Monthly hydropower prediction
at plant scale in data-scarce regions

Master thesis for the Industrial Ecology program
Leiden University
Delft University of Technology

Nils Boonstra
August 22, 2023

Monthly hydropower prediction at plant scale in data-scarce regions

Master thesis for the Industrial Ecology program
Leiden University
Delft University of Technology

Supervisors:
dr. V. Barbarossa
dr. S.J. Pfenninger
dr. S. Galelli

Cover Image: Grand Coulee Dam, created using http://maps.stamen.com and GIMP by the author.

Acknowledgments

I would like to thank my first supervisor, Valerio Barbarossa, for his guidance and support throughout the research and writing of this thesis. It would not have been possible to conclude this sometimes challenging project without his feedback during our regular meetings. He was able to guide the project when it needed direction.
Next, I would like to thank Stefano Galelli as external supervisor for his feedback on modeling and the field of hydropower, without which I would not have been able to finish this project. Our regular meetings were great for solving minor issues, to more general questions on what parts should be included and what not.
I am also thankful to my second supervisor Stefan Pfenninger for his feedback, keeping an eye on the broader picture and the project's usefulness. This made me step back at crucial moments to see if the project was still heading in the right direction.
I appreciate my family and friends' encouragement, patience, and understanding during the ups and downs of this journey.

In conclusion, this thesis would not have been possible without the abovementioned persons support, guidance, and encouragement. Thank you all for being a part of this project.

Nils Boonstra

Summary

Hydropower is currently the largest renewable energy generation method worldwide, being the third overall after coal and natural gas, and providing 15% of global electricity. However, generation data regarding hydropower is scarce. If available, it mostly exists at national and annual level. Only limited generation data is available at plant scale. For research on power grid decarbonization, electricity grid expansion development and electricity grid optimization, more data on hydropower generation is wanted at plant scale. Hydropower generation differs significantly throughout the year following weather patterns, which means monthly generation data would be beneficial for research. To fill in data gaps, two models were created for the prediction of hydropower, using plant capacity, discharge, and reservoir area as predictor variables for the monthly model. For the yearly model, reservoir area was not included in the final model. A linear mixed-effects regression model and a mixed-effects random forest model were fitted and compared to the Hydro Plant Generation Estimation Model. The models were created using data from the United States (US) and used for predictions with hydropower plants from the US and the European Union (EU). The median KGE for the monthly LMER model was -0.08 in the US. For the monthly MERF model, the median KGE was 0.12 in the US. In the EU the models were evaluated at an annual time step due to data limitations, resulting in the LMER model scoring better (-0.16) than the MERF model (-0.68) on median KGE. The prediction errors of the annual US model were comparable to the Hydro Plant Generation Estimation Model. Discharge and plant capacity were found to be important predictor variables, followed by reservoir area for the monthly model. The models were able to predict at plant scale in data-scarce regions and at a monthly time step, although they can produce large outliers. A purpose for the model could be to not use it at plant scale but at a larger scale, as the median KGE scores were around zero, showing that predictions over multiple HPPs are usable.

Keywords: Hydropower prediction, energy data scarcity, mixed-effects model

Table of Contents

1. Introduction

Hydropower is currently the largest renewable energy generation method worldwide, contributing 15% of the global electricity supply in 2022 (Ember, 2023). Many countries have plans to expand their hydropower capacity, which gives hydropower an important role in the energy transition to a low-carbon energy system (IEA, 2021). Hydropower functions as a relative flexible energy generation method, which becomes increasingly important as more intermittent renewable energy is added to electricity grids worldwide (Stoll et al., 2017). For many regions of the world, there is little publicly available data on hydropower generation (Larsen et al., 2019). If data is available on hydropower generation, it is at national and annual levels (Zhou et al., 2020). Data available at the power plant level are often found to be annually, with many years missing (Yin et al., 2020).

Hydropower electricity generation data at a sub-annual time step is wanted for power grid planning and analysis, as hydropower generation fluctuates throughout the year and is highly correlated with weather patterns (Turner & Voisin, 2022). More data on electricity generated from hydropower at sub-national level is important for research on the decarbonization of the power grid and grid expansion planning (Liu et al., 2019). Not knowing the amount of hydropower generated can lead to uneconomic electricity grid usage and development (Wei et al., 2023).

North America is seen as a relatively data-rich region regarding hydropower generation data, while Southern Africa and Central Asia are cited as regions with little data (Le et al., 2022). This is also reflected in data from the World Resource Institute, which tracks global data availability of electricity generation at the power plant level (see Table A1) (Yin et al., 2020).

To fill in existing data gaps, predictive models can be used. For hydropower, these predictive models can be divided into two main categories: short-term models, which can be used for managing power plants in the power grid, and long-term models, which can be used for electricity grid expansion planning (Chowdhurry et al., 2019). Hydropower models are underrepresented in the broader literature on renewable energy modeling, with most studies being on the topic of solar or wind, and 2% of the studies in this field being on the topic of hydropower (Lai et al., 2020). Of the examined studies using machine learning to predict renewable energy generation by Krechowicz, Krechowicz & Poczeta (2022), about 5% were on the topic of hydropower (Krechowicz, Krechowicz & Poczeta, 2022).

Existing hydropower models can be divided into four main types, distinguished by Turner & Voisin (2022). In the next section these four model types will be introduced, illustrated with hydropower models operating in a data-scarce region. The first type is a *Surface water to hydropower* model. This type of model uses a land surface model to convert climate data (precipitation, temperature or wind speeds for example) into runoff, which is then converted into hydropower. An example of a type one model is the pioneering study by Lehner, Czisch & Vassolo (2005). They created a model which calculated a country's gross hydropower potential (GHP) based on runoff. GHP is the total potential energy available in an area's hydrological system. Hydropower potential models are a subset of hydropower models, which evaluate the potential of hydropower in places where no HPP is realized yet (Hoes et al., 2017).

Besides a type one model, Lehner, Czisch & Vassolo (2005) created a type two model (*natural river flow to hydro*), which is an extended version of a type one model. In a type two model, runoff is converted into streamflow, which is converted into hydropower. Other examples of studies on hydropower potential type 2 models are from Hoes et al. (2017), who use the slope and discharge of rivers to assess the global hydropower potential, and Coskun et al. (2010), who combine remotely sensed precipitation, slope and elevation to assess the hydropower potential of a Turkish river basin with no gauged streamflow data. Existing models of type one and two currently do not solve data scarcity, as they often assess the hydropower potential in a situation with no realized hydropower. The time step of this type of model is commonly at annual interval, which means the variation in generation induced by weather patterns is not captured. Furthermore, these models need specific power plant characteristics such as hydraulic head to function, which might not be known in a data-scarce region.

Type three models (*reservoir storage and release*) are extended type two models, where reservoir storage and release characteristics are added to a type two model. An example of this type of model would be the study by Liu et al. (2019). They present a model which can predict hydropower generation at basin scale, using runoff, upstream basin area, reservoir inflow, and historical generation. The model is applicable worldwide due to their usage of globally available datasets. Due to the limited available data on reservoir storage and release, this model type was not found in the context hydropower data scarcity.

Type four models, *statistical and machine learning models*, directly model the relationship between climate data, streamflow, or reservoir discharge and hydropower. These models are commonly used for short-term forecasting, a subsection of hydropower models where historical electricity generation is combined with other features, such as precipitation, to predict future electricity generation from existing hydropower plants (HPPs) (Turner & Voisin, 2022). An example of a type four model is developed by Wei et al. (2023). They use a neural network model to predict short-term hydropower production in data-scarce regions. Wei et al. (2023) use historical power generation data, meteorological data, and environmental data as input. Statistical ARIMA & ARIMAX models are used by Barzola-Monteses et al. (2019) to create a short-term hydropower generation forecast. A limitation of this approach to prediction in data-scarce regions, is that these types of models require historical generation as model input. The models only function when data scarcity is defined as a lack of real-time generation data. These models cannot function independently if data scarcity is defined as having no historical generation data.

A model predicting hydropower in data-scarce regions is made by Falchetta, Kasamba & Parkinson (2019). They assess hydropower production in Malawi based on satellite data. Hydrological measurements from satellites are combined with nighttime light images to make predictions on hydropower generation without using historical generation data, showing the use case of remotely sensed data for filling data gaps.

Another instance of a type four model is the Hydro Plant Generation Estimation Model (HPGEM) developed by Yin et al. (2020). They applied a gradient boosting tree regressor model to predict annual hydropower at power plant level, based on plant characteristics such as capacity and local runoff, in a situation with limited or no generation data. The HPGEM is able to make predictions for power plants worldwide, as long as they are included in the Global Power Plant Database, which the model uses as input data (Global Energy Observatory, 2021). In a statistical model like the HPGEM from Yin et al. (2020), characteristics of a power plant, such as its type and capacity, are used to estimate correlations between these characteristics and known electricity generation data (Yin et al., 2020).

All model types from one to four can be physical models or statistical models. Physical models use equations that convert physical features such as precipitation and hydraulic head into hydropower output. Statistical models estimate the relationship between predictor and response variables. Type three and four models are also found as a combination of both, hybrid models (Yildiz & Açikgöz, 2021). An overview of the mentioned hydropower models can be found in Table 1, which summarizes the discussed models, their model type and model characteristics.

Table 1

Overview of related hydropower models.

Model type	Reference	Geographic location	Spatio-temporal scale	Hydropower plant type	Variable(s) used	Model approach
1: ***Surface water to hydropower.*** Climate → runoff → hydropower	Lehner, Czisch & Vassolo (2005)	EU	Country-level, yearly	Reservoir & ROR	Runoff	Physical model, gross hydropower potential
2: ***Natural river flow to hydro.*** Climate → runoff → streamflow → hydropower	Lehner, Czisch & Vassolo (2005)	EU	Country-level, yearly	Reservoir & ROR	Streamflow	Physical model, hydropower potential
	Hoes et al. (2017)	Global	Country-level, yearly	All	River slope, discharge	Physical model, hydropower potential
3: ***Reservoir storage and release*** Climate → runoff → streamflow → storage and release → hydropower	Van Vliet et al. (2016)	Global	Plant-level, yearly	All	Streamflow, water temperature, reservoir regulation, hydropower	Physical model
	Coskun et al. (2010)	Eastern Turkey	River basin-level, yearly	All	Elevation, slope, rainfall	Regression, hydropower potential
	Liu et al. (2019)	Global (with validation in China)	Provincial level, monthly	Reservoir	Reservoir inflow	Physical model
4: ***Statistical/ machine learning*** Climate → hydropower Streamflow → hydropower Storage and release → Hydropower	Barzola-Monteses et al. (2019)	Ecuador	River basin-level, Monthly	All	Hydroelectric power production (historical), precipitation	ARIMA/ ARIMAX
	Dabare et al. (2020)	Sri Lanka	Plant-level, monthly/yearly	Reservoir	Rainfall	Regression
	Falchetta, Kasamba & Parkinson (2019)	Malawi	River basin level, Monthly	All	Precipitation, temperature, nighttime light radiance	Regression, random forest
	Wei et al. (2023)	China	Plant scale, daily	ROR	Historical generation, meteorological data, environmental data	CNN-Bi LSTM, short term forecasting
	Yin, Byers, Valeri & Friedrich (2020)	Global	Plant level, yearly	Reservoir and ROR	Capacity, runoff, River size, annual capacity factor per country	Gradient boosting tree regressor

Note. Partly sourced from Barzola-Monteses et al. (2022) and Turner & Voisin (2022), extended with studies relevant to hydropower modeling in data-scarce regions.

Related studies show that most hydropower models function on annual time step, such as the HPGEM model from Yin et al. (2020). Models operating at shorter temporal time steps are found to be mostly forecasting models, using historical generation to predict future hydropower generation or hydropower potential models. An example is the model from Wei et al. (2023), predicting daily generation.

The spatial prediction resolution of the discussed hydropower models is mostly at global, national or river basin scale. The model van Vliet et al. (2016) developed is at plant scale, but uses an annual time step. Models

on finer resolution, such as the plant level, are found to be hydropower potential models or forecasting models. The HPGEM model from Yin et al. (2020) was the only model found to predict at plant scale in a data-scarce situation.
Few researchers have addressed the issue of hydropower modeling in data-scarce regions. Falchetta, Kasamba & Parkinson (2019) used remotely sensed data to estimate hydropower generation in Malawi. Hydropower models applicable in data-scarce regions were only found at annual time step (HPGEM), or were forecasting on the very short term (Wei et al., 2023). The HPGEM model is the closest to a model which can predict hydropower generation at plant scale in a data-scarce region, but it operates at an annual time step, which means the model is not capable of reproducing the hydropower generation fluctuations throughout the year. Furthermore, the HPGEM model can only predict using input data specifically provided by the GPPD, such as a GPPD capacity factor and a GPPD provided ID, limiting the model's usage to hydropower plants in this database. Only a selection of the complete set of hydropower plants worldwide is included in this database. The last version is from 2021, with no updates planned at this moment (Global Energy Observatory, 2021). This means a model currently capable of predicting at plant scale in data-scarce regions and using a monthly time step does not exist.

This study aims at investigating how accurately hydropower can be predicted at the power plant level, using a monthly time step, in data-scarce regions. Based on related studies and the research aim, the following sub-research aims were formulated:

I. *What explanatory variables are important for explaining hydropower generation?*
II. *How does the performance of a linear model compare to a nonlinear model?*
III. *How does the model perform on a monthly time step compared to a yearly one?*
IV. *How does the model perform when applied to an independent dataset?*
V. *How does the model compare to an existing model?*

In the next chapter, the materials and methods used to answer these questions will be presented.

