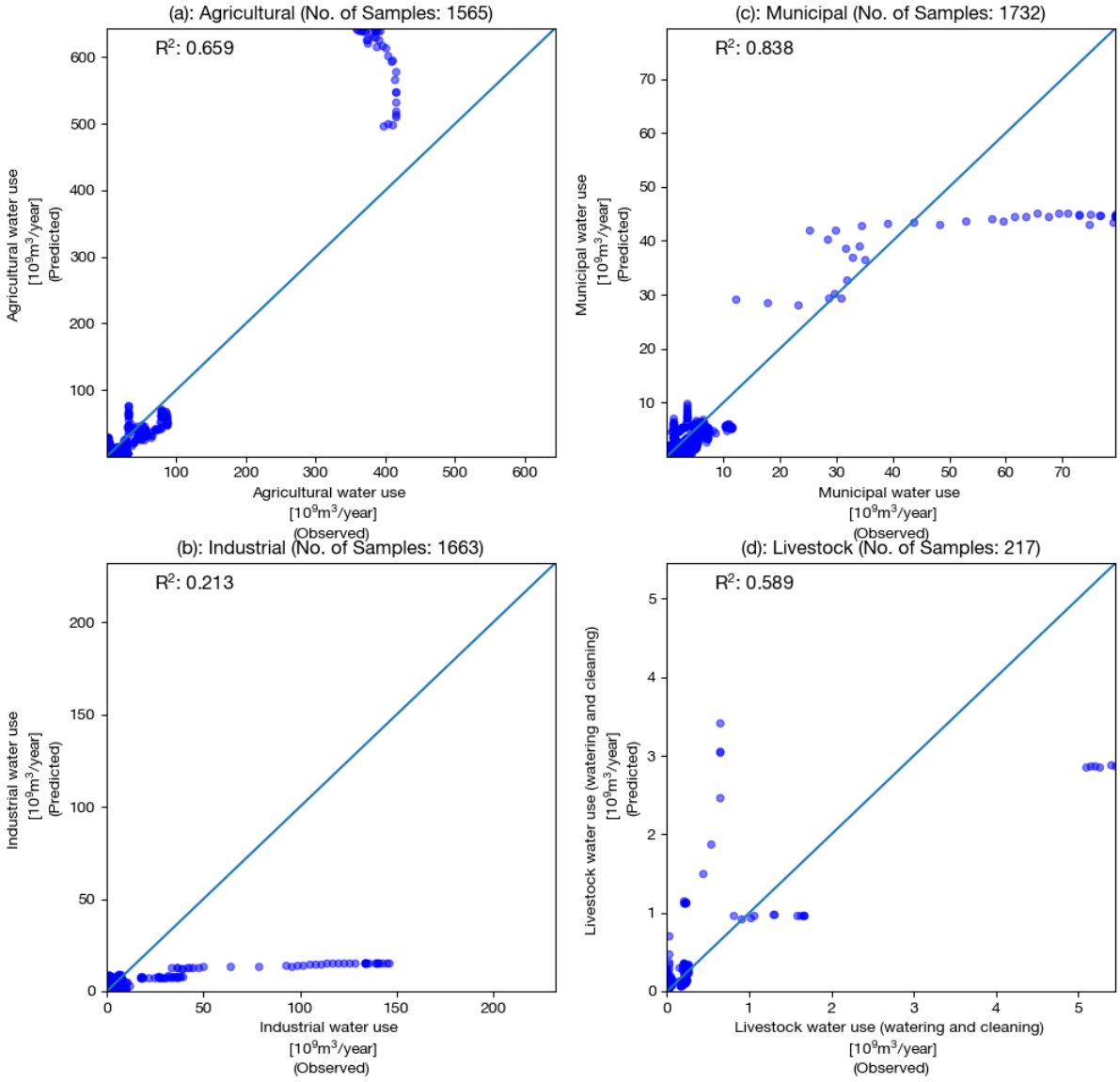

Figure 23 Predictions versus ground truth on validation dataset from country-basis validation. The line in each subplot is a 1:1 line representing perfect agreement or equality between predicted and observed values. The value

of the coefficient of determination between predictions and ground truth on the validation dataset is listed in each subplot

Figure 23 displays a comparison between the model predictions and the observations on the selected validation dataset. The model performs reasonably well for agricultural, municipal, and livestock water use, with the coefficient of determination generally exceeding 0.5. However, for Industrial water use, the coefficient of determination is only about 0.215, indicating a weaker performance.

