

Global and regional models for identification of cooling technology in thermal power generation for water demand estimations in water-energy nexus studies

Alena Lohrmann^{*}, Christoph Lohrmann, Pasi Luukka

School of Business and Management, LUT University, Yliopistonkatu 34, FI-53850, Lappeenranta, Finland

ARTICLE INFO

Handling Editor: Mingzhou Jin

Keywords:

Sustainability
Power plants
Water stress
Feature selection
Machine learning

ABSTRACT

Water-energy nexus studies aim to connect the process of power generation with the corresponding water demand. The lack of information on the currently installed cooling systems at individual power plants is a challenge for the assessment of the water use on the power plant-level, which complicates decision-making in water management, especially in water-stressed regions. In this study, we investigate the spatial and temporal trends in cooling technology installations globally. Based on that, we propose a machine learning model for cooling technology identification on a regional and global level, which uses a combination of feature selection and classification algorithms. The global model demonstrates an average test set accuracy of 85.42%, which corresponds to only a minor underestimation of the actual global water footprint of 1.78% when the cooling technology and water footprint of individual power plant units is unknown. Apart from that, a special emphasis was placed on regions characterized by high and extremely high water stress, where mistakes in water policy planning and water management may lead to an unsustainable water use or even to an overexploitation of water resources. In these regions, the calculated test set accuracy was 80.83%, which is significantly larger than the average accuracy of a majority class model. The results and the method proposed in this study enable cooling system identification in individual power units using information available from other sources, such as the water stress score or seasonal freshwater availability in the region.

1. Introduction

Since 2012, “water crisis” was consistently included in the Top-5 Global Risks by impact in the yearly Global Risks Report released by the [World Economic Forum \(WEF\) \(2021\)](#). Even according to very modest estimations, 47% of the global population (or 3.6 billion) already live in areas that suffer from water scarcity at least one month every year (United Nations World Water Assessment Programme (WWAP) and [United Nations World Water Assessment Programme and UN-Water, 2018](#)). This is the result of a constantly increasing demand for water, food, and energy for a growing population and the economy, as well as the depletion of water resources ([Boretti and Rosa, 2019](#)).

Moreover, the global power sector is currently the second largest consumer of freshwater resources after agriculture ([AQUASTAT, 2020](#)). According to some estimates, the energy-related water demand can reach a share as high as 40% of a country's total water demand ([Maupin et al., 2014](#)). Thermal power plants (such as coal-, gas-, oil-fired power

plants and nuclear power plants) use water for cooling purposes. While thermal power plants account for about 80% of the water demand in the global power sector ([WWAP, 2014](#)), their cooling is responsible for the largest share of their water demand. There are various types of cooling systems, which differ in their construction and in the principle of cooling: Studies typically differentiate surface-water cooling (once-through cooling, cooling ponds), wet cooling towers, as well as inlet and dry cooling systems ([Macknick et al., 2012](#)). The type of the cooling system (along with environmental conditions, such as plant elevation, air temperature, humidity, wind speed, and ambient water temperature) affects the total water demand of the power generation process as much as the generation technology itself ([Diehl et al., 2013](#)), and each cooling type is associated with a different water demand per unit of generated energy ([Macknick et al., 2012](#)).

[Larsen et al. \(2019\)](#) discussed the challenges related to cooling technology data availability in water-energy nexus studies. In particular, power plant operators typically neither report on the amount of water

^{*} Corresponding author.

E-mail address: Alena.Lohrmann@lut.fi (A. Lohrmann).

<https://doi.org/10.1016/j.jclepro.2022.134842>

Received 15 June 2022; Received in revised form 11 October 2022; Accepted 20 October 2022

Available online 27 October 2022

0959-6526/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

used during the power generation process nor the type of the installed cooling system (Luo et al., 2018). Moreover, many widely used inventories of electricity generation (for instance, European Network of Transmission System Operators for Electricity (ENTSO-E) (2022)) and power plant databases (for instance, Enerdata (2020)) do not include information concerning the installed cooling system. The online power plant database provided by Global Energy Observatory (GEO) (2018) as well as the commonly used databases of S&P Global (2016) and GlobalData Plc (2020) provide the cooling technology information only for a limited number of entries/power plant units. This poses a challenge for the assessment of the water use on the power plant-level and in the power sector in general.

Another challenge lies in the availability of the cooling technology data for planned and announced power plants, and for power units under construction. According to the information provided in the GlobalData database (GlobalData Plc, 2020), there are plans to commission at least 2.6 TW of fossil and nuclear power plant capacities worldwide until 2050 despite of the commitment of more than 190 countries to substantially reduce GHG emissions to tackle climate change and its negative impact (United Nations (UN), 2022). Generally, no information concerning the future installed cooling technology for these power plants is announced in open reports or databases.

Misclassification of cooling technologies may considerably affect the results of the regional water audit, which, consequently, may result in inadequate allocation of water resources between sectors of the economy and population. One example of such misclassification may be an incorrect assignment of the cooling technology to, for instance, individual gas power plants. Depending on the cooling type, gas power plants consume between 0.008 m³ (dry cooling, combined cycle) and 3.127 m³ (tower cooling, steam) of water per 1 MWh of generated electricity (Macknick et al., 2012). This example implies that the misclassification of the cooling technology on a regional- or global-level may lead to a considerable under- or overestimation of the amount of water used by power plants by a factor of up to (about) 400. This may be especially problematic for water footprint estimations of (1) power plants characterized by large capacity and high full load hours (base load power plants), which are generally associated with a high water demand for cooling, and (2) power plants located in countries facing high and extremely high water stress (for instance, Qatar, Israel, Lebanon, Iran and Jordan) (Water Resources Institute (WRI), 2021), where such miscalculations could potentially lead to an incorrect allocation of the already limited water resources between industries and the population.

Water-energy nexus studies aim to connect the process of power generation with the corresponding (historical, current, or projected) water demand. In order to overcome the cooling technology data limitation, in previous water-energy nexus studies the missing information on the cooling system for individual power plants was assigned by following one of the following strategies:

- i. Using the shares of cooling technologies in the region found in various literature sources (Larsen et al., 2019)
- ii. Using the shares of cooling technologies for each generator-fuel combination derived from a set of known power plants in each specific country (Spang et al., 2014)
- iii. Using the shares of cooling technologies for different installation years derived from a set of known power plants in each specific country (Vassolo and Döll, 2005)
- iv. Based on the geographical location of individual power plants in relation to various water bodies: the cooling type was assigned to all power plants based on their proximity to large rivers (Vassolo and Döll, 2005) and the ocean coastline (Larsen and Drews, 2019)
- v. Based on the visual investigation of aerial images of individual power plants (Luo et al., 2018)
- vi. Using the cooling technology classification models based on machine learning (Lohrmann et al., 2021b).

The first four above-mentioned methods typically do not consider all the individual (technological) characteristics of the power plant units and the geographical attributes of their location, although the combination of these features may affect the choice of the cooling system in each specific case. The fifth method, however, cannot be applied for planned and announced power plants since their aerial images are not (yet) available by the time of the scientific investigation. Moreover, previous research focused either on the country- or region-level allocation or only on the proximity to the coastline when assigning cooling technologies but did not consider (other) power plant specific information for the assignment. Thus, the accuracy of the cooling technology identification of the above-mentioned approaches on the individual power plant-level is rather low – about 35% for a five-class problem (Lohrmann et al., 2021b). The fifth method implies the need to identify the cooling technology of each power plant by an expert, which is a time-intensive process. However, it does not require the imputation of these unknown values based only on cooling technology shares in a region/globally or the proximity to water bodies, as these may be oversimplistic and likely more inaccurate than a power plant unit specific procedure. A previous study by (Lohrmann et al., 2021b) proposed a hybrid method for cooling system identification based on machine learning techniques. The cooling technology classification model (based on a decision tree classifier) developed in this study, has demonstrated low performance for power plants located in regions/countries with a small number of observations (power plants) due to the challenge this poses for training this type of model. Moreover, it does not consider local factors (such as, for instance, availability of sufficient water for cooling at the power plant location), which may be useful for the cooling technology identification. Hence, a more universal approach should be developed to overcome the above-mentioned limitations and, in addition, to ensure a high accuracy of the cooling technology identification for individual power plants.

To address this call, the aim of the study is to propose a new model for cooling system identification, which can be widely applied for water demand estimations in water-energy nexus studies. This model should:

- (1) Use widely available information from open sources
- (2) Be applicable for the cooling technology identification on both, the global and regional level
- (3) Enable the cooling technology identification for already existing (active) power plants and for planned capacities.

In contrast to the existing approaches, the model presented in this study is novel in three aspects. First, to our knowledge, this is the first study to apply machine learning techniques for the cooling technology identification globally and specifically in water stressed regions. Second, the presented model enables the identification of the cooling type on an individual power plant-level (compared to other approaches, where the assignment is done for aggregated capacities). This allows to conduct an assessment of the water footprint on all levels: from individual power plant to river-level, province-level, country-level and, finally, globally. Third, the proposed classification algorithm considers local factors (such as the availability of water for cooling, climate conditions, economic development of the country, etc.), which were not included in previous studies, and which may be useful for the cooling technology identification.

The study also contributes to build a better understanding of factors that are important for the cooling system identification on global and regional levels. In addition, it broadens the area of machine learning application.

This study will cover two model perspectives: the global model and the region models. The idea behind introducing both global and region-specific models was to give practitioners two modelling options, depending on the information available to them. The region models are specific to each region and, thus, may find region-specific patterns and only require information on power plants in that region, but may not be

robust when only few power plants are available in the region to train the model on. In contrast, the global model may be capable to find globally generalizable patterns that span regions and will not encounter the problem of having too few power plants for model training. However, it requires information on power plants globally and may be oblivious to patterns within a region that do not follow the global patterns. By differentiating between these two modelling perspectives, it will also be possible to analyse whether different features may be relevant for the cooling technology identification in different parts (regions) of the world.

This study is structured as follows: First, the data and variables contained in this study are depicted. This includes a presentation of features, which might influence the cooling system installations, and which can be used for the cooling technology identification but were not applied in previous studies. Second, the spatial and temporal trends in the global cooling technology installations are discussed. Then, based on the results of our investigation, we propose a model for the cooling technology identification, which applies machine learning techniques. In that context also the relevant features for the classification are presented and discussed, which are obtained through feature selection. The proposed classification model is also tested on the available power plant dataset (Lohrmann et al., 2019) and its performance is compared with the accuracy of a commonly used so-called majority class approach. We also investigate the performance of the model for regions associated with high and extremely high baseline water stress. Lastly, we identify the limitations of the proposed model and discuss its possible applications.

2. Methods

2.1. Local features influencing cooling technology selection

This study analyses features that may be relevant for the cooling technology identification on the global and/or regional level. However, it is crucial to mention that on the power plant level, the selection of the cooling system represents a non-trivial task since every cooling technology has advantages and drawbacks, which should be balanced before selecting the cooling system.

Several reports investigate features influencing the cooling technology selection. For instance, the report provided by the company EcoFys (2014) discusses features that may influence the optimal design of the cooling system, which can be, but are not limited to:

1. Thermal efficiency of the power plant.
2. Temperature and quantity of the heat medium, which needs to be cooled in the power generation process.
3. Environmental considerations, for instance, fish entrainment, emissions to surface water, air emissions, noise.
4. Economic considerations, meaning the cost of the cooling system. This includes capital costs, operation and maintenance (O&M) costs, and so-called "penalty costs". The penalty costs indicate the influence of the choice of the cooling system on the efficiency of the power plant (California Energy Commission and Electric Power Research Institute (EPRI), 2002).
5. Regulatory or permission constraints.
6. Geography of the power plant location and seasonal effects: e.g., type of water available (sea- or freshwater), availability and temperature of the water used for cooling and its seasonal variation, ambient temperature and humidity, and their seasonal variation.