2. Material and methods

A linear and a nonlinear statistical (type four) model were developed to answer the research questions, predicting hydropower generation at a monthly and yearly time step at plant scale in data-scarce regions. This means a total of four models were created. The input data was taken primarily from global spanning datasets, making the models generalizable to many regions.
For model training, a dataset was created with HPP data from the United States (US), as this region has the most hydropower-related data available (Table A1) (Yin et al., 2020). Next to the US dataset, an independent dataset was made containing HPPs from the European Union (EU), as this region has a relatively large amount of HPPs while having few HPPs with generation data (Yin et al., 2020). Both datasets were created on monthly and annual time steps, resulting in a total of four datasets.
The models were used for prediction on datasets with HPPs for the US and from the EU. The assumption was that a model trained on US data was generalizable to other regions, due to the US's heterogeneous climate and terrain characteristics. For the models, predictor variables had to be chosen. Each one will be introduced in the next section.

2.1. Predictor variables

Predictor variables were selected based on variables used in related studies and their availability in data-scarce regions. All related hydropower modes use a predictor variable related to hydrology in some form, like precipitation, runoff or discharge. Since the model needs to predict at the plant scale, a predictor variable for the water flow at the plant location was needed, with data globally available. This led to the decision to use river discharge as a predictor variable, which is defined as the volume of water flowing through a river channel.

In an ideal situation, river discharge data should represent the volume of water flow available to an HPP (reservoir outflow), but this data was not available globally. If data were available on the reservoir inflow related to the specific HPP, that would have been the second-best option, but this was not found at a global scale. There was globally available data on river discharge (hereafter called discharge), as it would have been without anthropogenic changes to the water flow, like reservoirs, dams, or other infrastructure interfering with the water flow. Therefore, discharge was used as a predictor variable.

Reservoir storage and release behavior are important for a hydropower model at a sub-annual time step, as water can be stored throughout multiple months, but in most cases not for longer than a year (Turner & Voisin, 2022). Actual reservoir storage and release patterns are bound by regulations and often impossible to model globally due to unknown regulations. As a proxy for reservoir storage and release, reservoir area was selected as a predictor variable.

Plant capacity was used as predictor variable in comparable models as the HPGEM model from Yin et al. (2020), which is similar in model scope (globally applicable) and purpose (predicting power generation at plant level in a data-scarce situation). Plant capacity is a predictor variable that is assumed to be known for all HPPs, thus was not limiting the model with data availability, and was therefore chosen as a predictor variable.

Hydropower plants show significant differences between types of plants, such as reservoir or run of river (Levasseur et al., 2021). The broad availability of data on the plant type led to the selection of the plant type as a predictor variable.

Dam height was selected as a predictor variable, due to the inclusion of this variable in related hydropower studies. In physical models, the hydraulic head is one of the main parts of model equations, as it drives hydropower generation together with discharge. Hydraulic head data was unavailable for the EU and many HPPs in the US. To include the hydraulic head variable in the model, dam height was chosen as a proxy, as dam height is generally possible to obtain for most HPPs. In the next section, a dataset to use as a source for each of the chosen variables is presented.

2.2. Data sources

Discharge, reservoir area, dam height, capacity, type and historical hydropower generation were used for modeling and for each a dataset was selected to use as source. If possible, the data was taken from datasets globally available. Only when a global dataset was not available, a dataset on a finer spatial scale was used.

There are several global discharge datasets, which can be roughly grouped into grid-based datasets and vector-based datasets. According to Lehner & Grill (2013), vector-based hydrology models can be preferable over grid-based models at finer spatial scales. To accurately model a river network in a grid-based model, the resolution needs to be high, with a single grid cell being able to differentiate between rivers. A vector-based network implements rivers as lines that connect, which leads to a more precise representation of the river network. The selected discharge dataset *Global Reach-level Flood Reanalysis* (GRFR) consists of 3-hourly discharge data for over 2.94 million river reaches worldwide from 1979-2019 (Yang et al., 2021). GRFR takes the river reach vectors from MERIT, which is a vector-based globally spanning dataset, with a median length of 6.8 kilometers for each river reach (Lin et al., 2019; Yang et al., 2021). The GRFR assumes no anthropogenic changes were made to the water flow, meaning dams, reservoirs, and other infrastructure that changes the water flow are not modeled. The discharge is calculated at the most downstream point of a river reach, and assigned to the whole river reach.

Remotely sensed reservoir area was used as a proxy for reservoir behavior. The reservoir area *Global Reservoir Surface Area Dataset* (GRSAD) was selected since it consists of data on reservoirs globally (Gao & Zhao, 2019).

This dataset contains the monthly reservoir surface area of 6817 reservoirs worldwide, from 1984 to 2018. The reservoirs represented in GRSAD are taken from the *Global Reservoir and Dam Database* (GRanD) (Lehner et al., 2011).

For plant locations in the US, the *Hydropower Infrastructure - LAkes, Reservoirs, and RIvers* (HILARRI) dataset was used, which has data on 1652 HPPs (Hansen & Matson, 2023). The dataset was created to link different datasets related to hydropower together, and contains the most important plant identifiers for that reason. The European HPP locations are taken from the *JRC PBBD OPEN* dataset, which contains locations of 1526 HPPs (Kanellopoulos et al., 2019).

Historical hydropower generation in the US (2001-2020) for 1505 HPPs was taken from RectifHyd (Turner, Voisin & Nelson, 2022). RectifHyd contains monthly reported data from the EIA. Hydropower generation data in the EU is taken from the *Global Power Plant Database* (GPPD) (Global Energy Observatory, 2021). This dataset contains data on 7156 HPPs worldwide, of which 372 can be linked to HPPs in the EU. The GPPD generation data is available at a yearly time step, meaning that the validation of the model on EU data can only occur yearly.

Dam height data was taken from the *Global Reservoir and Dam Database* (GRanD) (Lehner et al., 2011). The GRanD dataset contains 6862 dams and reservoirs worldwide, with 1902 in the US and 914 in the EU. For the US, dam height data from GRanD was combined with dam height data from the National Inventory of Dams (NID, 2023).

Capacity data was taken from the *Existing Hydropower Assets Capacity Plant* dataset in the US, for the years 2005-2022, and includes 1480 power plants, with a minimum of 1 MW in capacity (Johnson, 2023). Because hydropower plant capacity can change over the years due to plant modifications, plant capacity for the US was implemented as a value that can differ from year to year.
For the EU, capacity data was taken from the GPPD (Global Energy Observatory, 2021), and missing values were filled in by capacity data from JRC PBBD OPEN and JRC Hydro-power dataset (JRC Hydro-power database, 2019). For the EU no data was found on the yearly historical development of the capacity, which meant capacity was implemented as a fixed value.

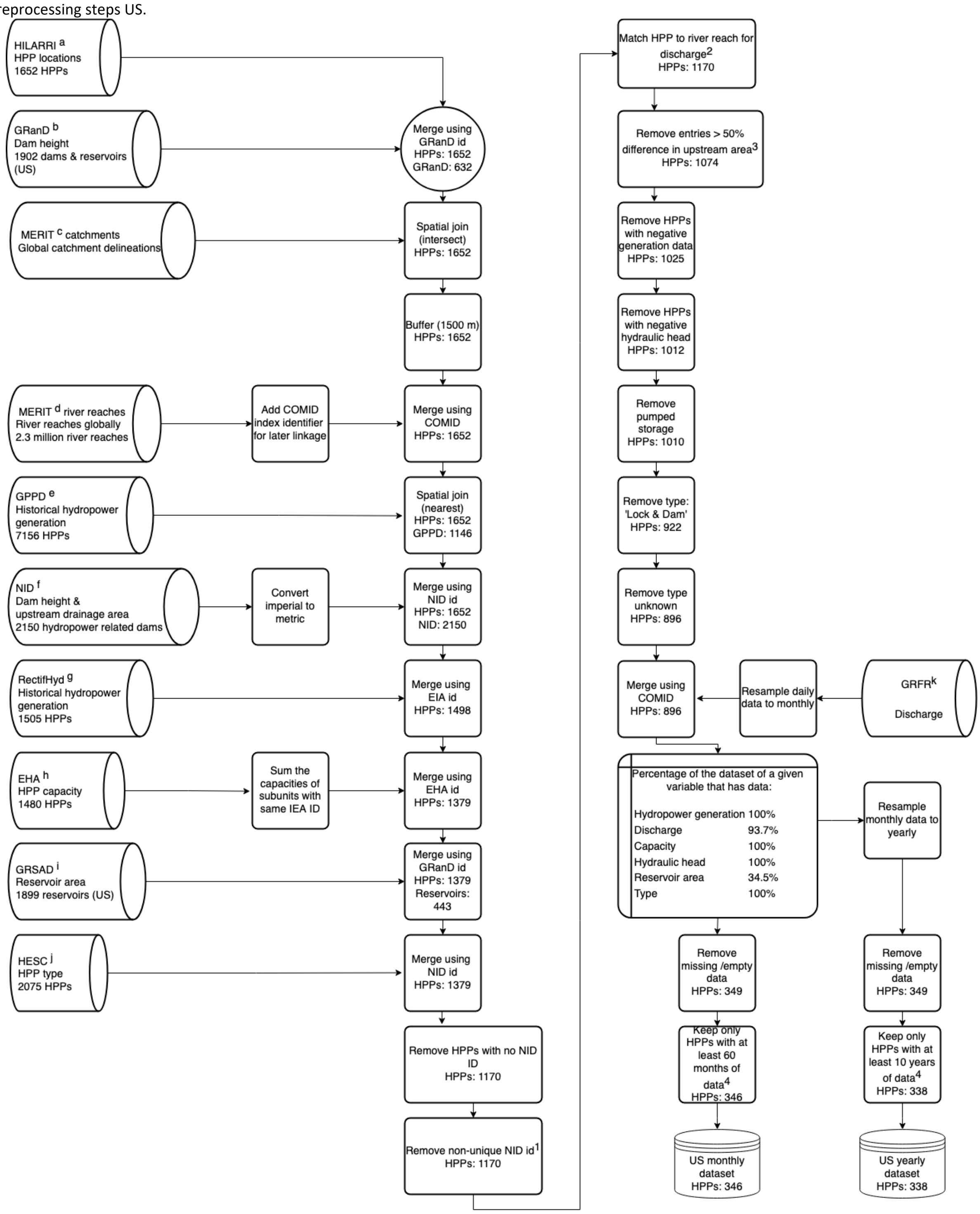
For the US, the *Hydropower Energy Storage Capacity* dataset (HESC) was used as the source for the Type (Hansen, Ghimire & Gangrade, 2021). The type was taken from the JRC PBBD OPEN dataset for the EU dataset. All datasets had to be combined, to create input data for model fitting and predictions. The preprocessing steps taken will now be introduced.

2.3. Data preprocessing

In the US, the HILARRI dataset was used as starting point. This dataset contains the location, EIA ID, GRanD ID, and NID ID of each HPP in the dataset, and is therefore ideal for linking other datasets together.
The EIA ID of an HPP was chosen as the main identifier and was used to distinguish a single HPP unit, as the EIA and RectifHyd present their generation data based on EIA IDs. There can be multiple HPPs with their own EIA ID belonging to the same infrastructural complex, since it is common for a large HPP to consist of multiple units with their own EIA ID and their generation data. These were kept as separate plants, as in some cases, they can be separated by a considerable distance and may have different plant characteristics such as capacity and dam height. Figure 1 shows all preprocessing steps taken, starting at the HILARRI dataset.

Figure 1.

Preprocessing steps US.



[a] HPP locations from *Hydropower Infrastructure - LAkes, Reservoirs, and Rivers* (HILARRI), Hansen & Matson (2023)

[b] Hydraulic head from *Global Reservoir and Dam Database* (GRanD), Lehner et al. (2011)

[c] River reaches taken from MERIT, Lin et al. (2019)

Footnotes continue on the next page.

---

[1] HPPs are linked to a NID id (dam identifier), but NID id's are not unique. If they belong to the same reservoir, multiple dams can have the same NID id. To select the correct dam, the distance between each NID dam sharing the same ID, and the linked HPP is calculated. The NID dam with the smallest distance to the HPP is chosen as the correct match.

[2] See Figure 2

[3] See Appendix B1

[d] River catchments taken from MERIT, Lin et al. (2019)
[e] *Global Power Plant Database (GPPD),* Global Energy Observatory (2021)
[f] National Inventory of Dams, NID (2023)
[g] Historical generation from RectifHyd, Turner, Voisin & Nelson (2022)
[h] Capacity from *Existing Hydropower Assets Capacity Plant* (Johnson, 2023)
[i] Reservoir area from Global Reservoir Surface Area Dataset (GRSAD), Gao & Zhao (2019)
[j] HPP type from *hydropower energy storage capacity* dataset (HESC) (Hansen, Ghimire & Gangrade, 2021)
[k] Discharge data from Global Reach-level Flood Reanalysis (GRFR), Yang et al. (2021)

HPPs had to be linked to the river reach which would most closely represent their actual source of water. This was done based on the similarity of the upstream drainage area reported by an HPP and the nearby river reach.