Overall, the model's performance in the country validation is not as good as demonstrated in the cross-validation and time-basis validation. It particularly struggles with accurately predicting high water use samples. In the case of agricultural water use, the model tends to overestimate high water use samples. In contrast, for the other three water use target variables, it generally underestimates samples with high water use (especially in the case of Industrial water use).

However, if we focus on the prediction of samples with relatively lower water use (as presented in Figure 24), the models performance appears better than for the whole group (Figure 23). The scatter plot of prediction versus observation for lower water use samples exhibits a more balanced distribution around the 1:1 diagonal line.

In conclusion, results indicate that the model has certain limitations in predicting water use across diverse geographic regions, mainly due to the substantial magnitude difference in water use levels across different countries. The model's performance varies for different water use types, and it struggles with accurately predicting high water use countries.

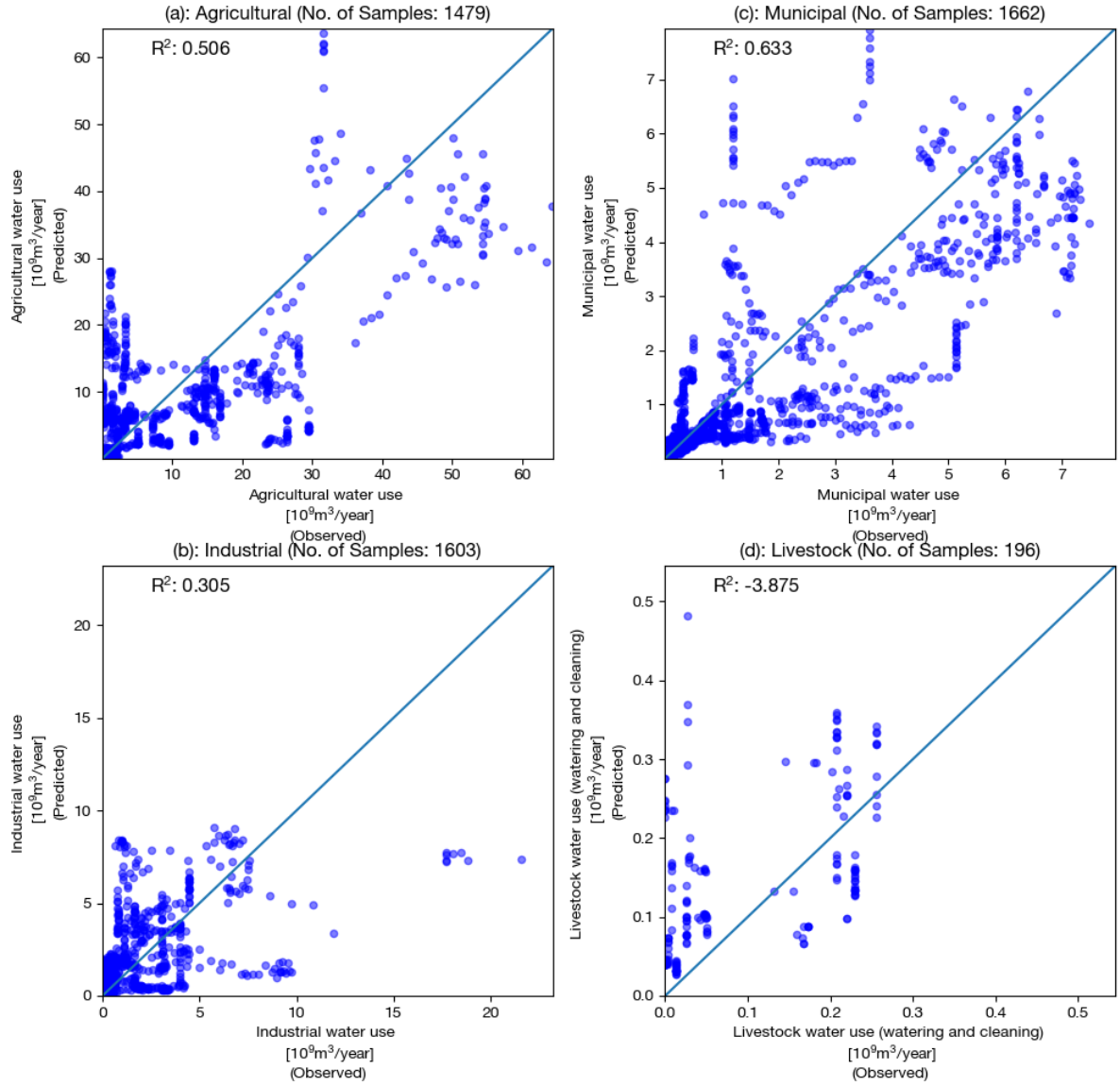

Figure 24 Predictions versus observations on validation dataset from country-basis validation (Zoomed out of Figure 23). The line in each subplot is a 1:1 line representing perfect agreement or equality between predicted v and observed values. The value of the coefficient of determination between predictions and ground truth on the validation dataset is listed in each subplot