Most thermal power plant operators consider information concerning the thermal efficiency, load and the characteristics of the heat medium as commercially sensitive information. This information is typically neither available in free nor commercial power plant databases and, thus, cannot be obtained for the cooling technology identification on the global and/or regional levels.

For different regions/countries the set of the most relevant features for the cooling system determination may vary. For instance, the report by EPRI and the California Energy Commission and Electric Power Research Institute (2002) indicate that the amount of water consumed, and the cost of the cooling system are relevant for power plants in California (which is classified as a water-constrained region (Water Resources Institute, 2021)).

Including the above-mentioned features into the classification model might potentially benefit the accuracy of the cooling system identification, especially on the individual power plant level. However, as stated previously, information on some of the presented features is not available for analysis. Hence, while selecting the features for the cooling system identification for this study, we:

1. Selected the features that can be found in open data sources on global level (this mostly concerns point 6 "Geography of the power plant location and seasonal effects" of the above-mentioned features)
2. Combined strategies, which were used for the cooling system identification in previous studies (i-v, listed in the *Introduction* section)
3. Added other features to evaluate their relevance for the cooling system identification.

More information concerning the features selected for the analysis is presented in the next section.

2.2. Data

The GlobalData database (GlobalData Plc, 2014), which is used in this study, contains information regarding 13'865 individual thermal power units with a capacity of 50 MW and higher that correspond to 5'778 power plants worldwide. The database was complemented by Lohrmann et al. (2019) using satellite imagery to add information concerning the cooling technology that is implemented in each power plant unit. From the database, the following features (= variables) were selected: the "Cooling technology" as the discrete dependent variable to be explained and to be predicted, and as explanatory variables the "Power plant capacity (total active)", "Fuel used in power plant", "Year Online", "Type of boiler" and "Seawater-cooling". Apart from the assumption that these features are relevant for the development of the cooling technology identification model, the above-mentioned features are typically available in free and commercial power plant databases, which supports the future implementation of a model based on these variables.

In addition to the features acquired from the initial dataset, the other explanatory variables were obtained from other open data sources. For convenience of presentation, these variables were divided into two groups. The first group of features is related to the availability of freshwater resources at the power plant location. The assumption for the selection of these variables is that locations with less freshwater resources available will correspond to cooling technologies that will require and strain water resources less.

- "Freshwater total" is the estimated amount of yearly available renewable water resources per country, in cubic metres (Food and Agriculture Organization of the United Nations (FAO), 2021). According to Lohrmann et al. (2019), approximately 66.6% of the currently active global thermal power plant capacity is estimated to use freshwater for cooling. This share is unlikely to change in the future, even in case of a future energy transition towards a more extensive use of renewable energy technologies. Therefore, the total amount of freshwater resources in the country might be a determining factor for the selection of suitable cooling technologies for at least two thirds of the thermal power capacities worldwide. This variable, however, does not consider the changes in the availability of freshwater resources, which may vary over time (seasonally)

and/or which might be impacted by other external factors (such as other water-intensive sectors of the economy e.g., agriculture).

- “*Freshwater per capita*”, which, in contrast to “*Freshwater total*”, indicates the availability of freshwater resources in cubic metres *per capita*. The country-specific timeseries data was obtained from (Food and Agriculture Organization of the United Nations, 2021). This variable was selected since it depicts potential changes and challenges of water management in each country.
- “*Seasonal variability*” describes the region-specific variation in the freshwater availability (including both renewable surface water and groundwater) between the months of the year. Seasonal variability is caused by changes in monthly precipitation and evaporation. It ranges from zero to five, and higher values indicate regions characterized by stronger variations in the available water supply within a year. The values of the seasonal variability for different countries were obtained from Water Resources Institute (2021).
- “*Water stress score*” is an indicator developed by Water Resources Institute (2021), which characterizes the potential exposure of a specific area to various water risk factors, such as flood and drought risks, and baseline water stress. As with the seasonal variability, the water stress score ranges from zero (no water stress) to five (the highest water stress). This study considered both, the country-specific and province-specific water stress scores derived from Water Resources Institute (2021). Initially, the country-specific water stress scores were used to impute missing values in the province-specific data when (1) the information concerning the province-specific water score was not available in the WRI database, or when (2) the initial GlobalData dataset did not contain information concerning the province-specific location of a power plant.
- “*Agricultural water withdrawal as percent of total renewable water resources of the country*”. In the database obtained from Food and Agriculture Organization of the United Nations (2021), this variable ranges from 0.01% (e.g. in Croatia) to 2460% (in Kuwait). Values above 100% indicate that the agricultural sector uses non-conventional sources of water (desalinated water or wastewater). It may appear counter-intuitive to consider agricultural water withdrawals instead of industrial water withdrawals. However, as reported by Food and Agriculture Organization of the United Nations (2020), agriculture is the largest freshwater user in the world, accounting for more than 70% of the global freshwater withdrawals. Hence, the agricultural water use in a specific country may indicate how much water is “left” for other purposes, such as municipal water use or power generation.

The second group includes variables, which do not fall within the above-mentioned category linked to the availability of freshwater resources but are also assumed to be relevant for the cooling system identification.

- “*Urban population*”, which stands for the percentage of the population in a country residing in urban areas, was obtained from (United Nations, 2018). Previous studies have demonstrated that the ongoing increase in the world’s urban population, rising prosperity and changes in the water and energy supply impact both, energy consumption (Zhao and Zhang, 2018) and water use (Flörke et al., 2018). This might indirectly affect the choice of cooling technologies for thermal power generation due to (1) limited water availability for thermal power generation (as well as for other industries), and (2) the preference for larger (and, typically, more water-demanding) power generation facilities in urban areas than in rural areas.
- “*GDP per capita*”. The rising prosperity of countries may influence the selection of cooling technologies in thermal power generation. In particular, the implementation of certain cooling systems (such as dry cooling) can be associated with substantial capital and operational costs (Maulbetsch and Stallings, 2012). The values of “*GDP per capita*” were obtained from (The World Bank, 2020) and were for

most countries only available from 1960 to 2015. Since the power plant database used in this study covers the period from 1923 to 2015, the data was complemented with information gathered from other sources (for instance, for the United States the data was collected from (Statista. The Statistic Porta and I, 2021)). In cases when such information was not available, the historical trend (yearly changes) of “*GDP per capita*” for the United States was used to estimate past values of “*GDP per capita*” for a country.

- “*Days of warm weather*”. The technologies currently used for thermal power generation are sensitive to air temperatures, wind speed, water flow and water temperatures (Diehl et al., 2013). These factors determine the cooling system design at the stage of the power plant construction and the actual amount of water required for cooling. The authors of this study acknowledge that the use of the actual site-specific temperature and wind data might provide a more accurate representation of these factors. However, this information is typically not widely available. Hence, in order to overcome this limitation and deploy information that may function as proxy for the site-specific temperature and can be applied in other studies, it was decided to use the values of the average annual number of days with temperature higher than 20 °C, which were calculated by Sadowska et al. (2019) for each country.
- “*Electricity price*”. The country-specific electricity prices in US dollar per kWh were derived from (GlobalPetrolPrices, 2020). The applied electricity prices include all components of the electricity bill such as the cost of electricity, transmission and distribution costs, and taxes. Since electricity prices are typically different for households and businesses, the average of the electricity prices of households and businesses was used. This variable was included in the study since the water costs for cooling are typically included in the operating costs of power generation facilities. According to the U.S. Energy Information Administration (EIA) (2021), operating costs are one of the components affecting electricity prices.
- “*Corruption Perceptions Index*”, or CPI, is reported by Transparency International (2019). It characterizes the perceived level of public sector corruption in each country and ranges from zero (highly corrupt countries) to 100 (so-called “very clean” countries). The reason of including the CPI indicator in this study was based on the assumption that some decisions concerning the design of a future power plant (in particular, the choice of the cooling system type) might be affected by corruption (and cannot be explained by other features). For this study, the CPI variable was transformed into a categorical variable through the following steps: First, the mean CPI for each specific country for the years 2012–2015 was calculated. The reason of not including the CPI scores from previous years was based on the fact that, according to Transparency International (2019), the CPI scores from previous years were calculated using a different methodology and are not compatible with the recently reported CPI scores. Second, based on the calculated CPI scores, all countries were divided into three categories: countries with high level of public corruption (1st quartile, CPI lower than 30.1), countries with low level of public corruption (4th quartile, CPI higher than 60.1), and countries that have neither very high nor very low public corruption (between the 1st and 4th quartile).

All continuous variables were normalized using the z-score normalization technique, whereas all categorical variables were converted into dummy variables, which were assigned the values 1 and -1.

2.3. The scope (geography of study)

As mentioned previously, many variables considered in this study represent country-specific data. However, for large countries which are characterised by diverse climate conditions and different availability of freshwater resources within provinces/states, it is advisable to account for such differences since the choice of the cooling system for power

generation might be affected by these features. In this study the world is divided into 145 regions, which either are equal to countries, denote provinces/states in a country, or represent multiple countries. For instance, the twelve largest countries (by area and population) were split into smaller regions, which represent either already existing individual provinces or combinations of neighbouring provinces. In particular, the United States was divided into 14 regions and China was split into 8 regions. In contrast to that, some regions are combinations of countries that have shown high levels of economic and energy cooperation in the past (e.g., the Baltic region (BLT) combines Estonia, Latvia and Lithuania). This geographical division was previously used in other water-energy nexus studies, such as (Lohrmann et al., 2019, 2021a). Supplementary Table SM1 contains information concerning the countries/provinces included in each specific region.

In addition to the region-based models that consider each of these regions separately, the cooling system classification in this study is also performed on the global level using the data on all power plants worldwide.

2.4. Model training, selection and assessment

Feature selection is commonly referring to the removal of features (=variables) from a data set that are irrelevant or redundant, with the aim to retain only those features that are relevant for the task at hand (Bolón-Canedo et al., 2012). There are multiple motivations to apply feature selection and, thus, reduce the number of features used for model training. These include (1) improving (or at least not considerably decreasing) the model performance (e.g. classification accuracy) by removing features that may act as noise, (2) reducing data storage requirements and computational cost, (3) obtaining more easily interpretable and visualizable data, and (4) increasing the speed of model training and obtaining more simple as well as robust models from the data (Guyon and Elisseeff, 2003).

In this research, a hybrid feature selection approach will be pursued,

which means that multiple feature selection algorithms of different types (Liu and Yu, 2005) will be used sequentially to combine the advantages of the different types. In particular, linear correlation will be used as a fast filter method to initially remove redundant features in order to reduce the computational cost and time for the subsequent application of the more sophisticated and computationally more demanding “differential evolution feature selection” (DEFS) (Khushaba et al., 2011) wrapper method.

For the filter, features which show high absolute pairwise correlation (>80%) with other features in the data were iteratively removed (see Supplementary Figure SM1) since they are unlikely to add additional information to the set of features. In particular, if two features had an absolute correlation of 80% or more, then the feature with the higher average correlation with all other features was removed. The resulting smaller feature set was then used as input to the computationally more expensive DEFS wrapper. The K-Nearest Neighbor (KNN) algorithm is used as the classifier in this study and is also used with the DEFS wrapper as this wrapper method uses the classification accuracy to guide its selection of the best feature subset. The KNN classifier uses a distance measure (e.g. Euclidean distance) to determine for each new observation (e.g., power plant unit) the closest k observations from the training data and to assign each new observation to the class that most of these nearest neighbors belong to (majority voting) (Hastie et al., 2009). This classifier is very intuitive, relatively fast - especially with not too large sample sizes as in this study, and, in contrast to many classification algorithms (e.g., Naïve Bayes, neural networks, support vector machines, decision tree, random forest), is a “lazy” learner that does not require parameter training and does not need to be retrained when new observations become available.