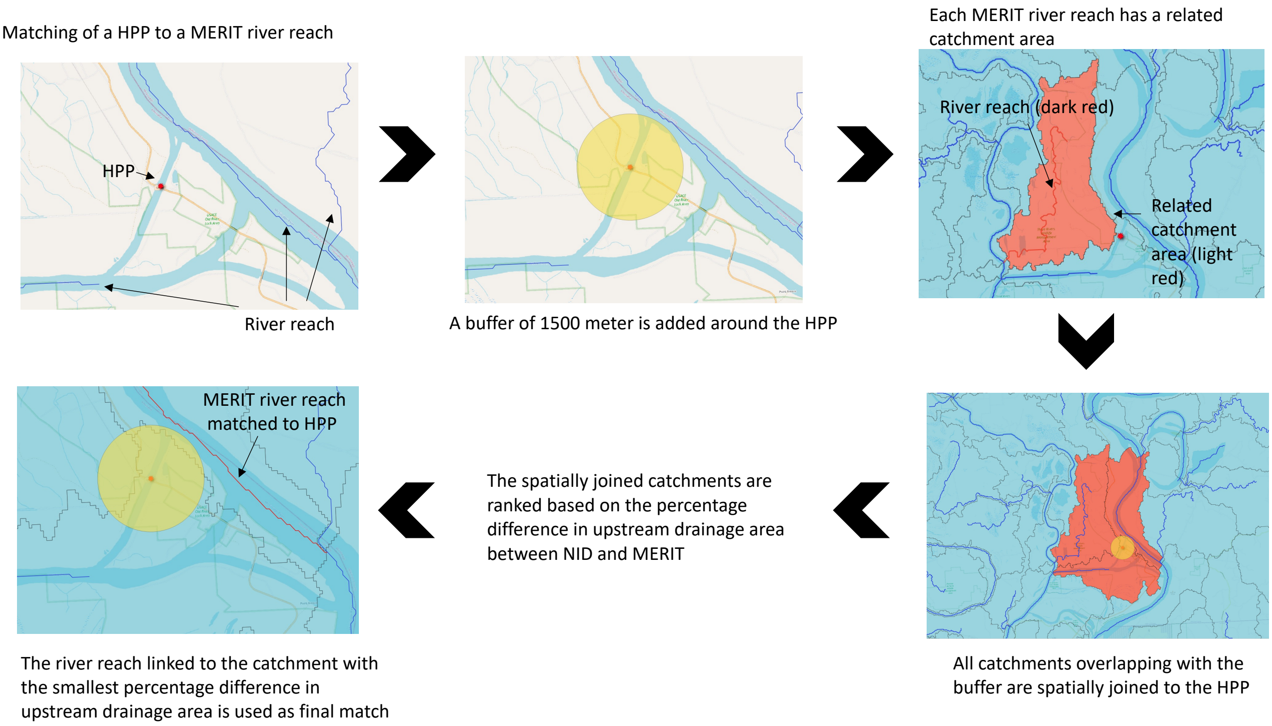
The river reaches included in MERIT are for some complex river systems not representative. When a river splits into multiple parts, only one of the branches is represented in MERIT (Figure 2, top left). If a river branch is not represented in MERIT and hosts an HPP, this HPP still needs to be matched to a river reach.

The first step for the matching process was adding a buffer of 1500 meter around each HPP, based on visual inspection in QGIS of the maximum distance from an HPP to the correct river reach (Figure 2).

After the buffer was added, a spatial join was applied between the buffered HPP locations, and the river catchment area related to a river reach.

Figure 2.

Matching of a HPP to a MERIT river reach



If a final match resulted in more than 50% difference in the upstream drainage area between the HPP and the river reach, the HPP was removed from the dataset.

Not all HPPs had the same number of measurements of historical generation data from RectifHyd. To find a balance between including as many HPPs as possible and removing time series with too few data points to be useful for modeling, a histogram was made of the number of data points for each HPP (Figure B2). This revealed that the most common number of data points for the monthly dataset was 168 months. An upward trend started around 60 months, which was chosen as the minimum number of data points for an HPP to be included in the dataset. The same procedure was applied for the yearly dataset, resulting in a cutoff point of 10 years of data, below which the HPPs were removed from the dataset. HPPs with the type of reservoir and run of river (ROR) were kept in the dataset, while pumped storage, lock & dam, and HPPs with an unknown type were removed from the dataset.

For the US, two datasets were created, one with a monthly and one with an annual temporal resolution. The monthly dataset for the US had a summed capacity of 27.5 GW for the year 2018. This was 34% of the total capacity of the US, as reported by the Department of Energy (Uría-Martínez et al., 2021). The summed average reservoir area included in the dataset for the year 2018 was 38% of the total reservoir area in the US, or about 14,916 km2, as reported by GeoDAR (Wang et al., 2022). Table 2 presents a summary of data on all variables in the dataset. The table shows that the variables are skewed, which can lead to problems with heteroscedasticity during the modeling stage. The same preprocessing steps were performed for the monthly and annual datasets, as models created with both datasets were compared with each other and should have similar input data.

Table 2.

Summary statistics of the explanatory and response variables of HPPs in the period 2006-2018. The table is for the monthly dataset (346 HPPs), a similar table for the annual dataset can be found in Table B1.

Variable	Unit	Mean	Median	SD[a]	Min[a]	Max[a]	Skew	Source
Hydropower net generation	MWh	1.88E+04	5.10E+03	4.25E+04	0.00E+00	7.38E+05	5.67	RectifHyd[b]
Discharge	m^3/s	9.95E+01	3.39E+01	2.19E+02	0.00E+00	7.24E+03	7.67	GRFR[c]
Capacity	MW	7.99E+01	2.62E+01	1.57E+02	1.00E+00	1.31E+03	4.47	EHA[d]
Dam Height	meter	4.91E+01	3.78E+01	4.18E+01	0.00E+00	2.23E+02	1.44	GRanD[e]
Reservoir area	km^2	4.19E+01	9.46E+00	1.16E+02	1.00E-02	1.30E+03	6.85	GRSAD[f]

[a]SD = Standard deviation; Min = minimum; Max = maximum

[b]Turner, Voisin & Nelson, 2022

[c] Yang et al., 2021

[d] Johnson, 2021

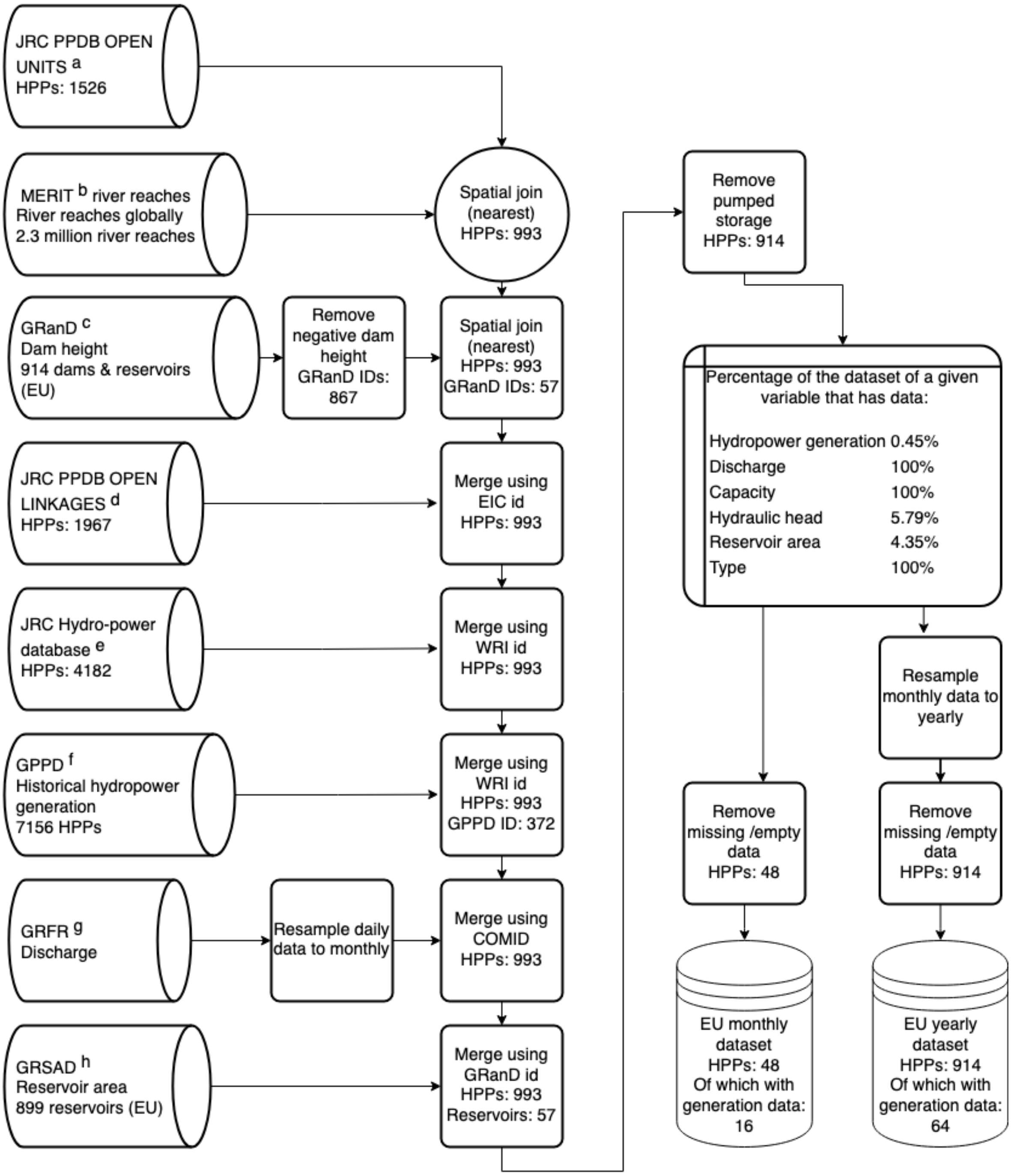
[e] Lehner et al., 2011

[f] Global Reservoir Surface Area Dataset (GRSAD), Gao & Zhao, 2019

For the EU, the JRC PBBD OPEN dataset was used as the starting dataset to which other datasets were linked. The EIC ID (Energy Identification Code, from ENTSO-E), was used to define a single HPP. The same procedure as for the US dataset was followed where possible. An exception was that matching based on the upstream drainage area was impossible due to a lack of data. The HPP to river reach match was done by a spatial join to the nearest river reach (Figure 3). Very few HPPs (57) could be linked to a GRanD ID. As the GRanD id was used for linking with the GRSAD reservoir area dataset and for dam height data, only 57 HPPs could be used for the EU dataset. Due to removing of pumped storage and other missing data, 48 HPPs remained in the EU dataset for the monthly dataset, of which 16 had generation data from GPPD (Figure 3). For the annual dataset, only capacity and discharge were included, due to the choices explained in Chapter 3.1 (model fitting). This meant no missing data for reservoir area had to be removed, resulting in 914 HPPs with capacity and discharge data, of which 64 had historical generation data. Predictions were created for all 914 (annual) and 48 (monthly) HPPs, while performance could only be calculated for the 64 (annual) and 16 (monthly) HPPs with generation data in the EU. Table 3 shows the summary statistics of the EU dataset, with skewed distributions for all variables, although considerably less than for the US dataset. The spatial distribution of the HPPs is evenly throughout the EU, as can be seen in Figure B3.

Figure 3.

Preprocessing steps for EU monthly and yearly datasets.



[a] JRC PBBD OPEN dataset, Kanellopoulos et al. (2019)

[b] River reaches from MERIT, Lin et al. (2019)

[c] Dam height from Global Reservoir and Dam Database (GRanD) (Lehner et al., 2011)

[d] JRC PBBD OPEN LINKAGES, Kanellopoulos et al. (2019)

[e] JRC Hydro-power database (2019)

[f] Global Power Plant Database (GPPD), Global Energy Observatory (2021)

[g] Discharge data from Global Reach-level Flood Reanalysis (GRFR), Yang et al. (2021)

[h] Reservoir area from Global Reservoir Surface Area Dataset (GRSAD), Gao & Zhao (2019)

Table 3.
Summary statistics of the EU dataset of hydropower plants in the period 2005-2018. The Table shows the summary for the monthly dataset containing 48 individual HPPs, of which 16 had generation data. A similar table for the annual dataset can be found in Table B2.