Conclusions

This deliverable (D2.1 Assessment of dynamic interactions of water use in the case studies and architecture of improved water use modules) investigates the water use dynamics globally and in the project's case studies, and explores the potential of machine learning algorithms, specifically random forest, for predicting country-level water use. The report considers four types of water use as prediction targets: agricultural, industrial, municipal, and livestock water use. Training the machine learning model involved using yearly data samples spanning 50 years (1965 to 2019) for agricultural, industrial, and municipal water use, while 20 years (2000 to 2019) of annual data samples were used for livestock water use. Seven input features were utilized to predict water use: population, GDP, average annual precipitation, GDP per capita, population density, aridity, and irrigation area.

Our analysis reveals that the global water use distribution is highly skewed and exhibits substantial magnitude variation. To address the prediction of water use, the report presents a 3-phase target transform model structure. This structure employs three separate models trained on groups of divided samples based on the input feature population. The model's performance is validated using cross-validation, time-based validation, and country-based validation.

Overall, the final 3-phase target transform model achieves good performance (with a coefficient of determination over 0.9) in both cross-validation and time-based validation for agricultural, industrial, and municipal water use prediction. However, the model's performance for agricultural, industrial, and municipal water use in country-based validation is not as strong, with a coefficient of determination values of 0.659, 0.213, and 0.838, respectively. The model particularly underestimates water use in samples with high water use. For livestock water use prediction, the coefficient of determination in cross-validation, time-based validation, and country-based validation are 0.923, 0.426, and 0.589, respectively. The relatively poor performance for livestock in time-based and country-based validation may be attributed to the limited amount of available training data.

Overall, the results indicate that the model generally performs well in terms of cross-validation, and the performance in time-basis validation is also promising (except for the prediction of livestock). This suggests that the model could be employed to make predictions for future scenarios as it is robust in time and shows good predictive power. However, it is essential to acknowledge that the model has certain limitations when predicting water use across diverse geographic regions.

The constructed machine learning model will be used to project future water uses under various future scenarios, which will be used as input of the water system models in T2.2 (Advancing state-of-the-art WSMs and linking with regional impact models) that set up model infrastructure at high (1km) spatial resolution. We will also do further validation analysis while downscaling the predicted country level water use data into water use data at high resolution local regions. In particular, this new model can contribute as a screening tool to select a few important scenarios, which will be analysed in a more detailed degree in water system analysis. Moreover, our model could contribute also to predicting water use in various designed scenarios and management pathways in T1.2 (Co-creation of future water scenarios), accounting for different climate, socioeconomic, land use and policy scenarios.

Bibliography

SOS-WATER. (2023, 08 17). Retrieved August 2023, from SOS-WATER: https://sos-water.eu/

van Beek, L. P. (2011). Global monthly water stress: 1. Water balance and water availability. *Water Resources Research, 47*, W07517.

Wada, Y. W. (2014). Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics, 5*(1), 15-40.

Burek, P. S. (2020). Development of the Community Water Model (CWatM v1.04) – a high-resolution hydrological model for global and regional assessment of integrated water resources management. *Geoscientific Model Development*, 3267-3298.

Flörke, M. K. (2013). Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study. *Global Environmental Change, 23*(1), 144-156.

Müller Schmied, H. E. (2014). Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. *Hydrology and Earth System Sciences, 18*(9), 3511-3538.

Stehfest, E. v. (2014). *Integrated assessment of global environmental change with IMAGE 3.0: Model description and policy applications.* Netherlands Environmental Assessment Agency (PBL).

Calvin, K. P.-L. (2019). GCAM v5.1: representing the linkages between energy, water, land, climate, and economic systems. *Geosci. Model Dev.*(12), 677-698.

Hanasaki, N. K. (2008). An integrated model for the assessment of global water resources–Part 1: Model description and input meteorological forcing. *Hydrology and Earth System Sciences, 12*(4), 1007-1025.