The complete process containing model selection (feature subset selection and hyperparameter selection) and model assessment is illustrated in Fig. 1. As shown in the flow chart, the partitioning of the data was conducted in two steps: Firstly, the data was split using a 5-fold cross-validation with stratified random sampling. Consequently,

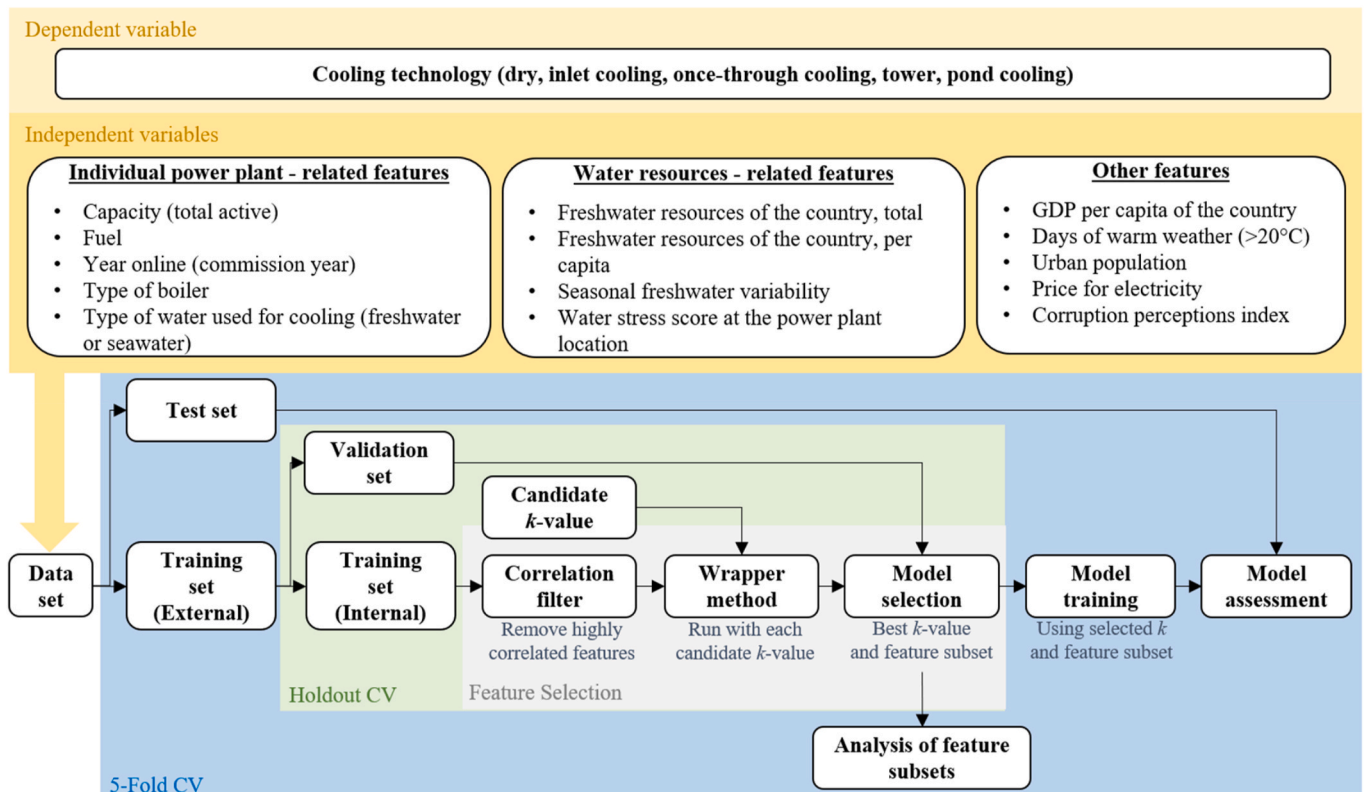


Fig. 1. Cooling technology identification model using hybrid feature selection and a k-nearest neighbor classifier.

regions with less than 5 observations (power plants), had to be excluded from the analysis. In addition, regions, which contain only one cooling technology (class), were left out to avoid training a classifier on data with a single class. As a result, 20 regions (or 14% of the total number of regions presented for the analysis) were excluded at this part of the process. Next, in each fold, the data was further divided into a training (75%) and validation (25%) set using random stratified holdout partitioning. Stratified sampling ensures that the training, validation and test sets all contain similar shares of the cooling technology classes.

It is noteworthy that the DEFS wrapper requires to specify the number of features that should be selected, and the KNN classifier requires the specification of the number of nearest neighbors k (hyperparameter). Since it is not known in advance how many features will result in the highest classification accuracy and the optimal number of features for the cooling technology classification may vary in each region, the number of features was varied from 1 to 25 (number of available features). For each number of features, the DEFS algorithm was run for 50 iterations. Since many regions contain a small number of power plants (observations) and some cooling technologies have a comparably small share in that region, the KNN classifier was run only with small nearest neighbor values of $k = 1, 3$, and 5. This ensures that (1) the classification can be performed even if only few observations are available and that (2) cooling technologies with a small share in the region can still be predicted (majority voting). The validation accuracy was used to select the best combination of feature subset and k hyperparameter. After this step of model selection, the KNN classifier with the selected k parameter and feature subset was trained on the training data (external) and deployed to predict the power plants in the independent test data set to calculate the test accuracy. More information concerning the model's input and output parameters is presented in Supplementary Information (Figure SM2).

One additional step in the flow chart is the analysis of the feature subsets, which were selected during model selection in each of the five folds. The frequency of each feature contained in the feature subsets was counted to determine which features were most often selected and, thus, appear most relevant. For categorical variables that were transformed into multiple dummy variables (for instance, three dummy variables representing four types of fuel in the categorical variable "fuel type"), the highest count for the most frequently selected dummy variable was used as the frequency for the original categorical variable. This was done to keep the results of numerical and categorical variables comparable when discussing the set of relevant features.

2.5. Water footprint estimations

The water footprint of thermal power plants was calculated using the formula:

$$\text{Water footprint} = \text{WUI} \times \text{Active Capacity} \times \text{Full load hours} \quad (1)$$

where WUI is the water use intensity factor, in m^3/MWh of generated electricity, Active Capacity is in megawatts, and Full load hours of power generation are in hours. WUI factors were derived from Macknick et al. (2012), WUI factors were assigned to each individual power plant in the dataset using the information concerning their type of fuel, generation technology and cooling system (actual and predicted by the model, respectively). The full load hours of power generation were obtained from the EIA International energy outlook 2021 (U.S. Energy Information Administration and U.S. Department of Energy, 2021) for the year 2020. For simplicity of calculations, it was assumed that all power plants presented in our database were active in the year 2020.

3. Results

3.1. Global trends in cooling technology installations

3.1.1. Historical trends in cooling technology selection

The evolution of global cooling system installations in relation to yearly installed power generation capacities during the period from 1923 to 2015 is illustrated in Fig. 2. The figure shows that during essentially the entire time period cooling towers were the most common cooling technology globally. From 2000 onwards, there appears to be a clear preference for investments in power plants equipped with cooling towers, which is reflected by a rapid increase in the commissioning of such power plants. Once-through cooling was the second main cooling technology, which shows almost stable annual installations during the given timeframe (showing peaks in the mid-1980s, mid-2000s and 2010s). Apart from that, annual installations of dry-cooled thermal power plants remained well below those of cooling towers and once-through cooling but, as with tower cooling, started to increase rapidly in the 2000s. This increase in installations of tower and dry cooling systems indicates a development towards the reduction of water abstractions for cooling purposes in power generation since they have a lower water demand per unit of generated energy than the other cooling technologies. This global "trend" of a more sustainable use of water resources was most likely driven by a globally rising demand for water resources and climate change, which have already led to forced reductions in power generation in many parts of the world (Roehrkasten et al., 2015). Opposed to that, annual investments in power plants equipped with cooling ponds and inlet cooling systems were considerably lower than those in the aforementioned cooling technologies. Major installations of inlet-cooled plants were recorded after 2003 (Fig. 2A), driven by investments in the United States, Australia, and countries in the Middle East and North Africa (MENA), Africa and South America.

3.1.2. Fuel-specific power generation and power plant capacity

Currently, 50.13% of active thermal capacities are equipped with cooling towers, 37.41% have once-through cooling systems and 8.32% are dry-cooled. Inlet cooling systems and cooling ponds were also considered in the analysis. However, their current share from a global perspective is rather low: 2.50% and 1.65%, respectively. Even though the current share of inlet and pond cooling systems is rather low, according to Macknick et al. (2012), these cooling systems have high water consumption factors. In particular, power plants equipped with cooling ponds consume, on average, 2.5 times more water than power plants that use once-through cooling. Therefore, their impact on local ecosystems might be considerable and should not be neglected. Fig. 3 (left side) illustrates the current shares of installed cooling technologies for each type of fuel for power generation. Once-through cooling systems represent the most common cooling technology currently utilized in nuclear (59.69% of the total capacity) and oil (57.87% of the total capacity) power plants. Cooling towers are the major cooling technology for coal and gas power plants with 62.96% and 44.77% of their total capacity respectively. Dry cooling systems are mostly used in gas- and oil-fired power generation but have a share of less than 20% for these fuel types (15.40% of gas capacity and 11.03% of oil capacity). In contrast to that, cooling ponds are mostly found in nuclear power generation: about 6.71% of the global nuclear capacity is pond-cooled. The use of inlet cooling systems was detected only for gas and oil power plants with a comparably small share of less than 10% for each type.

The next step was to analyse the link between the size (capacity) of power plants and the installed cooling technologies. According to Fig. 3 (right side), for power plants from 50 MW to 2000 MW, cooling towers represent the largest group with an average share of 44.77%. With the single exception of the largest class of power plants between 5'000–8'000 MW, no connection between the power plant size and the number of plants with cooling towers can be detected. Opposed to that, higher power plant capacities correspond to a higher share of once-

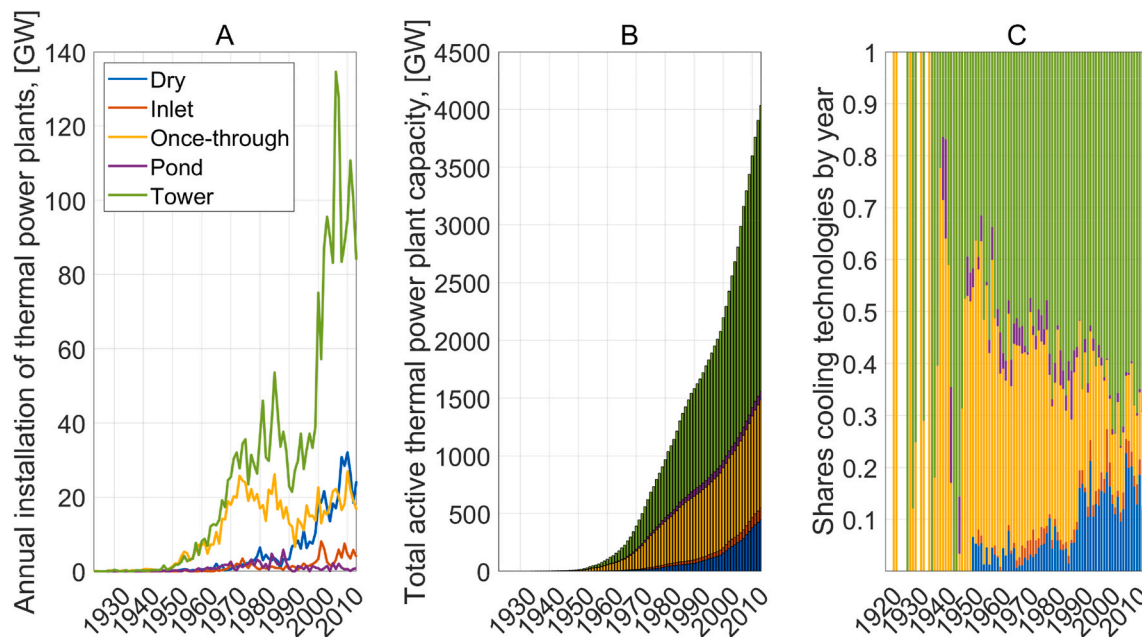


Fig. 2. Global trends in cooling technology installations, data presented for currently active power plants exceeding 50 MW: A – Annual installation of thermal power plants by type of cooling technology, in GW; B – Total active thermal power plant capacity by type of cooling technology, in GW; C: Shares of cooling technologies by year.