Variable	Unit	Mean	Median	SD[a]	Min[a]	Max[a]	Skew	Source
Hydropower net generation	MWh	5.48E+02	3.61E+02	5.29E+02	1.20E+02	2.35E+03	2.11	GPPD[b]
Discharge	m^3/s	5.34E+02	2.47E+01	1.79E+03	0.00E+00	1.96E+04	4.44	GRFR[b]
Capacity	MW	2.59E+02	1.91E+02	2.31E+02	1.20E+01	1.16E+03	2.28	JRC[d]
Dam Height	meter	9.01E+01	8.90E+01	3.45E+01	2.10E+01	1.68E+02	0.32	JRC[d]
Reservoir area	km^2	1.72E+01	4.83E+00	2.70E+01	3.00E-02	1.08E+02	2.16	GRSAD[e]

[a] SD = Standard deviation; Min = minimum; Max = maximum

[b] Global Power Plant Database (GPPD), Global Energy Observatory (2021)

[c] Yang et al. (2021)

[d] Kanellopoulos et al. (2019); JRC Hydro-power database (2019)

[e] Gao et al. (2019)

2.4. Model fitting

Two types of models were created, a linear model and a nonlinear model. A linear least square regression model was created, due to the relatively simple structure and the interpretable nature of the model. Linear least square regression is a well-known technique to evaluate the relationship between explanatory and dependent variables (Sharma et al., 2011).
The datasets for the US and EU exist out of longitudinal data, with a time series for each HPP. There are several methods to model longitudinal data. A simple solution would be to ignore the longitudinal structure of the data, which would remove the information that some data points belong to the same HPP. To use the fact that some data in the dataset belongs to the same individual unit (the HPP), a mixed-effects model was used (Bates, 2010). This allowed the model to learn general regression coefficients for the fixed effects (the chosen predictor variables), while learning a different random effect for each HPP, to account for unknown differences between HPPs. Each HPP in the dataset is seen as an individual unit and as a sample from a larger population of worldwide HPPs. The set of predictor variables was selected by calculating the Akaike Information Criterion (AIC) for each set of variables, and selecting the model with the lowest score. Due to the skewed nature of hydropower generation data, the models use a capacity factor as the dependent variable, an approach taken similar to the HPGEM model (Yin et al., 2020). The capacity factor is the ratio of the actual generation to the theoretically possible generation in a given timeframe. Would an HPP produce at maximum capacity during a whole period, the capacity factor for that period would be one. After predictions are made, the predicted capacity factors are converted to hydropower using the capacity of the HPP.

Besides a linear regression model, a nonlinear random forest model was selected as second model type. A second model was tried, as the assumption was that a linear model might be too simple to learn the relationship between predictor and response variables, and that the highly skewed nature of the dataset would lead to modeling problems for a linear model. Random forest models do not make any assumptions on normality of residuals and can therefore be widely used to learn nonlinear relationships between predictor and response variables (Schonlau & Zou, 2020). A mixed-effects random forest is used to account for the longitudinal structure of the dataset (Hajjem, Bellavance & Larocque, 2014).

2.4.1. Linear mixed-effects regression

The linear mixed-effects regression (LMER) model is implemented using the *lme4* library for R (Bates et al., 2014; R Development Core Team, 2005). Each HPP was modeled as an individual unit and is treated as the random effect (Bates et al., 2014). Multicollinearity was assessed by calculating the Variance Inflation Factor (VIF) values for each variable. These were between 1.01 and 1.26 for all predictor variables, well below the commonly used threshold of 5, which indicates no problem with multicollinearity should be expected (Menard, 2001).

The model formula used for the LMER model was in matrix notation:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \tag{1}$$

where:
y_i is the $n_i \times 1$ response vector for the *i*-th HPP.
X_i is the $n_i \times p$ model matrix for the fixed effects for observations in HPP *i*.
β is the $p \times 1$ vector of fixed-effect coefficients.
Z_i is the $n_i \times q$ model matrix for the random effects for observations in HPP *i*.
b_i is the $q \times 1$ random effect coefficients vector for HPP *i*, which is different for each HPP, but assumed to come from the same distribution.
ε_i is the $n_i \times 1$ vector of residuals for HPP *i.*
i is the HPP index.
p is the number of fixed effects variables.
q is the number of random effects variables.
n_i is the length of the time series belonging to HPP *i.*

A 5-fold cross-validation was performed to assess the model's performance on data for the US. The LMER model used for predictions in the EU was fitted using the complete dataset of the US. The model was then used to predict the hydropower generation for the EU.

2.4.2. Mixed-effects random forest

Beside the linear mixed-effects regression, a mixed-effects random forest (MERF) model was created. For this model, the Python package *merf* was used (Van Rossum & Drake, 2009; Breiman, 2001; Merf, 2023). The default model parameters were used. The following formula can describe the MERF model:

$$y_i = f(X_i) + Z_i b_i + \varepsilon_i \tag{1}$$

where:
y_i is the $n_i \times 1$ response vector for the *i* -th HPP.
X_i is the $n_i \times p$ model matrix for the fixed effects for observations in HPP *i*.
$f(X_i)$ is a nonlinear function, estimated using a random forest model.
Z_i is the $n_i \times q$ model matrix for the random effects for observations in HPP *i*.
b_i is the $q \times 1$ random effect coefficients vector for HPP *i*, which is different for each HPP, but assumed to come from the same distribution.
ε_i is the $n_i \times 1$ vector of residuals for HPP *i*.
i is the HPP index.
p is the number of fixed effects variables.
q is the number of random effects variables.
n_i is the length of the time series belonging to HPP *i.*

For the US MERF model a 5-fold cross validation was performed. The EU predictions were created using a model fitted on the complete US dataset. Predictions were performed on both monthly and yearly time step, while evaluation was done at a yearly time step due to data limitations.

To assess model performance, a feature permutation importance algorithm was used. Feature permutation importance can be applied to any model, as it assesses the model performance of a baseline (unaltered) model versus the model performance when one of the features is permutated by randomizing all data for the selected feature. This process was repeated for all features. The feature permutation importance algorithm was implemented as outlined by scikit-learn (Scikit-learn permutation importance, 2023).

A prediction ceiling was implemented as the plant capacity multiplied by 0.9 (the assumed efficiency of an average HPP), multiplied by the number of hours in a given period (month or year), to make sure no predictions were made which were physically impossible. Each HPP was assumed to have similar efficiency, as exact data were not available at the power plant level. In the case of negative predictions, these were set to zero. All other values were kept as they were predicted by the model. This means that predictions ranged from zero, to the physically maximum a HPP could generate.

2.5. Model evaluation

The evaluation metrics Kling-Gupta Efficiency (KGE) and Normalized Root Mean Square Error (NRMSE) were used. The advantage of KGE is that it is scale independent and can easily be compared across different HPPs. It is commonly used in hydrology (Gupta et al., 2009). KGE is dimensionless and falls on the interval of negative infinity to one, with a score of one being optimal. KGE is used outside of hydrology as well as a prediction metric, for example by Zhang et al. (2019) for evaluating a solar PV model (Zhang et al., 2019). The formula to calculate KGE is as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{2}$$

$$r = \frac{cov(y, \hat{y})}{\sigma(y) \cdot \sigma(\hat{y})} \tag{3}$$

$$\alpha = \frac{\sigma(\hat{y})}{\sigma(y)} \tag{4}$$

$$\beta = \frac{\mu(\hat{y})}{\mu(y)} \tag{5}$$

where y is the observed time series, $\hat{y}$ is the predicted time series, *cov* is the covariance, σ is the standard deviation, and μ is the mean.

RMSE is another well-known metric to evaluate predictive models (Hyndman & Koehler, 2006). RMSE is in the unit of the response variable, which would be MWh (hydropower generation). Due to large differences in electricity generation between HPPs, large differences in RMSE between HPPs occur. To account for these differences the RMSE was normalized (NRMSE) using the mean of observations, which is a common normalizer (Ssekulima et al., 2016). NRMSE ranges from zero to infinity, where zero would be the optimal score. NRMSE can be calculated in the following way:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{6}$$

$$NRMSE = \frac{RMSE}{\mu(y)} \tag{7}$$

Where y_i is the *i*-th observation of time series y, $\hat{y_i}$ is the *i*-th prediction of time series $\hat{y}$ and μ is the mean.

The resulting distribution of KGE and NRMSE values (one for each HPP) were visually and quantitively evaluated by plotting the histogram of the scores for individual HPPs, with the median and mean KGE and NRMSE added. Based on these measures, the model performance was evaluated.

3. Results

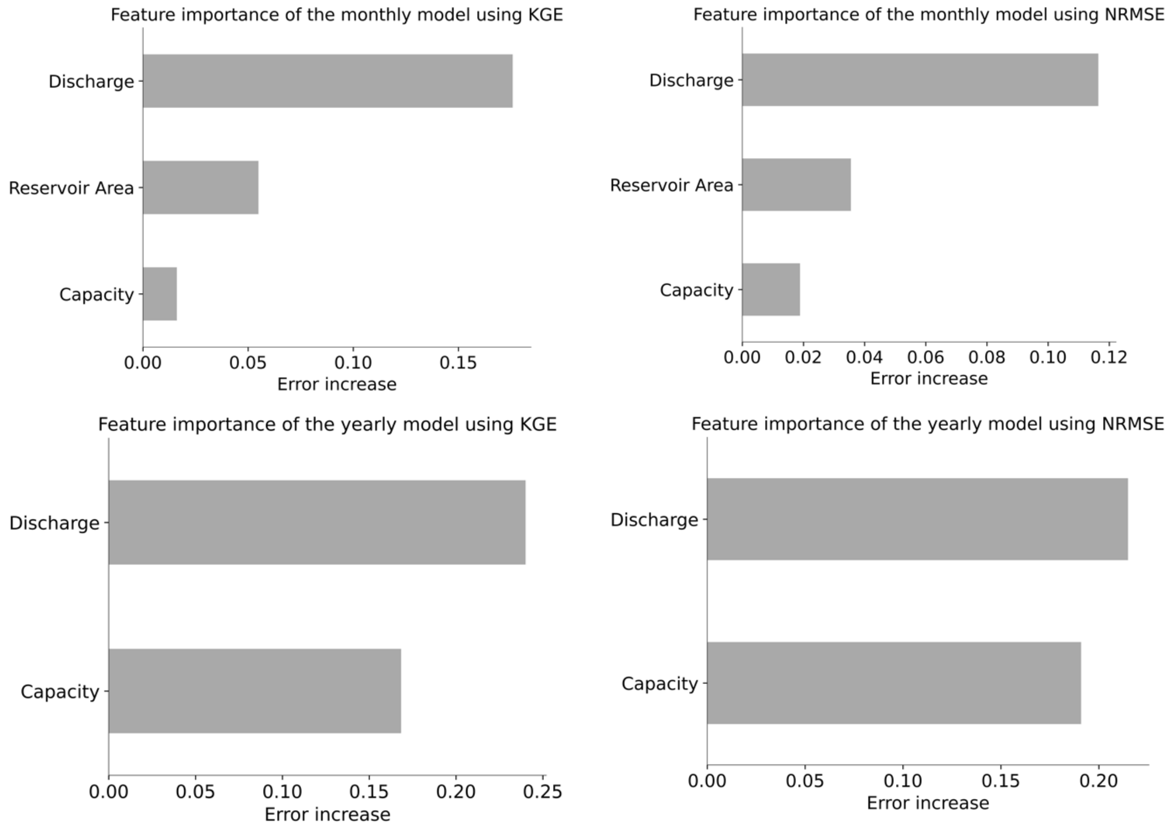
3.1. Model fitting

To determine the ability of a model to predict hydropower on a plant scale, on a monthly time step in data-scarce regions, first the chosen predictor variables will be examined on importance and significance.
To select the final model predictor variables from the set of chosen predictor variables, AIC scores were calculated for each set of variables. The monthly LMER regression model with discharge, capacity and reservoir area had the lowest AIC (Table C1), although differences were relatively small. For the annual LMER model the model with discharge and capacity had the lowest AIC score. The model with the lowest AIC scores had the best trade-off between complexity and accuracy, so these were chosen as final prediction variables. All regression coefficients for the chosen predictor variables were significant. (Table C2 and C3).

After evaluating the chosen predictor variables and selecting the predictor variables, the feature importance was examined. The feature importance of the LMER model revealed that for the monthly model, discharge was the most important variable with a distance, followed by reservoir area and at last capacity (Figure 4). The yearly model showed the same order as the monthly feature importance, but without the reservoir area as this predictor variable was removed from the final model. For the yearly model, capacity was relatively more important compared to the monthly model. There were no differences in the order of importance between the evaluation on KGE and on NRMSE.

Figure 4.

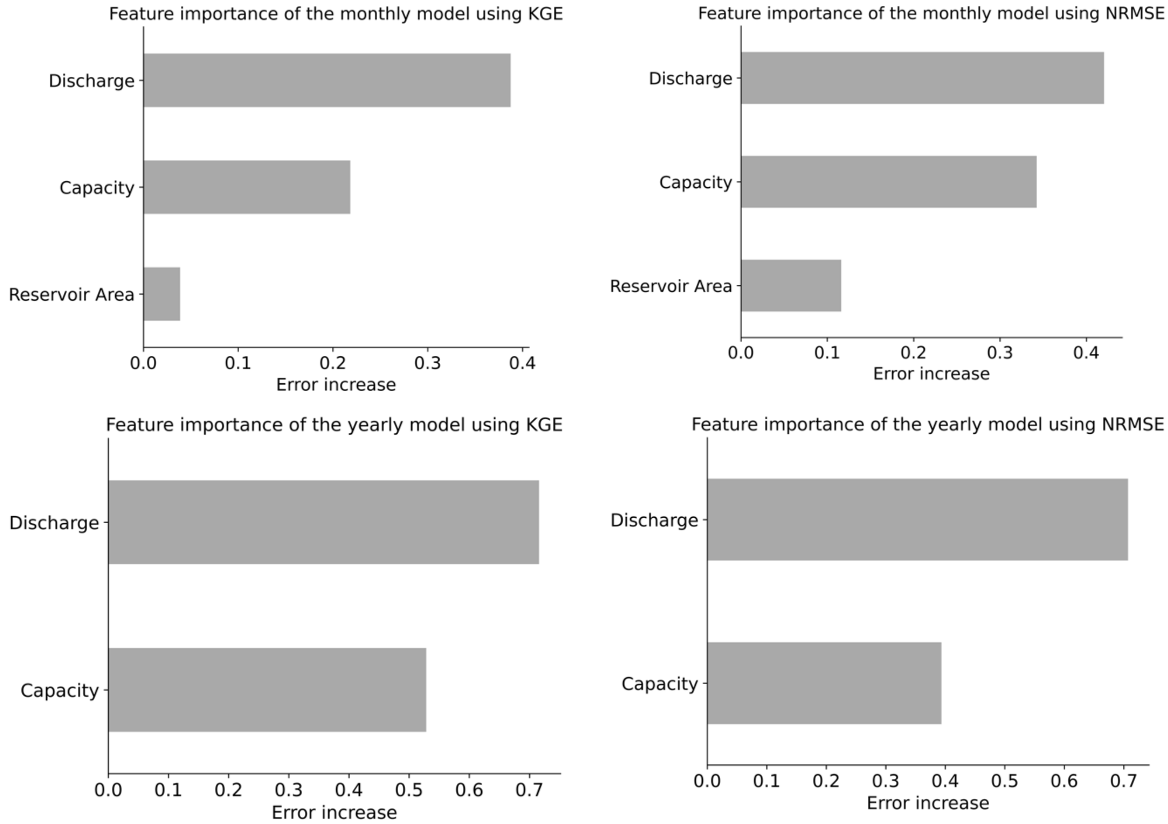
Feature importance results for the LMER model. Top row shows monthly model feature importance, bottom row shows yearly. Left shows feature importance based on KGE, right is based on NRMSE.