Muñoz Sabater, J. (2019). *ERA5-Land hourly data from 1950 to present.* Retrieved January 2023, from ERA5-Land hourly data from 1950 to present.: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview

Dirk N. Karger, S. L. (2022). *CHELSA-W5E5 v1.0: W5E5 v1.0 downscaled with CHELSA v2.0.* Retrieved from CHELSA-W5E5 v1.0: W5E5 v1.0 downscaled with CHELSA v2.0.: https://data.isimip.org/10.48364/ISIMIP.836809.2

FAO. (2023, 07 19). Retrieved from https://www.fao.org/faostat/en/#home

Meier, J. Z. (2018). A global approach to estimate irrigated areas–a comparison between different data and statistics. *Hydrology and Earth System Sciences, 22*(2), 1119-1133.

Burek, P. S. (2020). Development of the Community Water Model (CWatM v1.04)- a high-resolution hydrological model for global and regional assessment of integrated water resources management. *Geoscientific Model Development, 13*, 3267-3298.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*, 81-106.

Pedregosa, F. V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 2825-2830.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*(61), 85-117.

Disclaimer

Acknowledgement of funding

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101059264.

Appendix

A1 Model comparison (Decision Tree vs Random Forest)

We compared the performance of Random Forest with Decision Tree with 5-fold cross-validation (shown in Figure A1). As discussed in Section 2.1, Random Forest algorithm is an ensemble learning method that combines multiple decision trees. By comparing the two, we can determine if the ensemble approach of Random Forest outperforms the single Decision Tree and whether the complexity of Random Forest is justified for the given problem. Figure A1 indicates that Random Forest is consistently better than Decision Tree, especially for the prediction of agricultural, industrial, and municipal water use. The difference between Random Forest and Decision Tree is not as obvious as the prediction of the other three target variables, this might be because, for the prediction of livestock water use, there is much less training data compared with the amount of data available in the other three cases. Hence, the complex model does not gain much advantage. Overall, Random Forest is better than the Decision Tree for predicting water use in all cases, with a better mean model score and less cross-validation uncertainty.

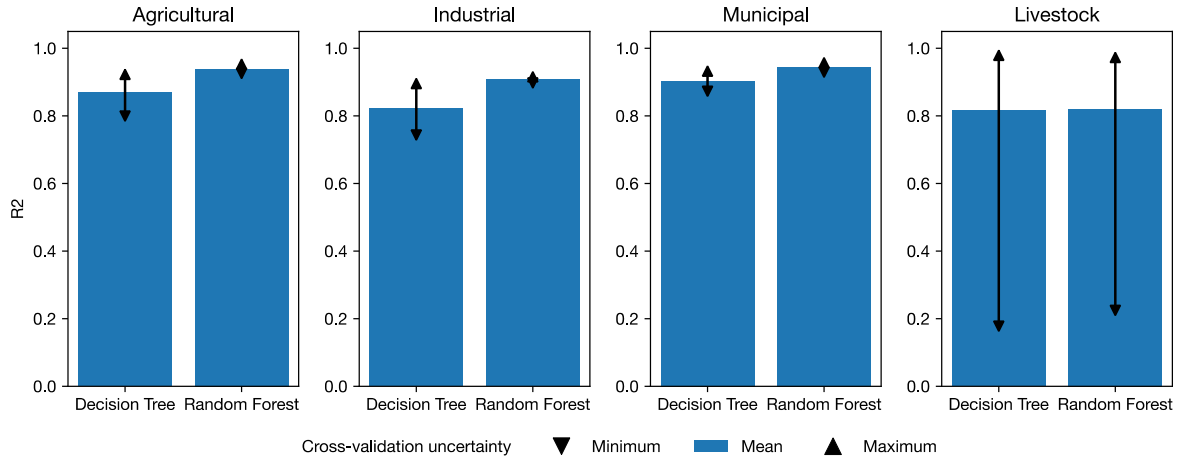

Figure A1 The model performance of Decision Tree and Random Forest for. The mean, maximum and minimum score (measured in coefficient of determination, R2) over 5-fold cross-validation is plotted.

A2 Model comparison (1-model vs 3-phase-target transformed model)

In this section, we present different model structures we have tested to address the prediction bias caused by the skewered distribution of the prediction target variables (as discussed in Section 2.1). We compared five different kinds of model structures: 1) original 1 Random Forest model; 2) two separate Random Forest models for low and high water use samples separately (divided based on criterion in Table A1); 3) three separate Random Forest models for low, middle, and high water use samples separately (divided based on criterion in Table 3); 4) original 1 Random Forest model with target transform; 5) three sperate Random Forest models for low, middle, and high water use samples separately with target transform (as described in Section 2.1.2).