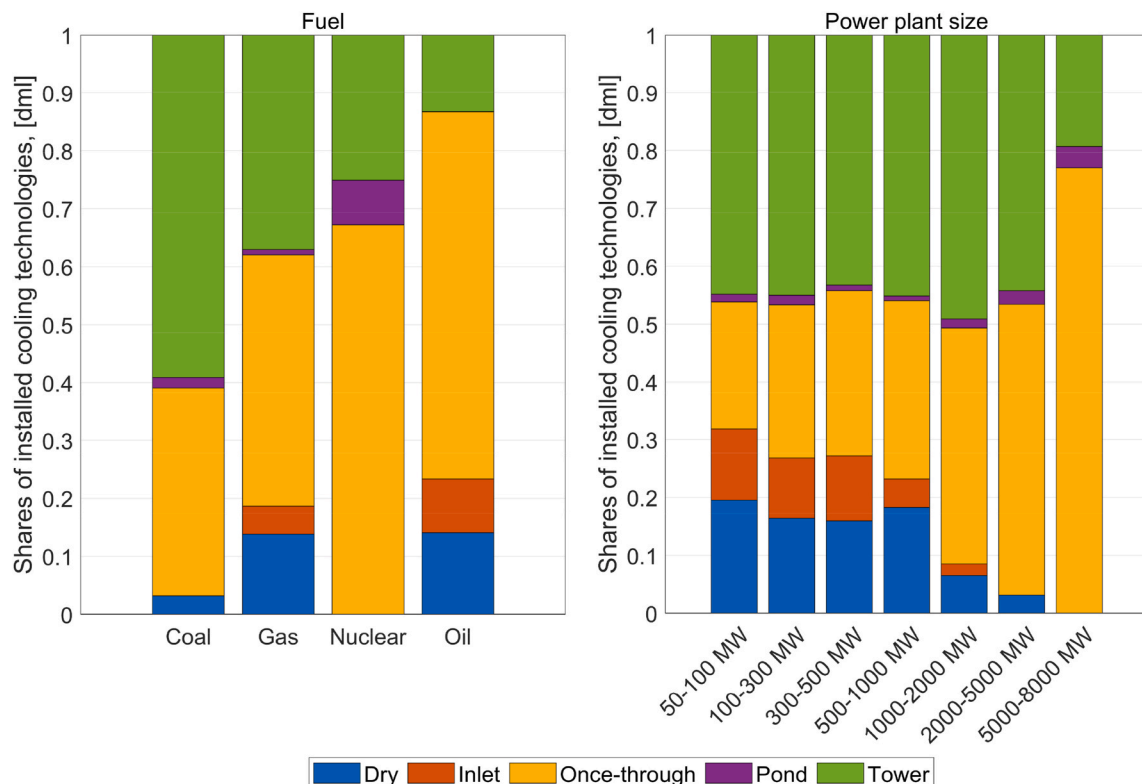


Fig. 3. Current shares of installed cooling technologies in relation to fuel-specific power generation (left) and size of individual power plants (right).

through cooling system installations. In particular, the share of once-through-cooled capacities is about 26.44% in the 50–100 MW size class, whereas for large-scale power generation (2'000–5'000 MW and 5'000–8'000 MW size classes) once-through becomes with 75.79% the major cooling technology due to its' low cost and increased heat discharge. Dry cooling is mainly utilized in power plants ranging from 50 to 1'000 MW: on average, 16.0% of the power plants of this size are

dry-cooled. However, due to its limited application, high material requirements and considerable costs, dry cooling is typically applied in cases of high thermal efficiency and low heat discharge. Therefore, it appears intuitive that the share of dry-cooled power plants in large-scale global power generation is negligible. Cooling ponds, in general, account for less than 1.90% of the global power capacity in different size classes. One exception is the 5'000–8'000 MW size class, where pond-

cooled power capacities have a share of 5.17%. These capacities exclusively represent nuclear power plants.

3.1.3. Geographical distribution of cooling technologies

Fig. 4 illustrates the geographical distribution of cooling technologies utilized by individual power plants.

The highest density of power plants equipped by cooling towers is recorded in regions with a high share of coal power capacities, for instance, in India and China. Once-through cooling systems are common in regions (1) with high potential of freshwater resources (availability of large rivers and lakes) or (2) with access to the ocean coastline. Examples for (1) are: Eastern Canada – the Great Lakes, Egypt – the Nile, Russian Urals and Siberia – the Ob and the Yenisei, Ukraine – the Dnieper, Austria and Hungary – the Danube, as well as Turkmenistan – the Caspian Sea. Examples for the latter, regions with ocean access (2), are: Japan, Indonesia, Malaysia, South Korea, and island countries. Unsurprisingly, the regions dominated by dry cooling correspond to regions showing a combination of three factors: prevalence of gas- and oil-fired power plants in the power generation system, hot climate, and low water resource availability. Therefore, there is a prevalence of the dry cooling technology in several countries in Africa, MENA, and some parts of Southern Europe (e.g., Italy). Inlet-cooled thermal power plants mainly occur in regions, which rely on gas- and oil-based power generation (South America, MENA and Australia), and cooling ponds are common in regions with high availability of water resources (e.g., Russia and some areas in the USA). Therefore, in addition to fuel type and the size of the power plant, such features as water availability, type of water used for cooling and climate conditions appear to be logically linked to the choice of the cooling technology.

Fig. 5 shows the connection between the current major cooling technologies in different countries and the average annual number of days with temperature higher than 20 °C in relation to the water resource availability.

The major cooling technology in the diagram corresponds to the cooling technology that has the largest overall capacity in the country. Water resource availability is represented by the water stress score, which was derived from the Water Risk Atlas of the [Water Resources Institute \(2021\)](#) for each country. The connection was separately studied for freshwater-cooling (Fig. 5, left side) and seawater-cooling (Fig. 5, right side). The source of water for cooling was assigned based on the location of individual power plants in respect to the nearest ocean's coastline. In particular, following the approach applied by [Biesheuvel et al. \(2016\)](#), all power plants located within 20 km of a coastline were assumed to be seawater-cooled.

As depicted in Fig. 5, different patterns for freshwater- and seawater-cooling technologies can be observed. In the case of freshwater-cooling (Fig. 5, left side), cooling towers cover almost half of the observed countries and, thus, represent the most common cooling technology currently utilized in thermal power plants. Moreover, they can be found in regions that experience any number of days with a temperature exceeding 20 °C. Once-through cooling is the prevalent technology for seawater-cooled power fleet, and its use seems not be clearly linked to the number of days with a temperature exceeding 20 °C or the water stress score. Unsurprisingly, dry cooling is widely implemented in hot regions and/or in regions with high water stress. Lastly, pond cooling is not presented as the main technology in any country.

3.2. Performance of the global model and region models

After the investigation of the general trends in the cooling technology installations worldwide, the next step was to create a global model as well as models for each region for the cooling technology identification and to analyse as well as contrast their performances.

The classification accuracy of the global model is 85.42% ($\pm 1.60\%$ standard deviation), which is very high for a five-class classification problem with relatively well-balanced cooling technology classes (the

calculated Gini coefficient¹ was 0.66). The stability² of 0.54 indicates that the best feature subsets selected by the algorithm are moderately stable.

In order to explore the performance of the model on different types of power plants, it was separately applied on four datasets, each containing only power plants of one fuel type: coal, gas, nuclear and oil. The decision to perform this step was based on the fact that power plants of the same fuel type have similar technical characteristics, which may potentially contribute to the classification performance of the model. The results of this investigation are presented in Table 1, including the actual and predicted consumption for each model (calculated using Equation (1)), the relative error of the prediction and the mean absolute error for the power plants. As demonstrated in the table, more than 90% of the cooling systems of coal and nuclear power plants were identified correctly. Since power plants of these fuel types generally use a considerably larger amount of water for cooling, mostly due to their large capacity and high full load hours compared to other power plants, it is desirable to be able to predict their cooling systems with a very high accuracy, which the presented model is capable to do. It is noteworthy that the mean absolute error (MAE) for nuclear power plants is about three to eight times larger than those of the other fuel types, which can be linked to the fact that these power plants are on average also multiple times larger (active capacity three to seven times larger). However, given the very high accuracy for nuclear power plants, the relative error is overall still comparably low (-2.45%).

The next step of the study was to train and cross-validate the region models. The average test set accuracy using the region models was 79.76% ($\pm 3.35\%$ standard deviation) (median 83.33%). When the test accuracy per region is weighted by the number of power plants in the region, the average test accuracy would amount to 83.47% ($\pm 1.40\%$ standard deviation), which is much closer to the accuracy of the global model – indicating that the region models perform almost as well as the global model, but each only requires information on the power plants in that region. Accounting for the number of power plants in each region for the comparison appears more suitable as following simple averaging e.g., the region US-MW (North America) with 681 plants and the region CAR (Africa) with 5 power plant would with simple averaging be given the same weight. The reason for the higher weighted average accuracy is that the average test accuracy is higher for regions with moderate to high number of samples (25 or more, $82.95\% \pm 1.57\%$)³ but comparably low for regions with few power plants (less than 25, $70.50\% \pm 8.62\%$). Similarly, it is apparent that the variation in the results, as measured by the standard deviation, tends to be lower the more power plants (observations) are available for the training of the model.

The performance of the models for each region is summarized in Fig. 6, which averaged the test set results after the feature selection and model training over the 5 folds. The figure helps to identify regions, for which the region models demonstrate high performance, and to detect regions, where the models have comparably lower test set accuracy.

The results presented in the figure indicate that for 98% of the regions (with the exception of the regions 'CAR', 'IOCE' and 'BR-N' – see Supplementary Table SM1) the average test set accuracy was higher than 50%. About 60% of regions demonstrated a test set accuracy of at least 80%, and about 25% of regions are associated with a test set accuracy of 90% and higher.

It may be expected that test accuracies are higher when there is a large imbalance (low Gini coefficient) in the data, meaning only a single

¹ For details concerning Gini coefficient used in this study see Supplementary Note 2: Inequality measure for class variable.

² For details concerning the stability measure used in this study see Supplementary Note 1: Feature subset stability.

³ For 25 to 99 power plants the average test accuracy is 81.51% ($\pm 2.31\%$), for 100 to 250 power plants it is 84.19% ($\pm 1.12\%$), and for over 250 power plants it is 85.81% ($\pm 1.30\%$).