The feature importance of the MERF models showed that the discharge was for both the monthly and yearly MERF model the most important feature, followed by capacity (Figure 5). Reservoir area was the least important for the monthly MERF model, while being second for the LMER model.

Figure 5.

Feature importance results for the MERF model. Top row shows monthly model feature importance, bottom row shows yearly. Left shows feature importance based on KGE, right is based on NRMSE.



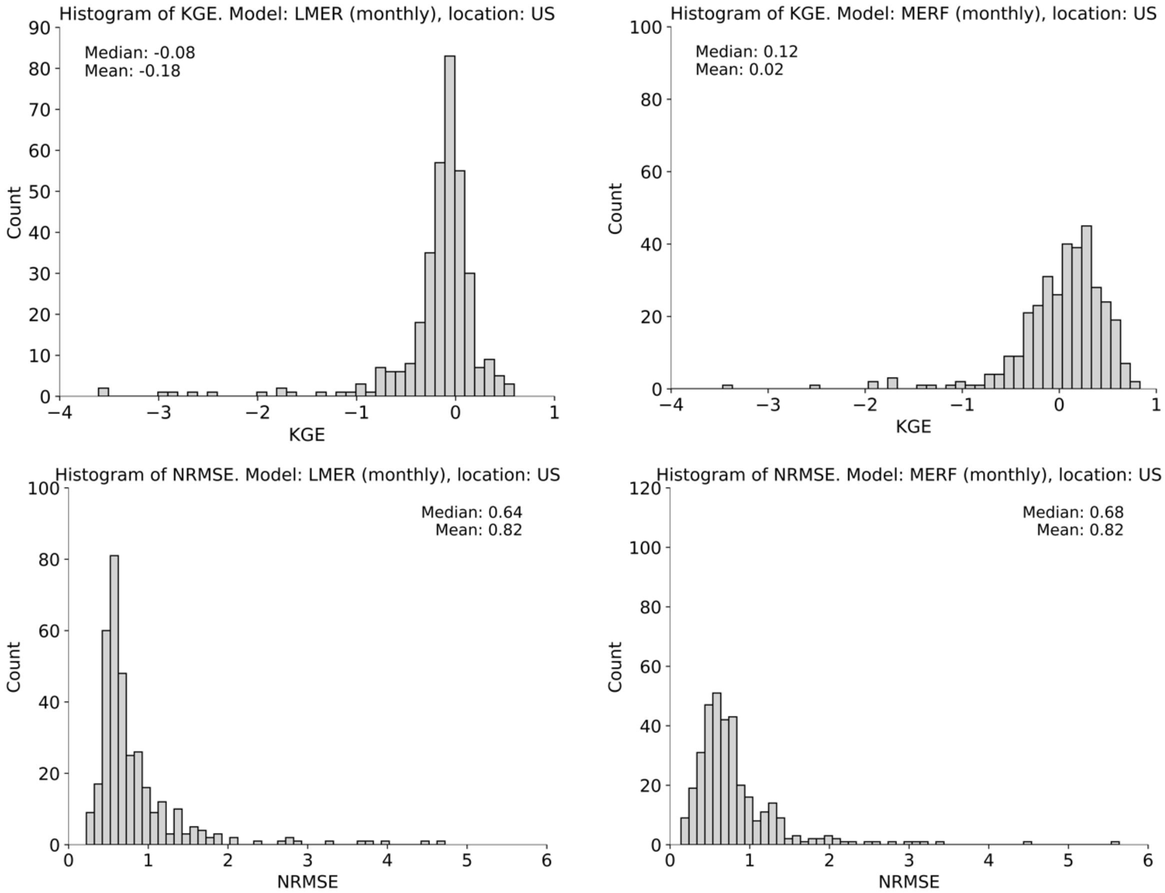
3.2. Model prediction

To test the predictive accuracy of the monthly LMER and MERF models, both models were compared to each other on KGE and NRMSE scores.

This resulted in only minor differences between both models on predictive accuracy (Figure 6). Figure 6 shows the KGE and NRMSE results at plant scale, for all HPPs. A single KGE and NRMSE score were calculated for each HPP, by using the predicted time series and the historical generation time series from RectifHyd. All calculated KGE and NRMSE scores are shown in the histogram below, giving an overview of model performance on all HPPs in the dataset. The MERF model had a slightly better median and mean score when evaluated on KGE (median KGE of 0.12, while the LMER model had a median KGE of -0.08), while when evaluated on NRMSE both models performed almost exactly similar (the MERF model had a NRMSE score of 0.68 while the LMER model had a NRMSE score of 0.64). The distribution of the LMER model had a higher peak (higher kurtosis), while the distribution of the MERF model was more evenly spread (lower kurtosis).

Figure 6.

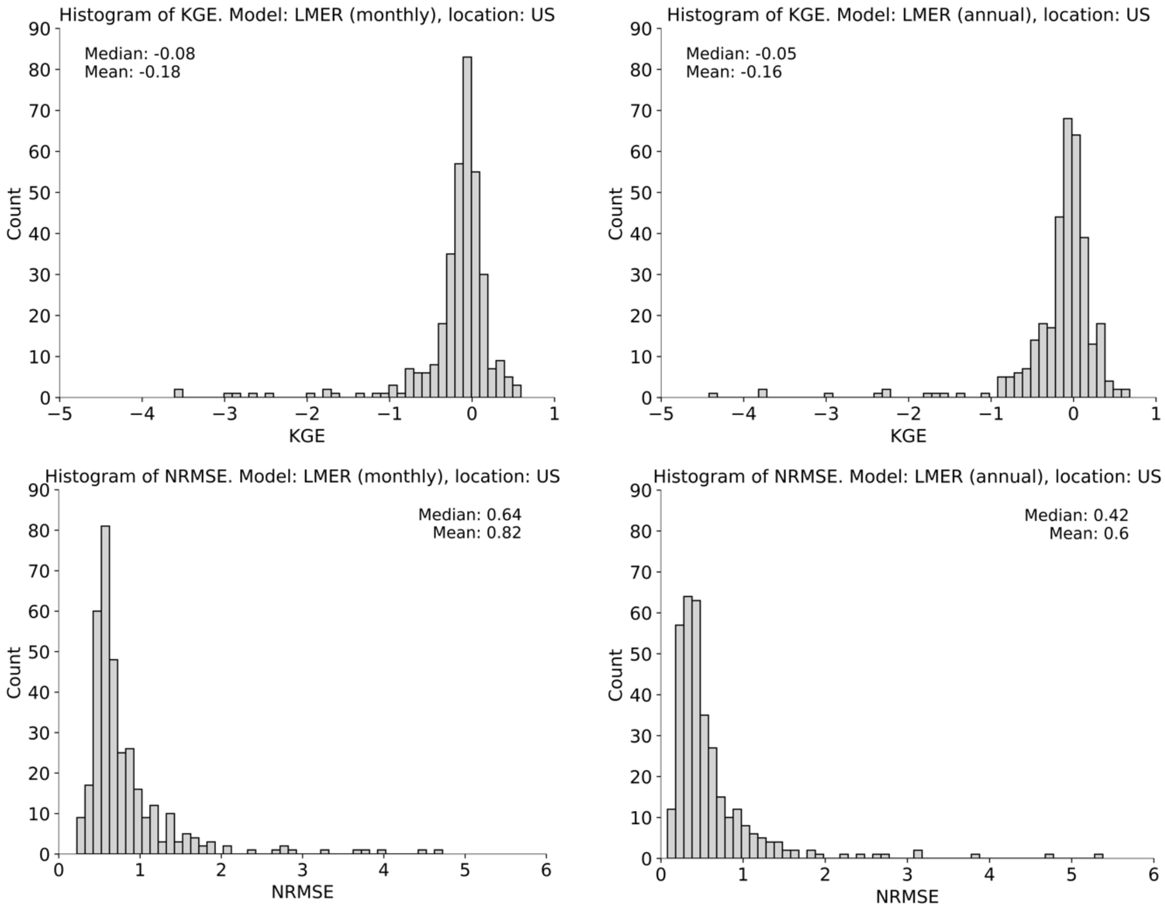
Left: LMER, right: MERF. The (m) denotes the monthly time step of the model. The US dataset is used for prediction. The KGE and NRMSE scores are calculated using the whole predicted and measured time series (from RectifHyd), and are at plant scale. The figures show the histogram of all KGE or NRSME scores for all HPPs in the dataset.



To assess the model performance of a model on a monthly time step compared to a model at an annual time step, KGE and NRMSE scores for both models were evaluated using the US models (Figure 7).
Differences were minor when the monthly LMER model was compared to the yearly LMER model. The median KGE and NRMSE score of the yearly model was overall a fraction better than that of the monthly model (KGE yearly was -0.05, while -0.08 for the monthly model).

Figure 7.

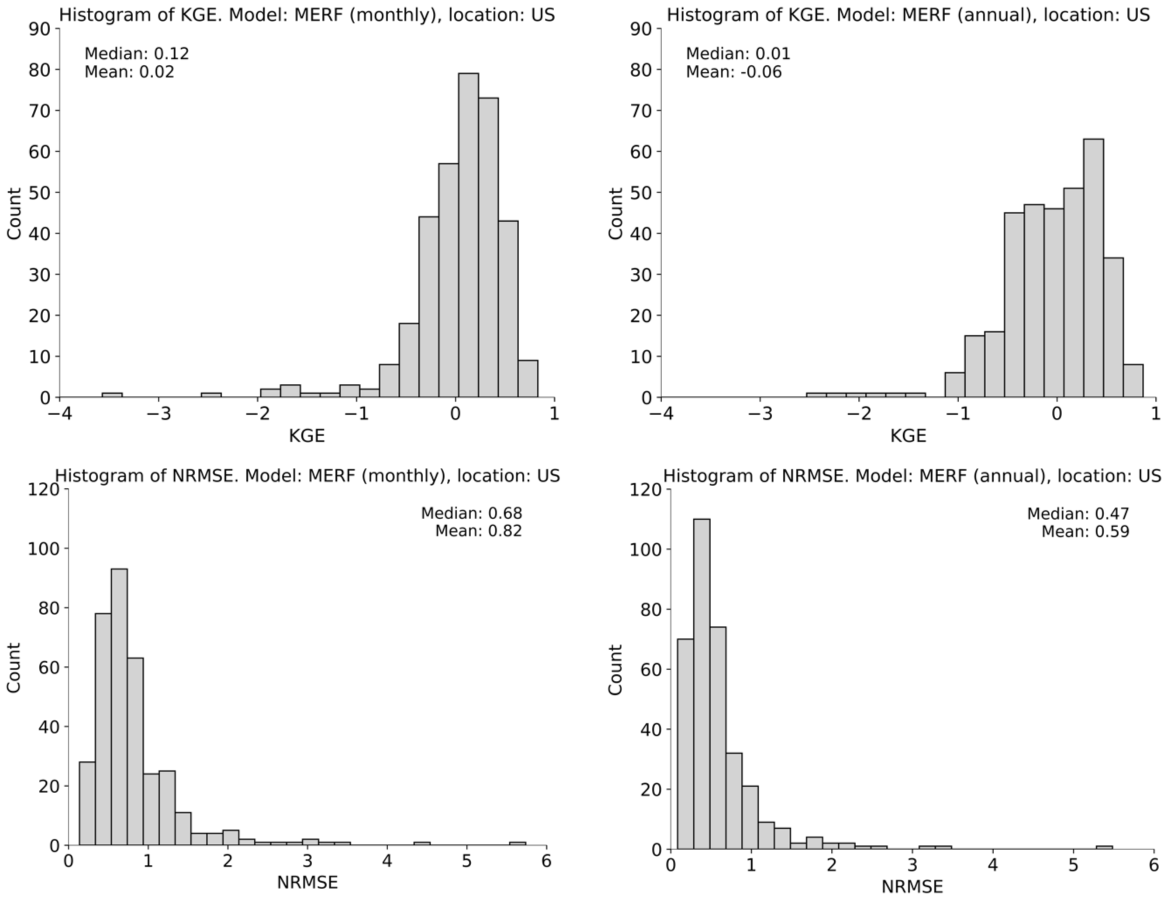
Monthly LMER model compared to yearly for the US. Between brackets the time step is given, monthly or annual. The KGE and NRMSE scores are calculated using the whole predicted and measured time series (from RectifHyd), and are at plant scale. The figures show the histogram of all KGE or NRSME scores for all HPPs in the dataset.