The result indicates that dividing samples into different groups and transferring target variables help improve the model performance (as shown in Figure A2). The combination of dividing samples into three

groups and transforming the target variable before training performs best if we focus on a large amount of low water use samples (as shown in Figure A3). The overestimation issue at high water use samples and underestimated issue at low water use samples have been resolved to a great extent. Hence, our final selected model structure is the three-phase target transform model.

Table A1 The number of samples in each divided group. The samples are divided based on the population size of the sample into low and high two different groups

No. Samples	Low (Population < 500M)	High (Population >= 500M)
Agricultural	5240	85
Industrial	5316	85
Municipal	5518	85

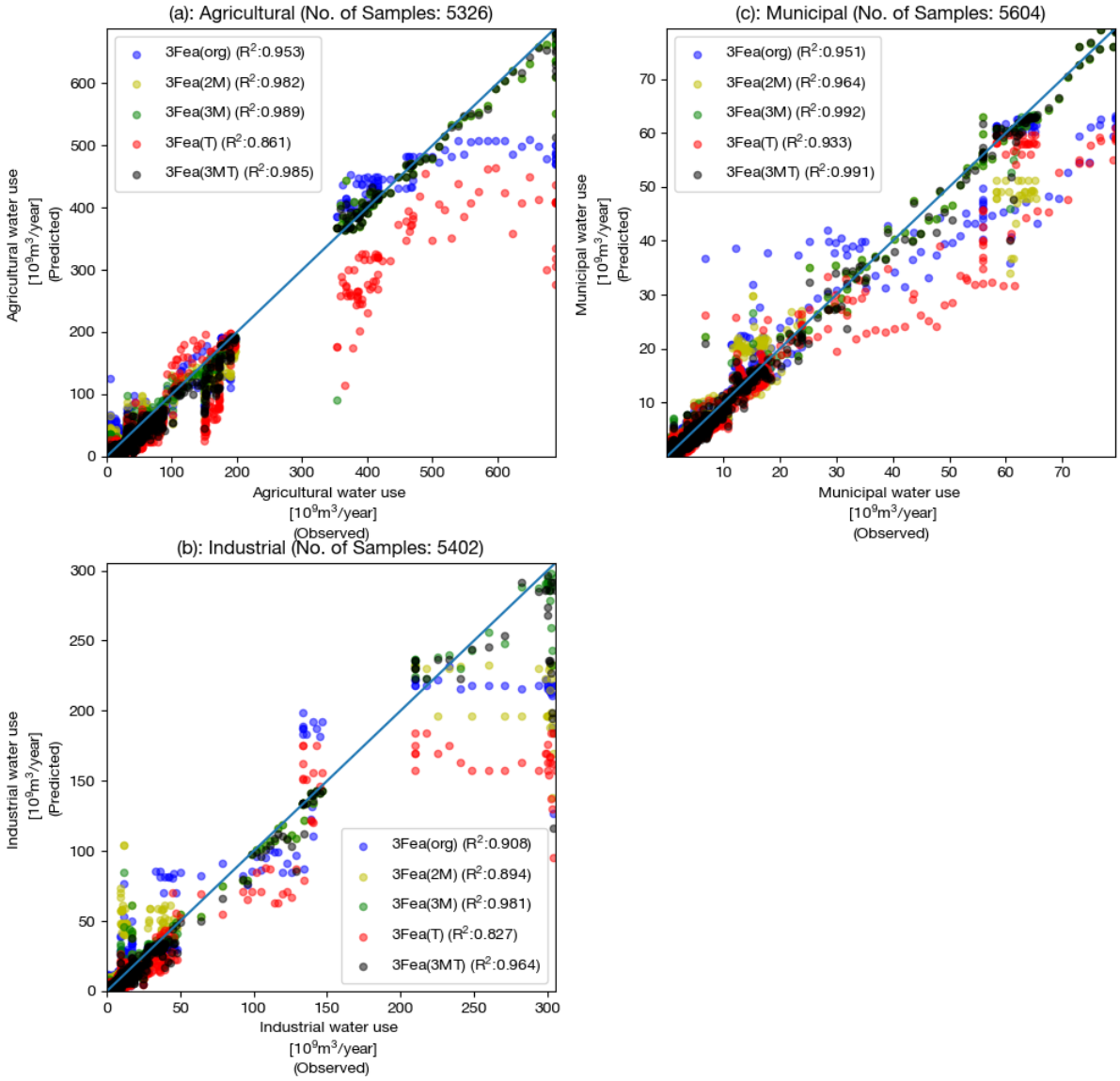

Figure A2 Predictions versus ground truth from 5-fold cross-validation of different models with 3 input features (i.e., GDP, Population, Precipitation). 3Fea(org): original 1 Random Forest model; 3Fea(2M): two separate Random Forest models for low and high water use samples separately; 3Fea(3M): three separate Random Forest models for low, middle, and high water use samples separately; 3Fea(T): original 1 Random Forest model with target transform; 3Fea(3MT): three sperate Random Forest models for low, middle, and high water use samples separately with target transform (as described in Section 2.1.2)