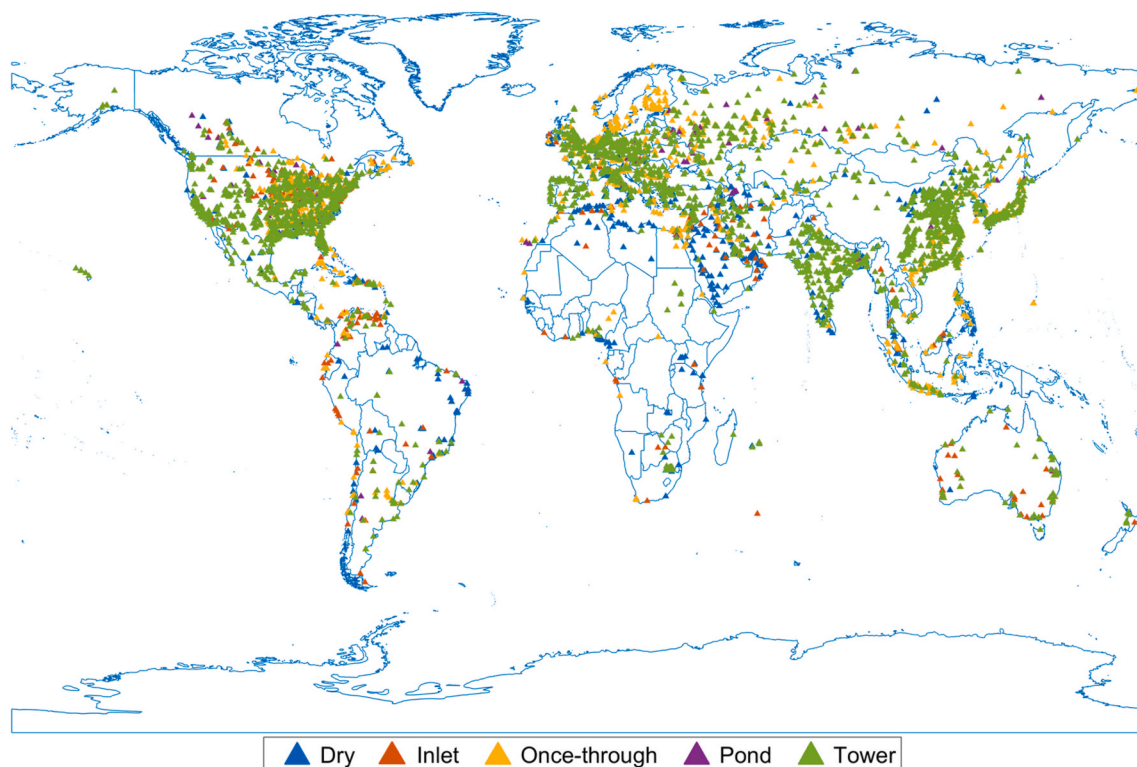


Fig. 4. Geographical distribution of cooling technologies utilized by individual power plants.

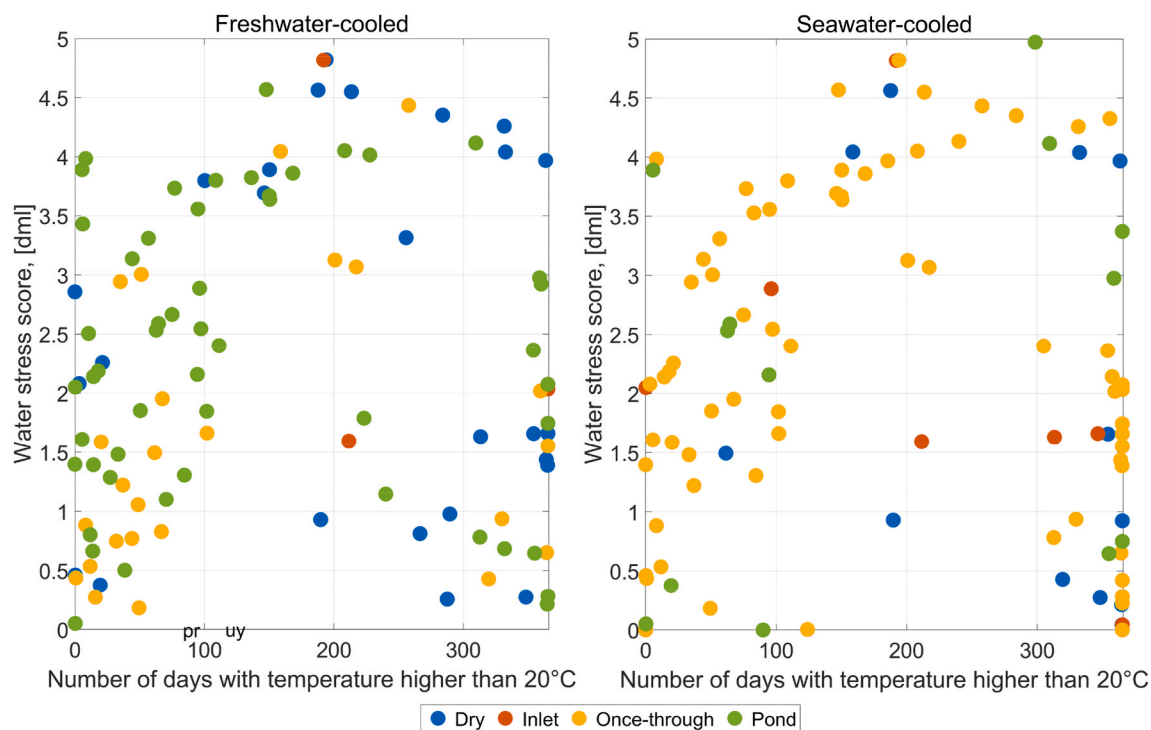


Fig. 5. Distribution of country-specific major cooling technology in relation to number of days with average temperature higher than 20 °C (Sadovskaia et al., 2019) and water stress score (Water Resources Institute, 2021).

cooling technology is prevailing. However, this only appears to be true to some degree, which is also illustrated by a low to moderate negative correlation (-0.4) between the test accuracy and the Gini coefficient. As depicted in Fig. 6, there are numerous regions (e.g., RU-SI, US-TVA, US-S, or AU-E) where a test accuracy of above 80% or 90% can be

accomplished which are also characterized by a low inequality (high Gini coefficient) of the cooling technology.

There are only five regions for which the model results in test accuracies of 50% or less. These regions have very few power plants (9.4 on average vs 100.5 on average for all regions), very low feature subset

Table 1
Comparison of global models.

Model	Performance (Accuracy $\pm$ Std)	Actual Consumption [km ³]	Predicted Consumption [km ³]	Mean Absolute Error [km ³]
Global model	85.42% $\pm$ 1.60%	16.656	16.360 (−1.78%)	0.00016
Coal power plants only	91.36% $\pm$ 1.82%	7.200	6.998 (−2.80%)	0.00015
Gas power plants only	81.34% $\pm$ 1.21%	4.608	4.625 (+0.37%)	0.00017
Oil power plants only	80.69% $\pm$ 1.79%	0.655	0.645 (−1.47%)	0.00007
Nuclear power plants only	93.57% $\pm$ 3.54%	4.194	4.091 (−2.45%)	0.00056

stability (average 0.10 vs average of all regions 0.43) and have selected exactly one feature each, which is the minimum to be able to conduct any model training. This illustrates that for these few regions the model was not capable to find any strong patterns in the data using any of the selected features. The fact that the number of power plants to train on was small and the average Gini coefficient of the cooling technologies was high (0.59) likely contributed to this outcome.

The next step is to analyse the results from the perspective of the major regions in which all regions displayed in Fig. 6 are situated. The regions located in Northeast Asia show, on average, the highest classification accuracies (the average accuracy is 89.46% ($\pm 3.57\%$)), whereas the regions of Africa are associated with the lowest average classification accuracy (the average accuracy is 66.33% ($\pm 26.32\%$)). This may be explained by the difference in samples available for model training and may be linked to the average feature subset sizes in these regions. The regions of Northeast Asia are characterized by, on average, 189.75

power plants per region, which is the highest value among all major regions, and a high number of features in the selected feature subsets. In contrast to that, the regions of Africa have, with on average 20.31 power plants per region, the smallest sample sizes, and also the smallest feature subsets among all regions considered in this analysis. The smaller sample size (power plants per region) may have negatively impacted the model training and the ability of the classifier to detect generalizable patterns in the data that link the features to the cooling technology. For these regions, the following strategies could be performed in order to obtain a higher sample size that may improve the classification accuracies: (a) redivision of regions or (b) merging some low sample regions. However, in this research, in order to be consistent with the previous water-energy nexus studies, it was decided to follow the previously applied geographical division of the world.

3.3. Performance comparison with the majority class model

An interesting comparison for the region models is that against a naïve classifier that assigns to all power plants the cooling technology that is most common in that region. The average test set accuracy of such a majority class classifier is 66.96% ($\pm 1.60\%$), which is highly significantly lower than the 79.76% ($\pm 3.35\%$) test accuracy accomplished with the region models (one-sided Welch's test, p-value < 0.1%). Fig. 7 compares the classification accuracy of the region models and the majority class models in terms of the Gini coefficient and the number of power plants.

First of all, the heatmaps show that the performance for regions where one cooling technology has a share of over 80% (Gini coefficient “Below 0.3”), the test accuracies of both approaches appear comparable, with the region models showing a slight outperformance for sample sizes of 25 and more, whereas the majority class models are slightly better for the smallest sample size of less than 25. However, it is apparent that with decreasing inequality (larger Gini coefficient) – meaning regions are having a more balanced use of different cooling technologies – the classification accuracy of the majority class models starts deteriorating considerably in comparison to the region models. This effect can be seen in the heatmap for the majority class models where the average test

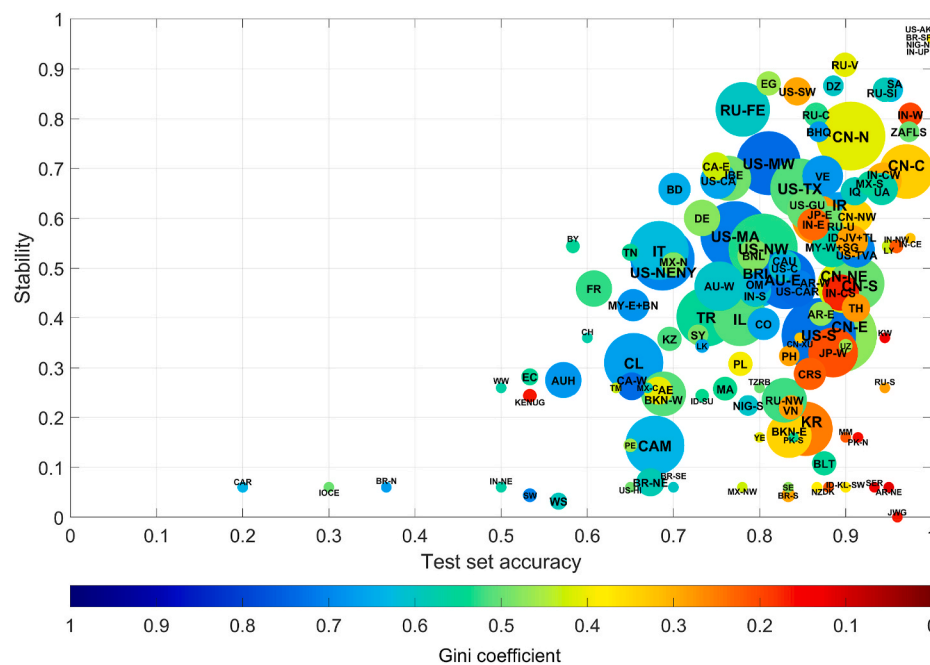


Fig. 6. Average test set accuracy in relation to stability of the “best feature subsets” and inequality among the class values (the Gini coefficient), per region. The size of the bubbles indicates the average size of the feature subset, with a minimum of 1 and maximum of 5 features per feature subset. Abbreviations used in the figure are provided in Supplementary Table SM1.