For the MERF model the annual model performs slightly better on median NRMSE, while the monthly model performs better on median KGE (Figure 8).

Figure 8.

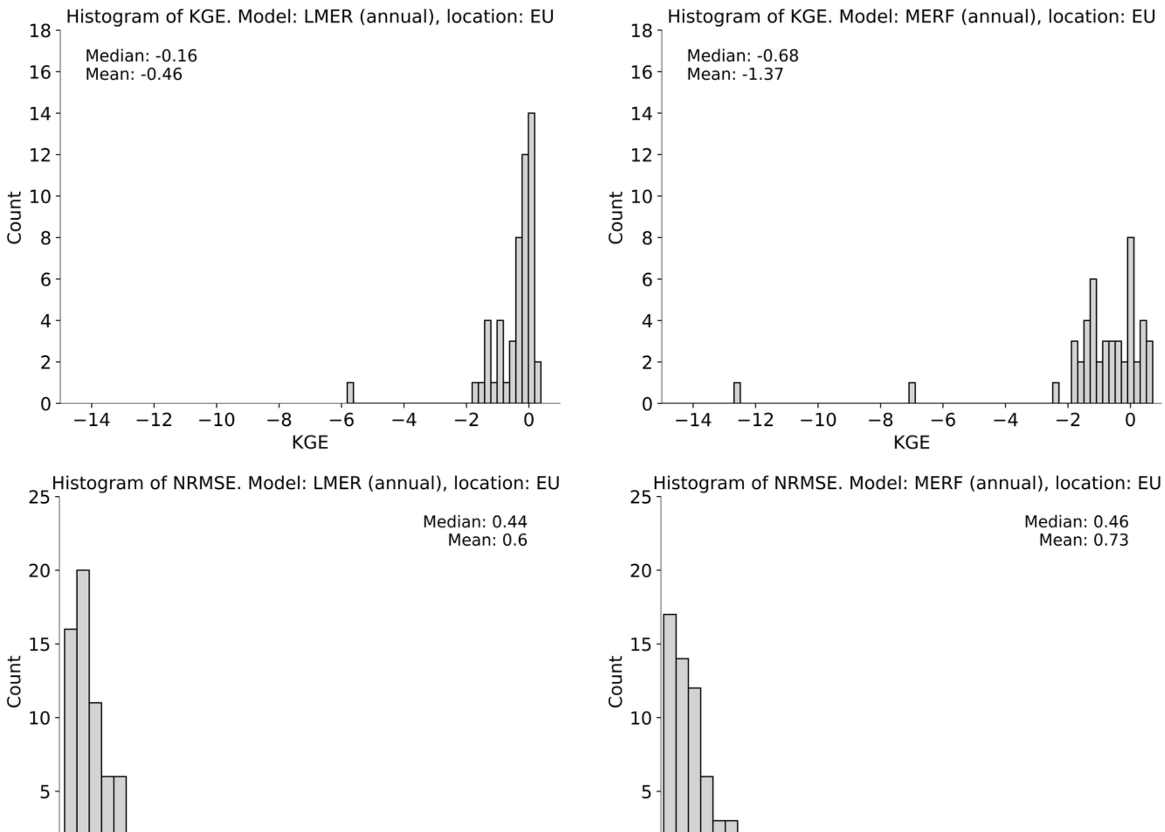
Monthly MERF model compared to yearly for the US. Between brackets the time step is given, (m) for monthly and (a) for annual. The KGE and NRMSE scores are calculated using the whole predicted and measured time series (from RectifHyd), and are at plant scale. The figures show the histogram of all KGE or NRSME scores for all HPPs in the dataset.



The LMER and MERF models were examined on an independent dataset data from the EU. This was done for models with an annual time step, due to historical generation data only being available on this temporal interval on a large scale for the EU. The model predictions for US and EU were compared to each other. This was done for LMER and MERF models (Figure 9). The LMER model performed better on median KGE and median NRMSE, with minor differences on median NRMSE. When evaluated on median KGE, the LMER model improved over the MERF model (-0.16 for LMER and -0.68 for MERF).

Figure 9.

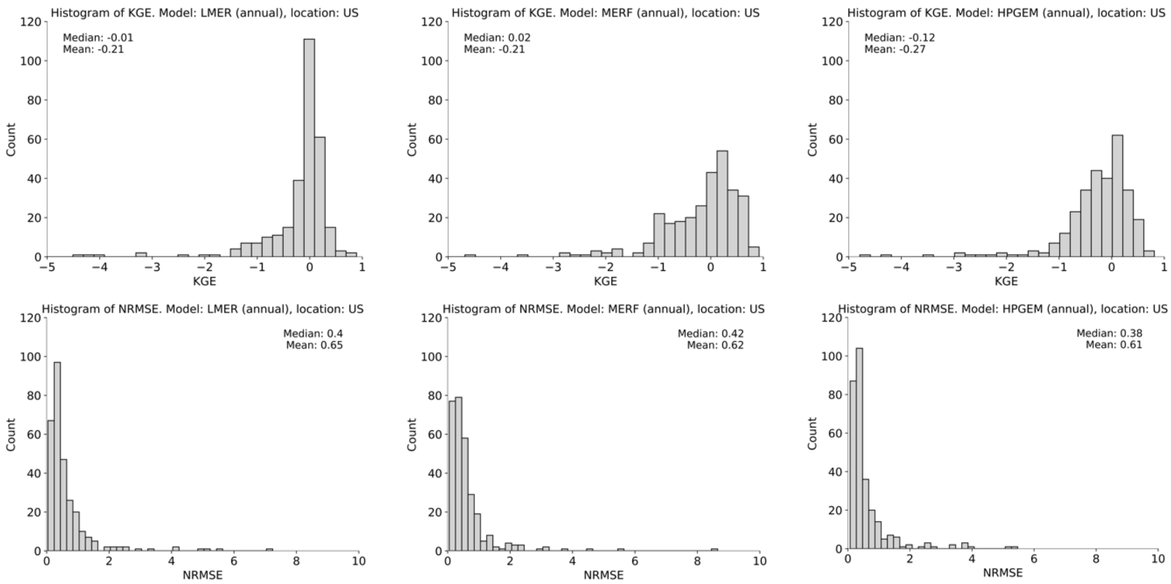
Prediction results evaluated on KGE using the LMER model (left) and MERF model (right). The top row shows the KGE results, while the bottom rows show NRMSE results. The KGE and NRMSE scores are calculated using the whole predicted and measured time series (from GPPD), and are at plant scale. The figures show the histogram of all KGE or NRSME scores for all HPPs in the dataset.



The LMER and MERF model performance was compared to HPGEM, which is the most closely related model. In the US, the models performed similarly, with the MERF model scoring best (KGE 0.02) when evaluated on median KGE, and the HPGEM model when evaluated on median NRMSE (0.38). The comparison was made with data from the US. The results show that no major differences are noticeable (Figure 10). The median KGE score of the HPGEM (-0.12) model was slightly worse than the MERF (0.02) and LMER (-0.01) model, while the HPGEM model scored the best on median NRMSE.

Figure 10.

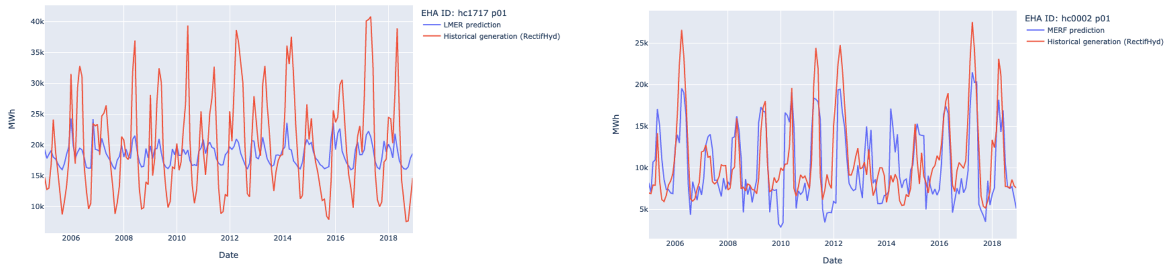
Comparing the LMER, MERF and HPGEM models. Left: LMER, middle MERF, right: HPGEM. Results are for both KGE and NRSME. The KGE and NRMSE scores are calculated using the whole predicted and measured time series (from RectifHyd), and are at plant scale. The figures show the histogram of all KGE or NRSME scores for all HPPs in the dataset.



To gain insight into how accurate the model predictions are for an individual HPP, two HPPs from the US were investigated in more detail (Figure 11). Both predicted time series are median scoring on KGE, on the left for the monthly LMER model, and on the right for the monthly MERF model. The LMER model can capture the trend of the hydropower generation, but not the peaks. The MERF model can reproduce the historical generation relatively accurately. This shows that while KGE scores are relatively similar, the underlying patterns can differ, and it differs between HPPS and between models how accurate predictions are on plant scale. The next chapter will discuss how these results can answer the research questions.

Figure 11.

Time series for two HPPs from the monthly US dataset. Left: prediction using the monthly LMER model, performance was KGE 0.16. On the right, created with the monthly MERF model, performance was KGE 0.18.



4. Discussion

The main research question, which was to evaluate the accuracy of a model predicting monthly hydropower generation at the power plant level in data-scarce regions, will be discussed using above-mentioned results. The LMER and MERF models showed consistent predictive results, with some outliers. The median KGE accuracy in the US was -0.08 for the monthly LMER model and 0.12 for the monthly MERF model, which is in both cases around zero, a commonly used threshold above which predictions are usable. In a data-scarce region, with no data available, this rate of error might be preferable above having no data at all, but it should be noted that model results can produce unreliable predictions.
When comparing to the HPGEM model, both LMER and MERF models performed relatively equal. Yin et al. (2020) mention that their HPGEM model shows large prediction errors, which were comparable to the LMER and MERF models.

When looking at the predictor variables, discharge followed by capacity were found to be the most important. Reservoir area was found to be useful for the monthly model, while this was not the case for the annual model. This confirms the hypothesis that reservoir behavior becomes more important when moving from an annual time step to a monthly time step, which was described by Turner & Voisin (2022).
Dam height was not found to be improving the model, while it was assumed that due to the large influence of hydraulic head on hydropower generation, and therefore dam height as well, it would be an important variable. This was not the case, which could be caused by the fact that dam height was not a good proxy for hydraulic head, or by the fact that hydraulic head is not needed as a predictor variable for a statistical model. Besides dam height, type was assumed to be an important variable due to significant differences between HPPs of different types (reservoir/ROR), but this was found not to be the case for the examined models. This could be explained by the fact that reservoir area already indirectly informs the model on the type. ROR plants will have a small reservoir area compared to reservoir HPPs. The HPGEM model showed similar feature importance results, with runoff and plant capacity as the most important variables (Yin et al., 2020).
The linear (LMER) and nonlinear (MERF) models performed similarly, with no overall better model. The average KGE and NRMSE scores for the monthly MERF model were better than for the monthly LMER model in the US, while the MERF model produced larger outliers. In the EU, the monthly LMER model performed better than the monthly MERF model. This went against the expectation that a nonlinear model would be more accurate than a linear model when predicting hydropower. The analysis by Dabare et al. (2020) showed that a linear regression between rainfall and hydropower showed a positive correlation, while a nonlinear analysis proved to be better fitting. The LMER and MERF models only partly confirm their results, as the MERF model was slightly better performing than the LMER model in the US. The Hydro Plant Generation Estimation Model (HPGEM) uses a tree-based model, and a random forest model was used by Falchetta, Kasamba & Parkinson (2020), showing the potential of a random forest for prediction in data-scarce regions. A reason for the relatively worse performance of the MERF model compared to the expected result might have been that the dataset's quality did not allow the random forest to learn a better model than a linear regression model. The discharge data from GRFR assumed no anthropogenic changes were made to the water flow, which might explain prediction errors.

Yearly model prediction results score slightly better than monthly model predictions, but show that the model can predict on a monthly timescale without losing substantial predictive accuracy.
The model result shows that the model does produce inaccurate predictions in about half of the cases (a median KGE of around zero). Comparing the monthly and yearly LMER and MERF models showed that both models perform equally.

The EU dataset contained less historical generation than the US dataset (1443 data points for the US yearly dataset and 171 for the EU yearly dataset). This limits the possibility of basing conclusions on the EU results.

The results showed that the model could produce substantial errors and some outliers in the EU. The average scores for EU models were lower than the score of the comparable model results in the US.
The predictive results using the EU dataset showed a decrease in model performance between the EU annual MERF and US annual MERF models. This decrease was not seen for the LMER models of the US and the EU. This could mean that the MERF model is overfitting on US data. If this is the case, that would mean that the MERF model is not generalizable to regions outside of the US. This model behavior could be resolved using different parameters or a different type of nonlinear model than a default MERF model. For the LMER model, the US and the EU performance was relatively similar, suggesting that the issue is rather with the MERF model configuration than with the datasets used. The results show that there are differences between KGE and NRMSE, which shows that it is important for these models to include multiple evaluation measures to gain more insight into actual model performance.

The existing HPGEM model produced similar results to the LMER and MERF models at an annual time step. HPGEM scored best when evaluated on NRMSE, while MERF scored the best when evaluated on KGE, indicating the models are similar in performance without a clear best model.
The next section will present limitations related to the performed research.