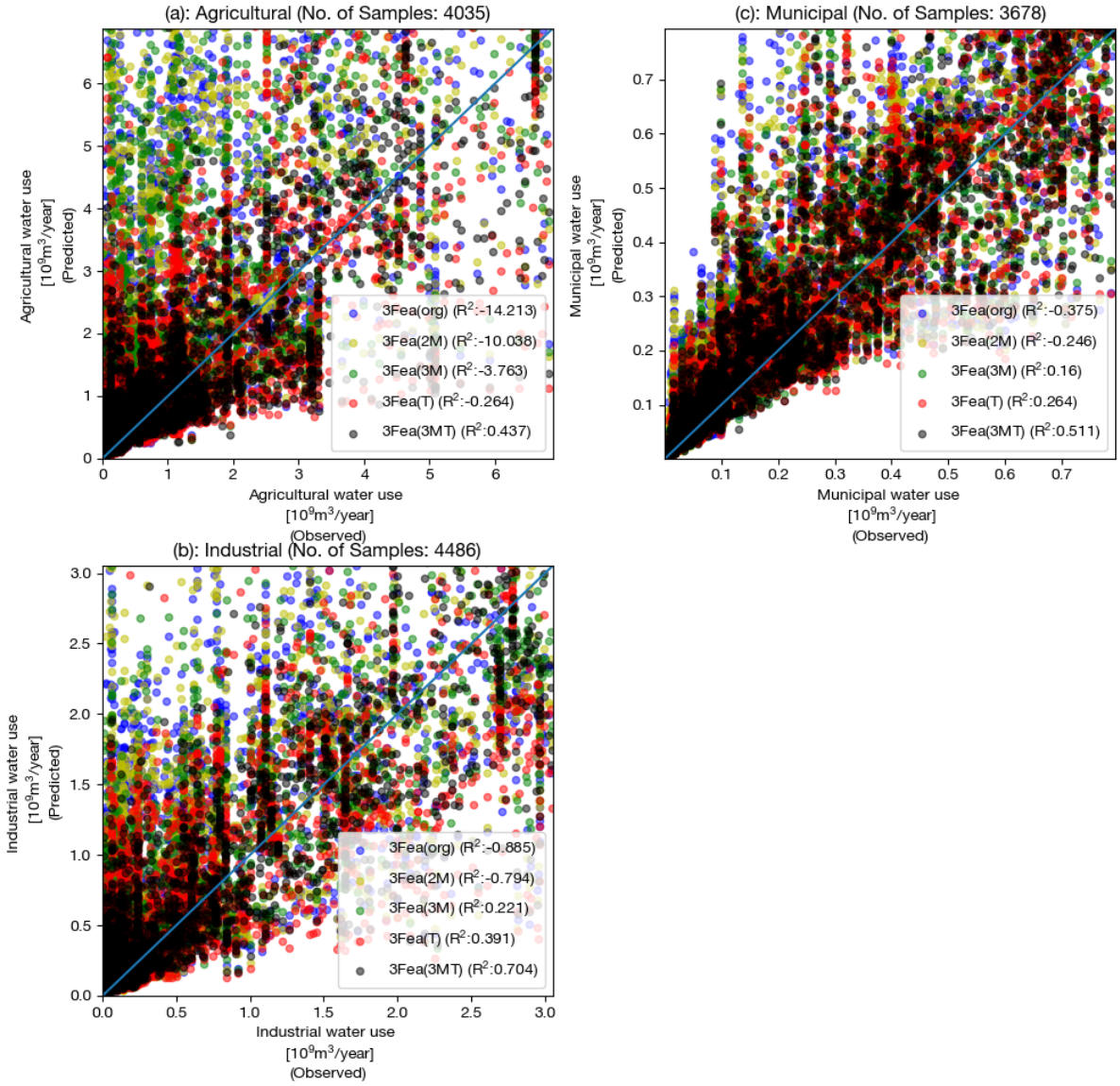

Figure A3 Predictions versus ground truth from 5-fold cross-validation of different models with 3 input features (i.e., GDP, Population, Precipitation) (Zoomed out of Figure A2). 3Fea(org): original 1 Random Forest model; 3Fea(2M): two separate Random Forest models for low and high water use samples separately; 3Fea(3M): three separate Random Forest models for low, middle, and high water use samples separately; 3Fea(T): original 1 Random Forest model with target transform; 3Fea(3MT): three sperate Random Forest models for low, middle, and high water use samples separately with target transform (as described in Section 2.1.2)