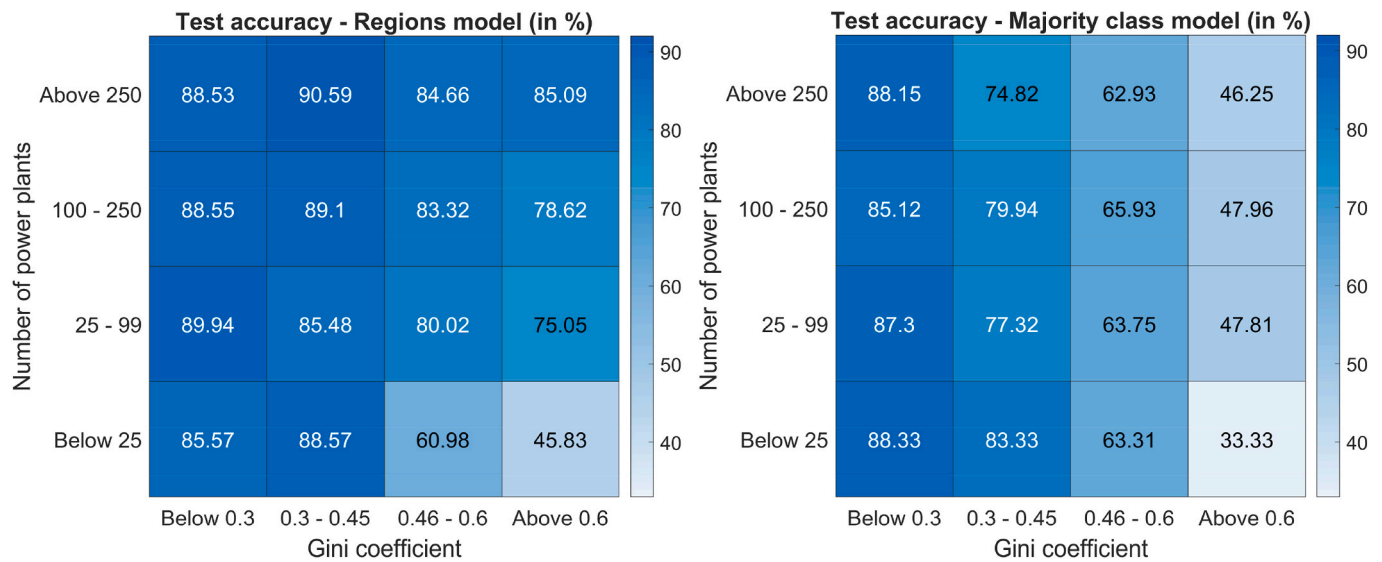


Fig. 7. Average test set accuracy for the region models and the majority class model by Gini coefficient and number of power plants.

accuracy deteriorates as the Gini coefficient increases (from left to right). It is noteworthy and intuitive that this development is not dependent on the number of power plants but only on the ratio of the majority class in the data as the key driver for the performance of a majority class model.

In particular, for the majority class models the linear correlation of the test accuracy with the Gini coefficient is strongly negative (-0.81) whereas the correlation between the test accuracy and the number of samples for these models is only weakly negative (-0.17). This indicates that the test set accuracy tends to decrease as the Gini coefficient increases, whereas there is no strong relationship between the number of power plants to train on and the classification accuracy accomplished. In contrast to that, the region models only experience a comparably small decrease in the average test set accuracy as the Gini coefficient increases. This highlights that the performance of the region models is clearly not as linked to the imbalance of the cooling technologies (correlation -0.4). However, it also reflects that the classification task's difficulty increased as the imbalance decreases (for instance due to more unique cooling technologies in those regions). Finally, for regions with more balanced cooling technologies (Gini coefficient ≥ 0.46) there is a noticeable decrease in the average test accuracy when the number of samples (power plants) decreases (see Fig. 7). This suggests that when the cooling technologies present in a region are more balanced, having a large number of samples is beneficial for model training. For regions with a strong focus on a single cooling technology (low Gini coefficient) this does not necessarily appear to be the case.

Overall, the comparison reveals that for 84.8% of the regions the proposed region models demonstrate a higher accuracy than the majority class models (see Supplementary Table SM1 for more information). The average improvement of the classification accuracy amounted to 17.2%-points. For the remaining 15.2% of the regions, the majority class model outperformed the proposed classification model. It is noteworthy that the majority class performs on these selected regions with an average test accuracy of 78.57% considerably better than it does on all regions (66.96%). In contrast to that, the region models perform here noticeably worse with an average test accuracy of only 66.78%, which is likely affected by the low number of samples for these regions. In particular, the average sample size is 100.5 for all regions in the data, whereas it is only 33.5 for these regions and even only 19.6 on average for regions for which the majority class model outperforms the region models by more than 5%-points. In addition, the average stability of the feature subsets of the region models for these regions is considerably lower (0.17) than the average over all regions (0.43) indicating the

challenge for feature selection to find a robust set of features.

Thus, it can be concluded that with exception of regions with very few samples and only a single prevailing cooling technology – a setup for which the majority class models perform well – the region models tend to outperform the majority class models considerably.

3.4. Cooling technology classification for different plant sizes

For researchers and practitioners, the identification of cooling technologies is typically a step that is needed to subsequently estimate the water demand for power generation. It is advisable to ensure a high classification performance for all power plant sizes, especially those groups that represent (1) the largest share of the overall capacity and (2) large power plants since their operation is associated with larger water demand, compared to small and medium-sized plants (on a per-power plant basis). In this regard, the following questions arise: How accurate is the process of the cooling technology identification for different power plant sizes? Will the model perform better for larger power plants than for smaller units? Fig. 8 demonstrates the results of this investigation.

As shown in Fig. 8, the global model's mean test accuracy (in relation to power plant sizes) varies between 81.53% (for power plants smaller

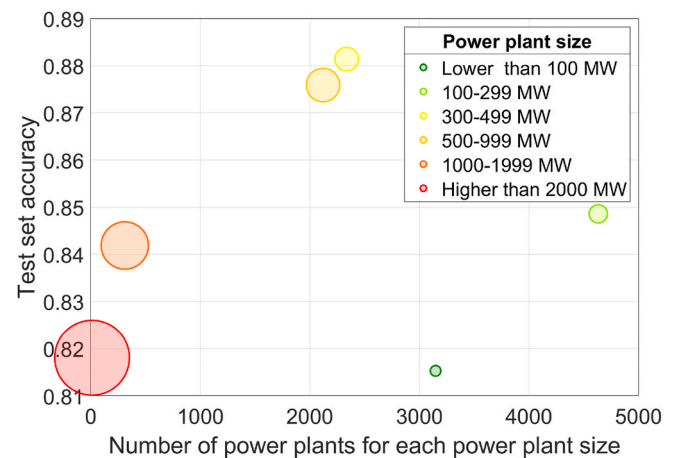


Fig. 8. Cooling technology classification accuracy (mean values) in relation to power plant size. The size and the colour of the bubbles indicate the size group of power plants.

than 100 MW of capacity) and 88.14% (for the power plant size of 300–499 MW). The power plants of 100–299 MW of capacity represent the largest group in the dataset used in this study (38.48% of the global thermal power capacity). For these power plants, the model demonstrates a mean test accuracy of 84.86%. According to the results of the analysis, the power plants exceeding 2000 MW of capacity are associated with an accuracy of 81.81%. This value is, in general, lower than the corresponding values of the test set accuracies for other power plant sizes (except for the power plant size lower than 100 MW) but is comparable with the average test accuracy of the region models. In this regard, it is crucial to mention that these power plants represent the plant size group with the lowest overall capacity accounting for just 0.73% of the global active power plant capacity. Hence, even though each power plant of this size may have a larger impact on the water demand than a power plant of smaller capacity, given the comparably small number of such power plants, their impact on the overall water demand associated with thermal power generation is negligible.

3.5. Cooling technology classification in water stressed regions

Ensuring a high accuracy of water demand estimates is especially crucial for regions associated with high and extremely high baseline water stress. In these regions, inaccuracies in water demand estimations (especially in cases when the demand values are underestimated) may affect the water management policies of the region and potentially lead to an overexploitation of the region's freshwater resources. Hence, the next step of this study was to investigate the link between the obtained classification accuracies and the reported water stress scores per region (Water Resources Institute, 2021) in order to identify potential bottlenecks of the proposed model for cooling technology classification. Fig. 9 depicts the distribution of regions based on their average regional water stress scores and the classification accuracy of the proposed model. The average regional water stress scores were calculated as the mean of the provincial water stress scores obtained from (Water Resources Institute, 2021) associated with the location of each specific power plant in the database.

A total of 46.2% of the global thermal power plant capacity is located in low and low-to-medium water stressed regions, and about 28.9% of

the capacities are in regions with medium-to-high water stress. However, the highest emphasis should be put on the most water stressed regions. The remaining 24.94% of the total global power generation capacity (corresponding to about 27.56% of the global power plant fleet) is located in regions characterized by high and extremely high water stress (with an average water stress score exceeding 3 obtained from (Water Resources Institute, 2021)). Moreover, these critical regions account for 32.0% of all regions considered. The classification approach for these regions demonstrated an average test set accuracy of 80.83% ($\pm 3.23\%$), which is even slightly higher than the average accuracy of 79.76% ($\pm 3.35\%$) of the region models overall. The test accuracy of the region models for these water-stressed regions is highly significantly higher than the test accuracy of the majority class model (one-sided Welch's test, p -value $< 0.01\%$). Hence, the proposed algorithm appears suitable to be applied for cooling system identification in water-stressed regions.

As shown in the figure, the results for the region 'IOCE', which combines power generation units of Reunion, Mauritius and Madagascar, demonstrated considerably lower classification accuracy (30%) than those of any other water-stressed regions. Hence, this region should be treated with caution when considering using a classification approach. However, it is crucial to mention that according to Lohrmann et al. (2019), 100% of power plants located in the 'IOCE' region are assumed to use seawater for cooling purposes. Therefore, a potential misclassification of cooling technologies in this region will likely have a minor effect on the freshwater usage estimation, especially compared to regions with a high share of freshwater-cooled power plants.