4.1. Limitations

Hydropower modeling can be challenging due to complex relationships between hydrology and hydropower infrastructure, which is mentioned by Hansen, Ghimire & Kao (2022). The limitations will be presented in the following section, starting with limitations related to general modeling of hydropower infrastructure.
HPPs can be in a cascading system, directly influencing each other's generation. In the models, all power plants are seen as individual units, without influence on downstream HPPs. Furthermore, not all data could be checked, and some hydropower plants may be incorrectly linked to other data such as discharge or reservoir area. There are difficulties in linking hydropower plants and reservoirs due to differences in data sets. For example, of the dams in the US with hydropower facilities in the NID and GRanD, about 25% cannot be linked to a hydropower plant (Hansen, Ghimire, & Gangrade, 2021). About 6% of US dams linked to an existing hydropower plant are not marked as hydropower in the NID or GRanD (Hansen, Ghimire, & Gangrade, 2021). Some datasets disagreed with each other, with different data for the same HPP. This happened for example with capacity data, which was available from multiple sources, in the US from EHA and GPPD, which did show some disagreements. If possible, data that differed above a certain level was removed, as was done with data with a larger than 50% difference in the upstream area from the NID dataset and the MERIT river reach dataset. The upstream area of each river reach is calculated at the downstream point of each river reach, which can be a different location than that of the linked HPP. This means the matching between an HPP and a river reach, which was done based on the similarity between the upstream area of both, might be wrong in some cases. For EU, matching HPPs to river reaches was done to the nearest river reach, which is not always correct (Figure 2). It was not possible to match HPPs to river reaches using a buffer for the EU dataset, as the upstream area which was used for matching an HPP to the river reach in the US was unavailable in the EU.

The discharge taken from the GRFR dataset is calculated at the most downstream point of each river reach, while the HPP might be at any point of the river reach it is matched to. This means the reported discharge will not exactly match the discharge at the location of the HPP unless the HPP is at the convolution point of the river reach. Discharge data (from GRFR) is from a reanalysis dataset, not from real measurements. The GRFR does assume a river flows uninterrupted, which means reservoir characteristics and other anthropogenic changes to discharge are not accounted for. The discharge data must thus be seen as an approximation.

Moving on to the other datasets, there was a low availability of reservoir area data in the US, as many data points in HILARRI have no linked GRanD id, which was needed for linking with the reservoir area dataset. Dam height is used as a proxy for the hydraulic head, due to little available data on the actual hydraulic head. Dam height and actual hydraulic head might differ substantially from each other, as the hydraulic head depends on the reservoir level. While dam height was not included in the final model, the fact that it was excluded could be caused by incorrect data used for model assessment.

The RectifHyd generation dataset is based on EIA-923 data, which is measured at an annual time step, and for some power plants at a monthly time step (Turner, Voisin & Nelson, 2022). The power plants with observations at an annual scale have been imputed following the method described by Turner, Voisin & Nelson (2022). This means that RectifHyd does not solely includes actual measurements. Some RectifHyd generation data was negative, and was removed from the dataset based on a recommendation from the dataset's author. For calculations of the maximal possible generation in a given timeframe, the HPP efficiency was assumed to be 90% in the LMER and MERF models.
The inspection of individual predictions made on the power plant level instead of evaluating a histogram shows that model KGE and NRMSE scores are not always representative of actual performance and that it is necessary to inspect individual predicted time series of the model, next to more general evaluation methods such as KGE and NRMSE (Figure 11). In Figure 11, the left panel shows a median scoring HPP, where the predicted trend is correct, but actual predictions are far removed from the actual generation range. This will be more severe for about half of the power plants since the showed HPP was median scoring, indicating large differences between actual generation and predictions. Therefore, KGE and NRMSE give insight into model accuracy but are not completely informative. KGE scores can be above zero, while the actual predictions are far removed from actual generation.
In the next section, recommendations for future research will be given, based on the presented research.

4.2. Recommendations for future research
A first recommendation is that a future iteration of the monthly model could use the reservoir area as an optional predictor variable, increasing the data available and thus the number of HPPs for which the model can be used. Since reservoir area was the second or third most important variable (out of three), it could be removed without decreasing model performance drastically. The current models depend on a GRanD id, to link an HPP to a reservoir. The reservoir area dataset uses the GRanD id, and dam height data is taken from the GRanD dataset. The GRanD id was often found missing in the EU dataset and was the limiting factor for data availability.

Secondly, creating a new dataset for the US with more HPPs included would be possible, only for prediction. The dataset currently used was constrained by the need for historical generation data for model fitting, but historical generation would not be needed for a dataset used for model prediction. For the presented US results, the US dataset for model fitting and prediction was the same. Creating a larger dataset only for predictions using the same input datasets would be possible. A downside of this approach would be that HPPs without generation data cannot be evaluated, which is why the approach of having two different datasets for model fitting and model prediction was not taken in this research.

Thirdly, the *merf* package allows the implementation of other model types besides a random forest, which could be implemented. The model parameters of the MERF model could be further optimized for increased performance, as the MERF results between the US and EU could suggest the model was overfitted on US data, and performing worse on EU data. The model could furthermore be validated in another region outside of the US and EU, to check if the model can be used worldwide and if results differ from EU results.

Concluding, the new LMER and MERF models can make useful predictions overall (measured by median KGE) on a monthly time step for HPPs in data-scarce regions, with substantial model errors and outliers at the individual plant scale. The annual model requires discharge and capacity as predictor variables, while the monthly model requires reservoir area data next to discharge and capacity. Hydropower prediction in data-scarce regions remains difficult, which is a problem the presented models do not solve. On an individual plant level, errors remain substantial.

When no data on hydropower generation is available, the models can provide a monthly generation prediction usable for on research for energy grid research, on larger spatial scales than the plant scale. On plant scale the model can produce prediction errors of KGE scores up to -2.5 for the US dataset and -13 for the EU. The outliers of the KGE and NRSME distributions show that individual predictions can be far removed from the actual generation. Therefore, the models should not be expected to predict the plant scale accurately, but given that median KGE scores are above or around zero, they are useful on larger scales.

5. Conclusions

The main goal was to assess the accuracy of a predictive model for monthly hydropower prediction at the power plant level in data-scarce regions. A linear mixed-effects model and a mixed-effects random forest model were created for this goal. About half of the predictions were reliable (median KGE around zero), while outliers showed predictions up to a KGE of -2.5 for the US dataset, with most predictions between a KGE of 1 and -1. For the EU dataset, there was an outlier up to a KGE score of -13 on the EU dataset, with most of the predictions between a KGE score of 1 and -2. This leads to the conclusion that hydropower can only for some HPPs be accurately predicted in data-scarce regions. The model results are usable on a larger spatial scale, as median KGE scores are around zero.

Discharge and capacity were found to be the most important predictor variables, while the type and dam height were removed from the final model.

Monthly models scored similar to models at a yearly time step, showing the model setup using monthly discharge and reservoir area is able to predict on a monthly time step. For the US predictions, the MERF model scored better than the LMER model, while for the EU predictions the LMER model scored better. This leads to the conclusion that in the current configuration, the LMER model performs best on independent data, while the MERF model could potentially be improved with different model parameters, as it scored best on the US dataset. The model performance on an independent dataset from the EU showed a decrease in model performance on KGE for the LMER and MERF models. The sample size was small (64 HPPs in the annual EU dataset) due to little generation data available at plant level in the EU. The LMER and MERF models were compared against the HPGEM model, resulting in similar performance between LMER, MERF, and HPGEM. The model produces unreliable results at the power plant level, at a comparable level to existing models like the HPGEM. The models predict at a monthly time step with only a minor decrease in model performance compared to a model at an annual time step. In a data-scarce region, the model could provide electricity grid researchers with a prediction of hydropower generation for HPPs with unknown hydropower generation, with the risk of considerable prediction error. The model cannot be used for accurate predictions at plant scale due to these substantial errors. A purpose of the model could be to not use it at plant scale but on a larger spatial scale, as the median scores of KGE were around zero, which shows that the median predictions over multiple HPPs are usable.

6. Code and data availability

Code is made available at Codeberg[4]. The README file explains how the results can be reproduced.

Datasets containing hydropower predictions for the US and EU created using the models, can be found in the folder *prediction_datasets*

[4] https://codeberg.org/nilsboonstra/hydropower_monthly_prediction

7. References

Barzola-Monteses, J., Mite-León, M., Espinoza-Andaluz, M., Gómez-Romero, J., & Fajardo, W. (2019). Time Series Analysis for Predicting Hydroelectric Power Production: The Ecuador Case. Sustainability, 11(23), 6539. https://doi.org/10.3390/su11236539

Bates, D. M. (2010). lme4: Mixed-effects modeling with R.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

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

Chowdhury, A. F. M., Kern, J., Dang, T. D., & Galelli, S. (2019). PowNet: a power systems analysis model for large-scale water-energy nexus studies.

Coskun, H. G., Alganci, U., Eris, E., Agıralioglu, N., Cigizoglu, H. K., Yilmaz, L., & Toprak, Z. F. (2010). Remote sensing and GIS innovation with hydrologic modeling for hydroelectric power plant (HPP) in poorly gauged basins. Water resources management, 24(14), 3757-3772.

Dabare, G., Gunathilake, M. B., Miguntanna, N., Laksiri, K., & Rathnayake, U. (2020). Climate variation and hydropower generation in samanalawewa hydropower scheme, Sri Lanka. Engineer: Journal of the Institution of Engineers, Sri Lanka, 53(3), 19-25.

Ember. (2023, July 30). Global Electricity Review 2022. Ember. URL: https://ember-climate.org/insights/research/global-electricity-review-2022/#supporting-material (Accessed: August 14, 2023)

Falchetta, G., Kasamba, C., & Parkinson, S. C. (2020). Monitoring hydropower reliability in Malawi with satellite data and machine learning. Environmental Research Letters, 15(1), 014011.

Gao, H., & Zhao, G. (2019). Global Reservoir Surface Area Dataset (GRSAD). https://doi.org/10.18738/T8/DF80WG, Texas Data Repository, V3

Global Energy Observatory, Google, KTH Royal Institute of Technology in Stockholm, Enipedia, World Resources Institute. 2021. Global Power Plant Database. URL: https://datasets.wri.org/dataset/globalpowerplantdatabase (Accessed: August 22, 2023).

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6), 1313-1328.

Hansen, C., Ghimire, G., & Gangrade, S. (2021). Hydropower Energy Storage Capacity Dataset (No. 1). Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).

Hansen, C., Ghimire, G. R., & Kao, S. C. (2022). Evaluation of nominal energy storage at existing hydropower reservoirs in the US. Water Resources Research, 58(11), e2022WR032210.

Hansen & Matson. (2023). Hydropower Infrastructure - LAkes, Reservoirs, and RIvers (HILARRI), V2. HydroSource. Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. DOI: https/doi.org/10.21951/HILARRI/1960141

Hoes, O. A., Meijer, L. J., Van Der Ent, R. J., & Van De Giesen, N. C. (2017). Systematic high-resolution assessment of global hydropower potential. PloS one, 12(2), e0171844.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.

IEA (2021). Hydropower Special Market Report. URL: https://www.iea.org/reports/hydropower-special-market-report. (Accesses: August 15, 2023)

Johnson. (2023). Existing Hydropower Assets (EHA) Capacity Plant Database, 2005-2022. HydroSource. Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. DOI: https://doi.org/10.21951/EHA_Capacity_FY2023/1989316

JRC Hydro-power database. (2019). [Data set]. European Commission, Joint Research Centre. http://data.europa.eu/89h/52b00441-d3e0-44e0-8281-fda86a63546d (Original work published 2019)

Kanellopoulos, K., De Felice, M., Gonzalez, I. H., & Bocin, A. (2019). Jrc open power plants database (jrc-ppdb-open). Zenodo.

Krechowicz, A., Krechowicz, M., & Poczeta, K. (2022). Machine Learning Approaches to Predict Electricity Production from Renewable Energy Sources. Energies, 15(23), 9146.

Lai, J.-P., Chang, Y.-M., Chen, C.-H., & Pai, P.-F. (2020). A Survey of Machine Learning Models in Renewable Energy Predictions. Applied Sciences, 10(17), 5975. https://doi.org/10.3390/app10175975

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 Reviews, 26, 100426.

Le, M. H., Kim, H., Adam, S., Do, H. X., Beling, P., & Lakshmi, V. (2022). Streamflow Estimation in Ungauged Regions using Machine Learning: Quantifying Uncertainties in Geographic Extrapolation. Hydrology and Earth System Sciences Discussions, 1-24.

Lehner, B., Czisch, G., & Vassolo, S. (2005). The impact of global change on the hydropower potential of Europe: a model-based analysis. Energy Policy, 33(7), 839-855.

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15), 2171-2186.

Lehner, B., C. Reidy Liermann, C. Revenga, C. Vorosmarty, B. Fekete, P. Crouzet, P. Doll, M. Endejan, K. Frenken, J. Magome, C. Nilsson, J.C. Robertson, R. Rodel, N. Sindorf, and D. Wisser. 2011. Global Reservoir and Dam Database, Version 1 (GRanDv1): Dams, Revision 01. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4N877QK. Accessed 08-08-2023.

Levasseur, A., Mercier-Blais, S., Prairie, Y. T., Tremblay, A., & Turpin, C. (2021). Improving the accuracy of electricity carbon footprint: Estimation of hydroelectric reservoir greenhouse gas emissions. Renewable and Sustainable Energy Reviews, 136, 110433.