A3 Model comparison (with different input features)

This section presents the model performance when using different input features. We show the model of both the original 1 random forest model (1M) as well as the final 3-phase target transform model (3MT). Figure A4 and Figure A5 (zoomed out of Figure A4) indicate that the model with 7 features (i.e., GDP, population, average annual precipitation, GDP per capita, population density, aridity, and irrigation area) generally performs the best. The use of the irrigated area (Dynamic) is only slightly better than the use of the irrigated area (Static). Considering that the established model will be used for the projection

of water use in future scenarios, and there is no projected irrigated area available, we selected the irrigated area (Static) as the irrigated area feature used in the final model.

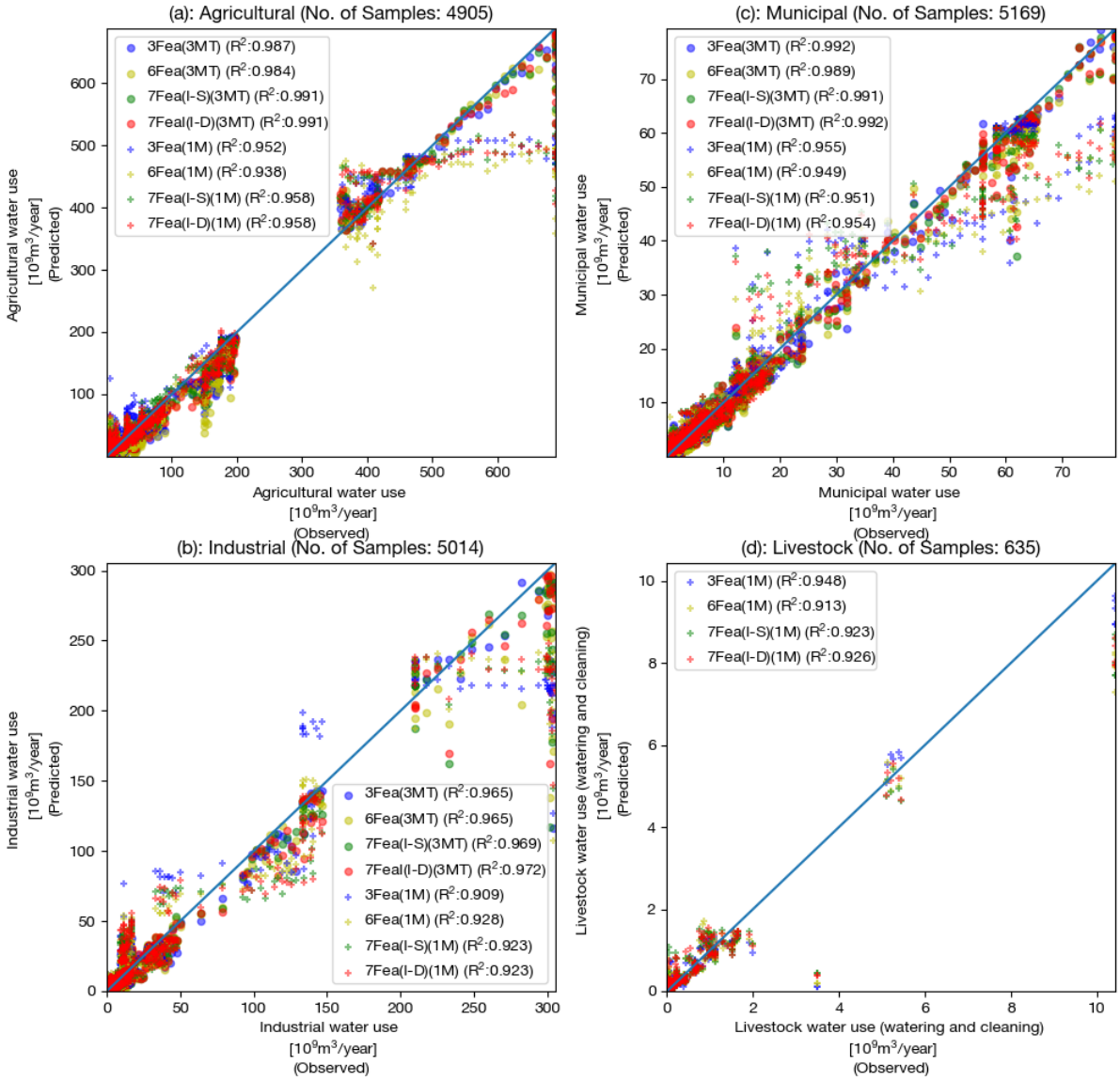

Figure A4 Predictions versus ground truth from 5-fold cross-validation of the original 1 model (1M) and 3-phase target transform model (3MT) with different input features