3.6. Identification of relevant features (variables)

For the global model, the analysis of the most commonly selected features for the cooling technology identification were "Power plant capacity (total active)" and "Seasonal variability"- these features were in each fold (during cross-validation) selected by the algorithm. It is crucial to mention that the variable "Seasonal variability" was not included in the original database, which highlights the importance of considering "external" features for the cooling systems identification. The feature "Electricity price" was selected in 80% of cases, and the features "Fuel

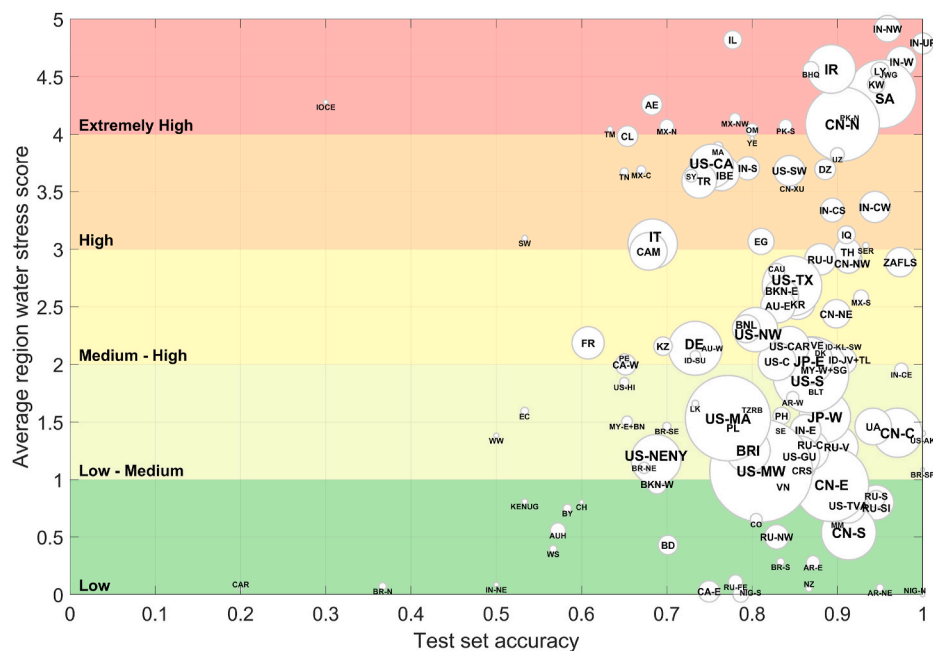


Fig. 9. Average test set accuracy in relation to average water stress coefficient (Water Resources Institute (WRI), 2021), per region. The size of the bubbles indicates the size of the region (number of power plants per region), with a minimum of 5 and a maximum of 681 power plants per region. Abbreviations used in the figure are provided in Supplementary Table SM1.

type” and “Seawater-cooling” were used in 40% of the selected feature subsets. The above-mentioned features were identified as the most important ones and, thus, are suggested to be selected in future studies for the cooling technology identification on the global level with, for instance, a KNN classifier.

For the region models, as can be expected, the relevant features selected differ to some degree between regions but show that certain features are frequently included in the models. Thus, each of the features that was the most relevant feature (most frequently selected in that region) for at least ten regions is analysed in more detail. Furthermore, the results for these six features (“Power plant capacity (total active)”, “Freshwater per capita”, “Water stress (province and country)”, “Fuel used in power plant”, “GDP per capita”, “Seawater-cooling”) are depicted for each specific region in Fig. 10. The figure displays the frequency with that these features were part of the selected feature subsets in each of the regions and, thus, illustrates their geographical distribution.

It is crucial to mention that only three out of these six often selected features were provided in the initial database (“Power plant capacity (total active)”, “Fuel used in power plant”, “Seawater-cooling”), whereas the other three (“Freshwater per capita”, “Water stress (province and country)” and “GDP per capita”) were included in this study as additional information, which was obtained from other open sources. Hence, collecting additional information related to the location of each specific power plant (such as the regional water stress score or the economic situation) might be crucial for the cooling system identification when the type of cooling system is unknown.

As shown in Fig. 10a, the feature “Power plant capacity (total active)”, which refers to the size of power plants, appeared most frequently in the selected feature subsets for the cooling system classification. In particular, this feature was included for 104 regions (or 83.2% of the total number of regions) in the selected feature subsets, and for 79 regions (or

63.2%), it was the most frequently selected feature in all the region’s feature subsets.

The second most important feature was “Water stress (province and country)”, which appeared as a relevant feature for 75 regions. This result implies that although this feature was not considered in previous studies, it might increase the accuracy of the cooling technology identification, especially on the individual power plant-level using a region model.

Unsurprisingly, the variable “Seawater-cooling” was selected only for regions, which have access to the sea coastline (Fig. 10f). However, it was not selected for all regions with coastline access. In particular, it was only selected for regions, where (1) seawater is used for the cooling of power plants (as estimated by Lohrmann et al. (2019)), and (2) not all power plants use seawater (in other words, some power plants which are located in the region withdraw freshwater for cooling). Similarly, the feature “Fuel used in power plant” was selected in regions, where thermal power plants of the same fuel type tend to have different cooling technologies.

The results also revealed that the features “Days of warm weather”, “Agricultural water withdrawal as percent of total renewable water resources of the country” and “Corruption Perceptions Index” were not included in any of the 125 regions (with at least 5 power plants) most frequently selected features and, on top of that, were rarely even included in any of the subsets once (maximum twice) during the cross-validation. That implies that, although these variables might be crucial for the selection of the cooling system design at the stage of the power plant construction, they are potentially irrelevant for the cooling technology classification task at the region-level. This could be explained by the fact that numerical values taken by these features do not vary considerably within specific regions.

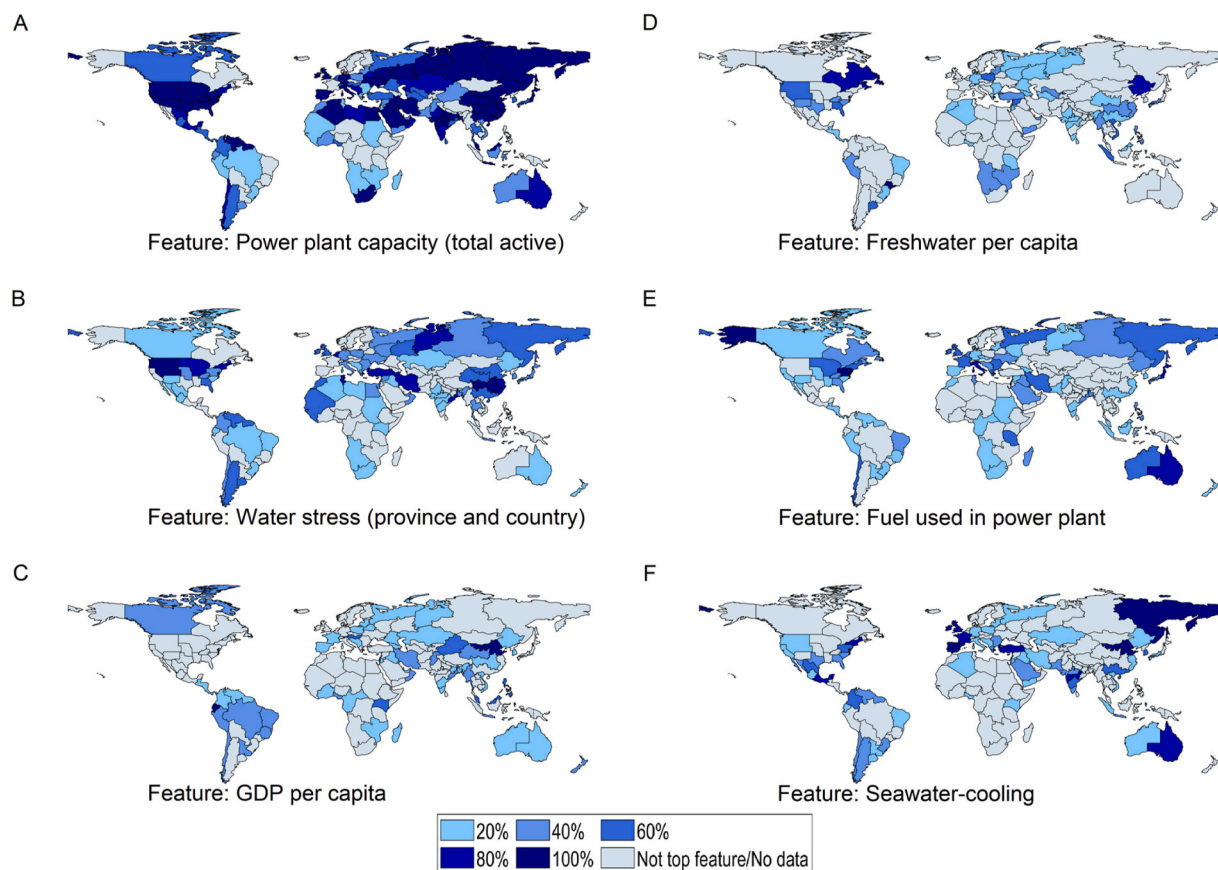


Fig. 10. Features that are most frequently selected during feature selection, per region. The colour indicates the share of CV folds in which a specific feature is selected for the feature subset in a region.

4. Discussion

Understanding the actual water demand of thermal power plants is the first step towards a sustainable use and allocation of water resources. This helps to raise the awareness of the amount of water used by the global thermal power fleet for cooling purposes and may potentially contribute to the discussion of the energy transition to a more sustainable and less water-dependent and water-demanding water resources, such as, for instance, solar and wind.

The models developed, presented and tested in this study represent the first step in the process of the water footprint assessment of thermal power plants. The study provides a theoretical basis for subsequent water demand estimations for thermal power plants by presenting the models and the corresponding “best feature subsets”, which could be used for the cooling system identification by practitioners in other studies.

The results of the study show that the models based on machine learning techniques are capable to deliver stable results and demonstrate a high performance. Overall, both the global and region models outperformed the commonly used majority class model for cooling technology identification and demonstrated a highly significant difference in their test accuracies to that benchmark. Apart from that the proposed machine learning-based method is not only considerably higher than the accuracy of the majority class model but also than other approaches used in water-energy nexus studies (85.42% versus about 66.96%, and 35% for other approaches on this five-class problem (Lohrmann et al., 2021b), respectively), which constitutes a further significant improvement in the identification and assignment of cooling technologies in water-energy nexus studies. As this point, the following question arises: How are these results related to the accuracy of the actual water footprint estimations? In order to answer this question, we calculated the water footprint of thermal power plants presented in the study for two cases: using the actual cooling technologies reported in the database, and cooling technologies, which were predicted by the model.

According to the results of this calculation, about 16.65 km³ of water was estimated to be consumed globally by the power plants, when the actual cooling technologies were considered. In contrast, the water footprint using the cooling technologies predicted by the model is about 16.35 km³ of water. This implies that although the accuracy of the cooling technology identification model is about 85%, it leads to only a slight underestimation of the water footprint of about 1.78%. This could be explained by the fact that coal and nuclear plants, which have a higher water footprint compared to other power plant types (due to their large size and high full load hours), also have a considerably higher accuracy of the cooling technology identification.

5. Conclusions

Identification of cooling technologies of thermal power plants is a fundamental step in the assessment of the water footprint of thermoelectric generation on both, global- and local-level. In this study, we introduce a new machine learning approach based on Differential Evolution Feature Selection and the K-Nearest Neighbor classifier for the cooling technology identification on a global and regional level.

The results demonstrate that the weighted average test set accuracy of region models (83.47% $\pm$ 1.40%) overall was comparable to that of the global model (85.42% $\pm$ 1.60%) while providing more specific results for each of the regions. Both a global model and region models offer practitioners a choice between models, depending on data availability (regional vs. global power plant data, number of observations (power plants)) and the desired perspective of their investigation (regional or global). Apart from that, a special emphasis in the study was placed on regions characterized by high and extremely high water stress, where errors in water policy planning and water management may lead to an unsustainable water use or even to an overexploitation of water resources.

The results of the study also revealed that the optimal set of features for the cooling technology identification varies in different regions around the world, which highlights the potential existence of region-specific patterns for the cooling technology assignment. According to the models, the most common features for the cooling technology identification were the total active capacity of the power plant (typically available in the power plant databases) and the water stress score (assigned based on the exact location of each specific power plant). In future research, the features identified as being relevant as well as machine learning models such as the KNN classifier used in this study, can be widely applied for the cooling technology identification on both, the global and regional level. The model can be widely used for the cooling technology identification for both, currently active and future power plants, when the size and location of the power generation unit is known, which is crucial for future energy system planning.