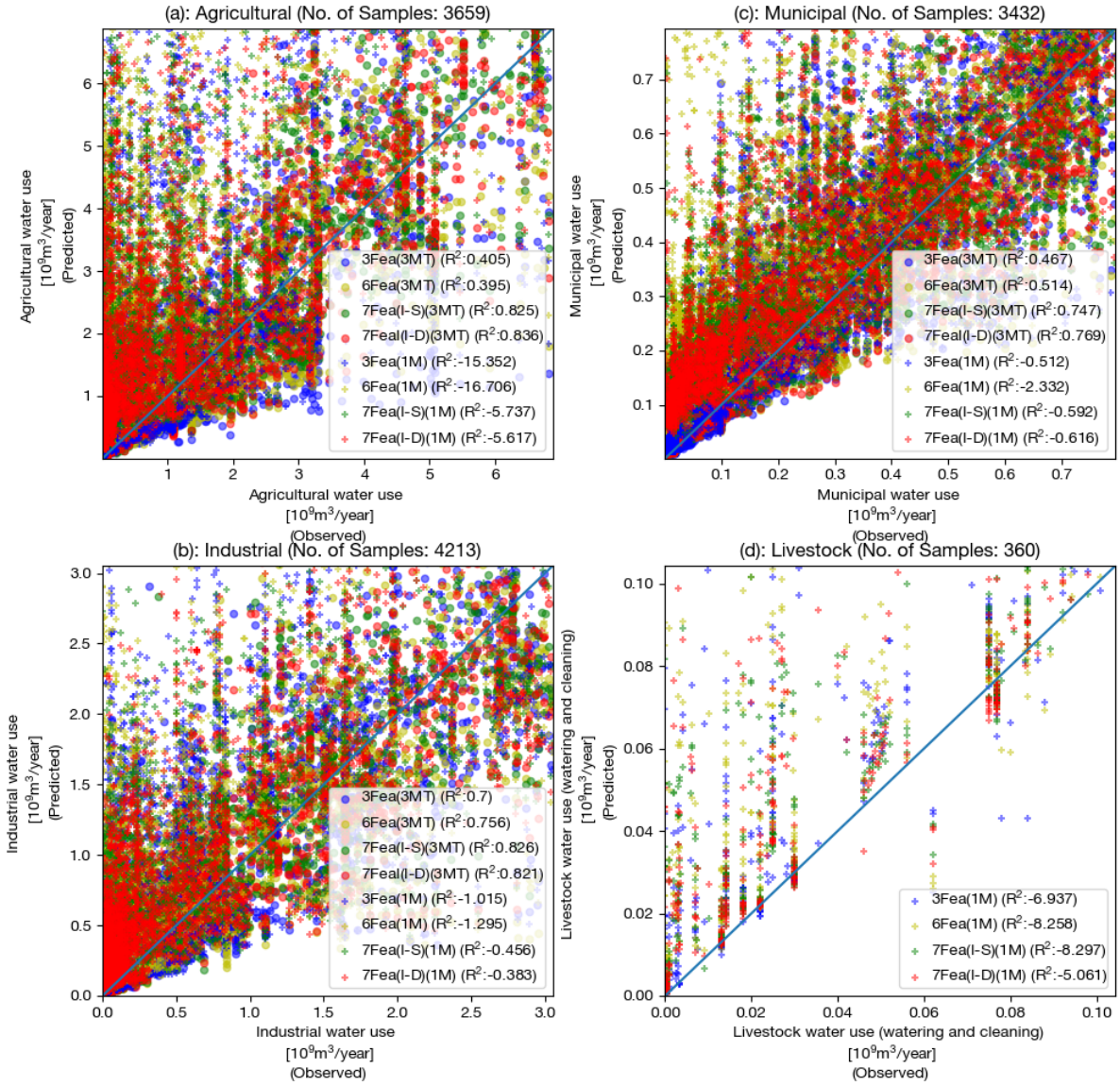

Figure A5 Predictions versus ground truth from 5-fold cross-validation of the original 1 model (1M) and 3-phase target transform model (3MT) with different input features (Zoomed out of Figure A4)