Since the actual size of specific power plants is identified as the most important feature for the cooling system classification for most of the regions worldwide, any classifier (including the KNN classifier applied in this study) cannot be applied in the same way for the missing value imputation of the cooling technology for so-called “aggregated capacities”. The term “aggregated capacities” usually refers to separate entries in power plant databases, which typically combine several power plants of unknown size (in this study, these capacities were removed from the initial power plant database prior to the feature selection to avoid their impact on the results of the feature selection). In addition, this has limited applicability for some energy transition scenarios, which provide values of “aggregated capacities” for power plants of different fuel types, but not of different plant sizes. Lohrmann et al. (2021b) reviewed existing methods for the cooling technology missing value imputation for these cases and suggested that the existing methods yield an average test set accuracy of about 35% (for a five-class problem). Hence, in terms of accuracy, the method proposed in this study outperforms the previous methods considerably.

Code availability

Matlab scripts used in the production of this analysis as well as the important features datasets, which were obtained for each specific region, are available from the corresponding author upon request.

CRediT authorship contribution statement

Alena Lohrmann: Conceptualization, Resources, Methodology, Software, Investigation, Visualization, Validation, Writing – original draft. **Christoph Lohrmann:** Formal analysis, Methodology, Software, Validation, Writing – original draft. **Pasi Luukka:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank Dr. J. Farfan for helpful discussions concerning later versions of the manuscript and proofreading of the article.

This work was supported by the Kone Foundation and the Finnish Academy of Science and Letters.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclepro.2022.134842>.

References

- AQUASTAT, 2020. AQUASTAT - FAO's Information System on Water and Agriculture. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. <http://www.fao.org/aquastat/en/overview/methodology/water-use>. accessed 10.5.21.
- Biesheuvel, A., Witteveen+Bos, Cheng, I., Liu, X., Greenpeace International, 2016. Methods and Results Report: Modelling Global Water Demand for Coal Based Power Generation. Witteveen + Boss & Greenpeace International. <https://wayback.archive-it.org/9650/20200430014936/http://p3-raw.greenpeace.org/international/Globa/International/briefings/climate/2016/Coal-Water/Technical-Report-Coal-Water-Methods-Results.pdf>.
- Bolón-Canedo, V., Sánchez-Maróño, N., Alonso-Betanzos, A., 2012. An ensemble of filters and classifiers for microarray data classification. *Pattern Recogn.* 45, 531–539. <https://doi.org/10.1016/j.patcog.2011.06.006>.
- Boretta, A., Rosa, L., 2019. Reassessing the projections of the world water development report. *npj Clean Water* 2, 15. <https://doi.org/10.1038/s41545-019-0039-9>.
- California Energy Commission, Electric Power Research Institute (EPRI), 2002. Comparison of Alternate Cooling Technologies for California Power Plants. Economic, Environmental and Other Tradeoffs. Palo Alto, California, USA. <https://www3.epa.gov/region1/npdes/merrimackstation/pdfs/ar/AR-1167.pdf>. accessed 12.18.21.
- Diehl, T.H., Harris, M.A., Murphy, J.C., Hutson, S.S., Ladd, D.E., 2013. Methods for Estimating Water Consumption for Thermoelectric Power Plants in the United States. <https://doi.org/10.3133/sir20135188>.
- EcoFys, B.V., 2014. Pilot Project on Availability, Use and Sustainability of Water Production of Nuclear and Fossil Energy. Geo-localised inventory of water use in cooling processes, assessment of vulnerability and of water use management measures. By the order of: the European Commission Directorate General Environment. <https://ec.europa.eu/environment/archives/water/adaptation/pdf/InventoryCoolingWaterUse.pdf>. accessed 3.18.20.
- Enerdata, 2020. Power Plant Tracker. Grenoble, France. <https://www.enerdata.net/research/power-plant-database.html>. accessed 12.18.21.
- European Network of Transmission System Operators for Electricity (ENTSO-E), 2022. Transparency Platform. Brussels, Belgium. <https://transparency.entsoe.eu/>. accessed 12.18.21.
- Flörke, M., Schneider, C., McDonald, R.I., 2018. Water competition between cities and agriculture driven by climate change and urban growth. *Nat. Sustain.* 1, 51–58. <https://doi.org/10.1038/s41893-017-0006-8>.
- Food and Agriculture Organization of the United Nations (FAO), 2021. AQUASTAT Main Database. Rome, Italy. <http://www.fao.org/aquastat/statistics/query/index.html;jsessionid=6C067D308754D4DA7FEF0D86786A4BAB>. accessed 10.5.21.
- Food and Agriculture Organization of the United Nations (FAO), 2020. The State of Food and Agriculture 2020. Overcoming Water Challenges in Agriculture. Rome, Italy. <https://www.fao.org/3/cb1447en/cb1447en.pdf>. accessed 3.18.20.
- Global Energy Observatory (GEO), 2018. Power Plants Database. New Mexico, USA. <http://globalenergyobservatory.org/>. accessed 12.18.21.
- GlobalData Plc, 2020. GlobalData Power Plant Database. London, UK. <https://www.globaldata.com/>. accessed 5.10.21.
- GlobalData Plc, 2014. GlobalData Power Plant Database. London, UK. <https://www.globaldata.com/>. accessed 5.2.16.
- GlobalPetrolPrices, 2020. Electricity Prices. https://www.globalpetrolprices.com/electricity_prices/. accessed 1.20.21.
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. <https://doi.org/10.1007/b94608>.
- Khushaba, R.N., Al-Ani, A., Al-Jumaily, A., 2011. Feature subset selection using differential evolution and a statistical repair mechanism. *Expert Syst. Appl.* 38, 11515–11526. <https://doi.org/10.1016/j.eswa.2011.03.028>.
- Larsen, M.A.D., Drews, M., 2019. Water use in electricity generation for water-energy nexus analyses: the European case. *Sci. Total Environ.* 651, 2044–2058. <https://doi.org/10.1016/j.scitotenv.2018.10.045>.
- Larsen, M.A.D., Petrovic, S., Engström, R.E., Drews, M., Liersch, S., Karlsson, K.B., Howells, M., 2019. Challenges of data availability: analysing the water-energy nexus in electricity generation. *Energy Strategy Rev.* 26, 100426 <https://doi.org/10.1016/j.esr.2019.100426>.
- Liu, H., Yu, L., 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.* 17, 491–502. <https://doi.org/10.1109/TKDE.2005.66>.
- Lohrmann, A., Child, M., Breyer, C., 2021a. Assessment of the water footprint for the European power sector during the transition towards a 100% renewable energy system. *Energy* 233, 121098. <https://doi.org/10.1016/j.energy.2021.121098>.
- Lohrmann, A., Farfan, J., Caldera, U., Lohrmann, C., Breyer, C., 2019. Global scenarios for significant water use reduction in thermal power plants based on cooling water demand estimation using satellite imagery. *Nat. Energy* 4, 1040–1048. <https://doi.org/10.1038/s41560-019-0501-4>.
- Lohrmann, A., Lohrmann, C., Luukka, P., 2021b. A region-based approach for missing value imputation of cooling technologies for the global thermal power plant fleet using a decision tree classifier. In: Luukka, P., Stoklasa, J. (Eds.), *Intelligent Systems and Applications in Business and Finance*. Springer Nature Switzerland AG, Cham, Switzerland.
- Luo, T., Krishnaswami, A., Li, X., 2018. A Methodology to Estimate Water Demand for Thermal Power Plants in Data-Scarce Regions Using Satellite Images. Technical note, Washington, DC.
- Macknick, J., Newmark, R., Heath, G., Hallett, K.C., 2012. Operational water consumption and withdrawal factors for electricity generating technologies: a review of existing literature. *Environ. Res. Lett.* 7 <https://doi.org/10.1088/1748-9326/7/4/045802>.
- Maulbetsch, J., Stallings, J., 2012. Evaluating the economics of alternative cooling technologies. *Power Eng.* 116, 120–128.
- Maupin, M.A., Kenny, J.F., Hutson, S.S., Lovelace, J.K., Barber, N.L., Linsey, K.S., 2014. Estimated use of water in the United States in 2010, U.S. Geol. Surv. Circular 1405. <https://doi.org/10.3133/cir1405>.
- Roehrkasten, S., Schaeuble, D., Helgenberger, S., 2015. Secure and Sustainable Power Generation in a Water-Constrained World. Institute for Advanced Sustainability Studies (IASS), Potsdam.
- S&P Global, 2016. Platts World Electric Power Plants Database. Washington DC, USA. <http://www.platts.com/products/world-electric-power-plants-database>. accessed 5.10.21.
- Sadovskaia, K., Bogdanov, D., Honkapuro, S., Breyer, C., 2019. Power transmission and distribution losses – a model based on available empirical data and future trends for all countries globally. *Int. J. Electr. Power Energy Syst.* 107, 98–109. <https://doi.org/10.1016/j.ijepes.2018.11.012>.
- Spang, E.S., Moomaw, W.R., Gallagher, K.S., Kirshen, P.H., Marks, D.H., 2014. The water consumption of energy production: an international comparison. *Environ. Res. Lett.* 9 <https://doi.org/10.1088/1748-9326/9/10/105002>.
- Statista. The Statistic Portal, 2021. Annual Gross Domestic Product and Real GDP in the United States from 1930 to 2020 (In Billion U.S. Dollars). Hamburg, Germany. <https://www.statista.com/statistics/1031678/gdp-and-real-gdp-united-states-1930-2019/>. accessed 10.5.21.
- The World Bank, 2020. GDP (Current US\$) - All Countries and Economies. Washington DC, USA. <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD>. accessed 10.5.20.
- Transparency International, 2019. Corruption Perceptions Index 2019. Berlin, Germany. <https://www.transparency.org/en/publications/corruption-perceptions-index-2019>. accessed 10.5.21.
- U.S. Energy Information Administration (EIA), 2021. Electricity Explained. Factors Affecting Electricity Prices. Washington DC, USA. <https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php>. accessed 10.5.21.
- U.S. Energy Information Administration (EIA), U.S. Department of Energy, 2021. International Energy Outlook 2021. Washington, DC, USA. https://www.eia.gov/outlooks/ieo/pdf/IEO2021_Narrative.pdf. accessed 8.20.22.
- United Nations (UN), 2022. The Paris Agreement: Status of Treaties. New York, USA. https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtsg_no=XXV11-7-d&chapter=27&clang=en. accessed 9.9.22.
- United Nations (UN), 2018. World Urbanization Prospects: the 2018 Revision - Percentage of Population at Mid-year Residing in Urban Areas by Region, Subregion and Country, 1950-2050. United Nations Department of Economic and Social Affairs, Population Division, Washington DC, USA. <https://population.un.org/wup/Download>. accessed 3.18.20.
- United Nations World Water Assessment Programme (WWAP), 2014. The United Nations World Water Development Report 2014: Water and Energy. United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris, France. <https://unesdoc.unesco.org/ark:/48223/pf0000225741>. accessed 3.18.20.
- United Nations World Water Assessment Programme (WWAP), UN-Water, 2018. The United Nations World Water Development Report 2018: Nature-Based Solutions for Water. United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris, France. www.unwater.org/publications/world-water-development-report-2018. accessed 3.18.20.
- Vassolo, S., Döll, P., 2005. Global-scale gridded estimates of thermoelectric power and manufacturing water use. *Water Resour. Res.* 41, 1–11. <https://doi.org/10.1029/2004WR003360>.
- Water Resources Institute (WRI), 2021. AQUEDUCT Water Risk Atlas. <https://www.wri.org/applications/aqueduct/water-risk-atlas>. accessed 3.18.20.
- World Economic Forum (WEF), 2021. The Global Risks Report 2021, sixteenth ed. http://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2021.pdf. accessed 9.09.22).
- Zhao, P., Zhang, M., 2018. The impact of urbanisation on energy consumption: a 30-year review in China. *Urban Clim.* 24, 940–953. <https://doi.org/10.1016/j.uclim.2017.11.005>.