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., et al. (2019). Global reconstruction of naturalized river flows at 2.94 million reaches. Water Resources Research, 55, 6499–6516. https://doi.org/10.1029/2019WR025287

Liu, H., Andresen, G. B., Brown, T., & Greiner, M. (2019). A high-resolution hydro power time-series model for energy systems analysis: Validated with Chinese hydro reservoirs. MethodsX, 6, 1370-1378.

Menard, S. (2001). Collinearity. Applied logistic regression analysis Second Edition (Quantitative applications in the social sciences).

MERF. Mixed-effectss random forest (python), Manifold AI. URL https://github.com/manifoldai/MERF. (Accessed: July 23, 2023).

NID (2023). URL: https://nid.usace.army.mil/#/downloads. (Accessed: July 23, 2023).

R Development Core Team. R (2005) A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

Scikit-learn permutation importance, 2023. Permutation feature importance. URL: https://scikit-learn.org/stable/modules/permutation_importance.html#id1 (Accessed: July 23, 2023).

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29.

Sharma, N., Sharma, P., Irwin, D., & Shenoy, P. (2011, October). Predicting solar generation from weather forecasts using machine learning. In 2011 IEEE international conference on smart grid communications (SmartGridComm) (pp. 528-533). IEEE.

Ssekulima, E. B., Anwar, M. B., Al Hinai, A., & El Moursi, M. S. (2016). Wind speed and solar irradiance forecasting techniques for enhanced renewable energy integration with the grid: a review. IET Renewable Power Generation, 10(7), 885-989.

Stoll, B., Andrade, J., Cohen, S., Brinkman, G., & Brancucci Martinez-Anido, C. (2017). Hydropower modeling challenges (No. NREL/TP-5D00-68231). National Renewable Energy Lab.(NREL), Golden, CO (United States).

Turner, S. W., & Voisin, N. (2022). Simulation of hydropower at subcontinental to global scales: a state-of-the-art review. Environmental Research Letters.

Turner, S. W., Voisin, N., & Nelson, K. (2022). Revised monthly energy generation estimates for 1,500 hydroelectric power plants in the United States. Scientific data, 9(1), 1-11.

Uría-Martínez, R., Johnson, M. M., Shan, R., Samu, N. M., Oladosu, G., Werble, J. M., & Battey, H. (2021). US Hydropower market report (No. DOE/EE-2088). Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States).

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Van Vliet, M. T., Wiberg, D., Leduc, S., & Riahi, K. (2016). Power-generation system vulnerability and adaptation to changes in climate and water resources. Nature Climate Change, 6(4), 375-380.

Wang, J., Walter, B. A., Yao, F., Song, C., Ding, M., Maroof, A. S., ... & Wada, Y. (2022). GeoDAR: georeferenced global dams and reservoirs dataset for bridging attributes and geolocations. Earth System Science Data, 14(4), 1869-1899.

Wei, Z., Shen, X., Qiu, G., Liu, Y., & Liu, J. (2023). An Integrated Transfer Learning Method for Power Generation Prediction of Run-off Small Hydropower in Data-scarce Areas. IEEE Transactions on Smart Grid.

Yang, Y., Pan, M., Lin, P., Beck, H. E., Zeng, Z., Yamazaki, D., ... & Wood, E. F. (2021). Global reach-level 3-hourly river flood reanalysis (1980–2019). Bulletin of the American Meteorological Society, 102(11), E2086-E2105.

Yildiz, C., & Açikgöz, H. (2021). Forecasting diversion type hydropower plant generations using an artificial bee colony based extreme learning machine method. Energy Sources, Part B: Economics, Planning, and Policy, 16(2), 216–234. URL: https://doi.org/10.1080/15567249.2021.1872119

Yin, T., Byers, L., Valeri, L. M., & Friedrich, J. (2020). Estimating Power Plant Generation in the Global Power Plant Database.

Zhang, N., Li, Z., Zou, X., & Quiring, S. M. (2019). Comparison of three short-term load forecast models in Southern California. Energy, 189, 116358.

Zhou, F., Li, L., Zhang, K., Trajcevski, G., Yao, F., Huang, Y., Zhong, T., Wang, J., & Liu, Q. (2020). Forecasting the Evolution of Hydropower Generation. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2861–2870. DOI: https://doi.org/10.1145/3394486.3403337

8. Appendix A: dataset information

Table A4.

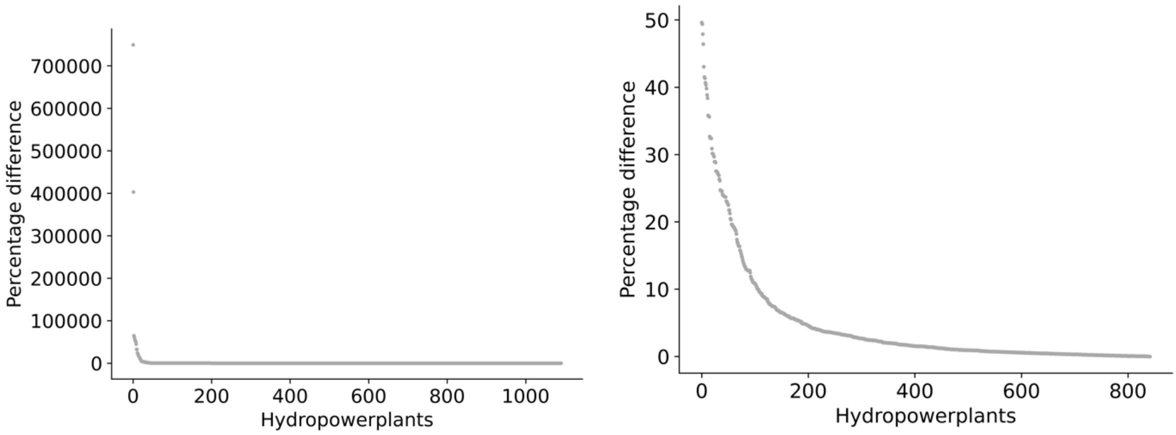
Estimating Power Plant Generation in the Global Power Plant Database (Yin et al., 2020).

	Hydroelectricity plants by region	Hydroelectricity plants by region (%)	Hydroelectricity plants with reported generation data	Hydroelectricity plants with reported generation data (%)
North America	1,474	39.50%	1,364	77.40%
South America	633	16.90%	0	0.00%
Europe	841	22.50%	62	3.50%
Africa	80	2.10%	25	1.40%
Asia	690	18.50%	311	17.70%
Australia/Oceania	18	0.50%	0	0.00%
Total	3,736	100%	1,762	100%

9. Appendix B: dataset summary statistics

Figure B1

Figure showing the percentage difference of upstream area reported by HPPs and their matched river reaches.



Percentage difference between upstream area between NID and MERIT right after matching, showing two extreme outliers.

Percentage difference between upstream area between NID and MERIT after removal of differences larger than 50%.

Figure B2

Histogram of the number of data points for historical generation per HPP, for the USA monthly dataset (left) and yearly dataset (right). The histogram shows that most HPPs have 168 months of historical generation data, and 14 years of historical generation data.

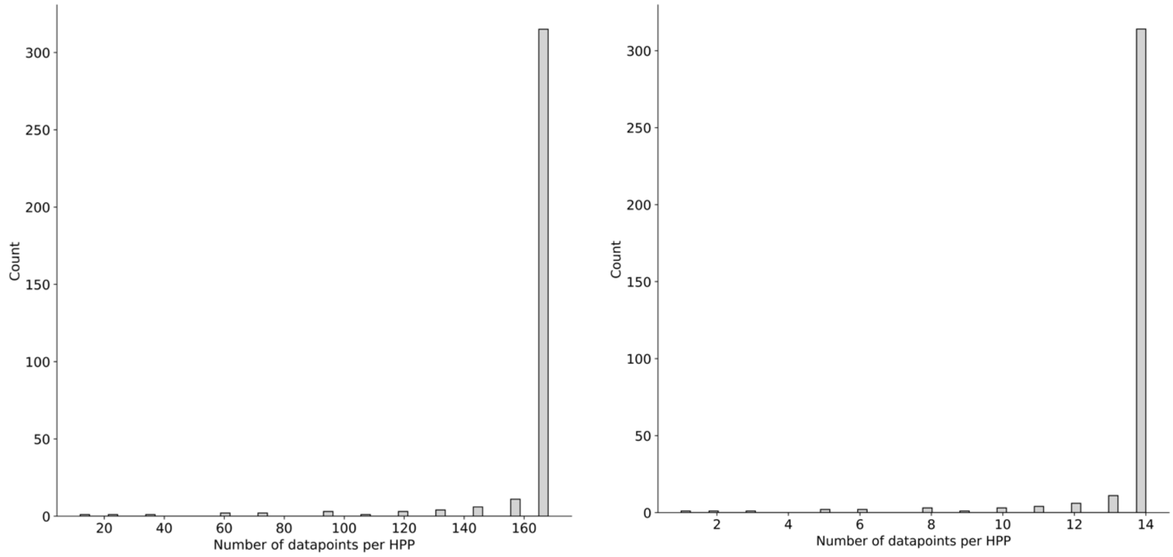


Figure B3.

Locations of all HPPs in US (left and EU (right) monthly datasets.

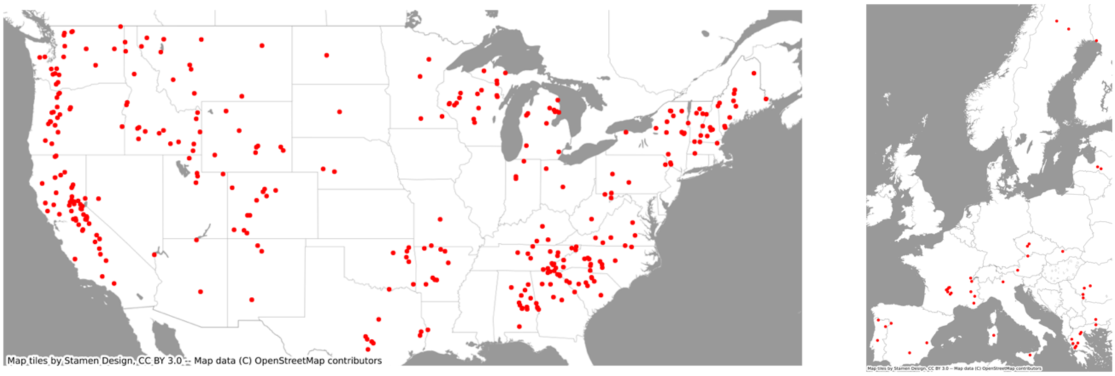


Table B1.

Summary of US yearly dataset, containing 338 HPPs

Variable	Symbol	Unit	Mean	Median	SD[a]	Min[a]	Max[a]	Skew	Source
Hydropower net generation	MW h	MWh	2.28E+05	7.71E+04	4.66E+05	6.00E+00	6.25E+06	4.83	RectifHyd[b]
Discharge	Q	m^3/s	1.00E+02	4.55E+01	1.60E+02	1.95E-02	1.86E+03	3.7	GRFR[c]
Capacity	W	MW	8.08E+01	2.79E+01	1.58E+02	1.00E+00	1.31E+03	4.45	EHA[d]
Hydraulic head	H	meter	4.92E+01	3.80E+01	4.19E+01	0.00E+00	2.23E+02	1.44	NID[e]
Reservoir area	A	km^2	4.22E+01	9.52E+00	1.17E+02	2.44E-01	1.28E+03	6.82	GRSAD[f]

[a]SD = Standard deviation; Min = minimum; Max = maximum

[b]Turner, Voisin & Nelson, 2022

[c] Yang et al., 2021

[d] Johnson, 2021

[e] NID, 2023

Table B2.

Summary of EU yearly dataset, containing 914 (64 with generation data) HPPs.

Variable	Symbol	Unit	Mean	Median	SD[a]	Min[a]	Max[a]	Skew	Source
Hydropower net generation	MW h	MWh	6.69E+02	4.61E+02	5.53E+02	1.05E+02	2.49E+03	1.41	GPPD[b]
Discharge	Q	m^3/s	1.04E+02	1.82E+01	2.87E+02	2.62E-01	2.24E+03	5.00	GRFR[c]
Capacity	W	MW	1.63E+02	9.60E+01	2.19E+02	1.00E+00	1.44E+03	3.09	JRC[d]
Hydraulic head	H	meter	8.93E+01	9.50E+01	3.91E+01	2.60E+01	1.68E+02	0.05	JRC[d]
Reservoir area	A	km^2	1.96E+01	1.18E+01	2.23E+01	1.20E+00	8.00E+01	1.42	GRSAD[e]

[a]SD = Standard deviation; Min = minimum; Max = maximum

[b] Yin et al., 2020

[c] Yang et al., 2021

[d] Hildalgo-Gonzalez et al., 2019; JRC Hydro-power database. (2019).

[e] Gao et al., 2019

10. Appendix C: Results

Table C1.

Model AIC performance for incrementally added variables.

Model variables	AIC (monthly)	AIC (annual)
Discharge	-13,672.57	-4,378.69
Discharge + Capacity	-13,734.64	**-4,469.51**
Discharge + Capacity + Reservoir area	**-13,854.12**	-4,451.33
Discharge + Capacity + Reservoir area + Dam Height	-13849.68	-4,464.09
Discharge + Capacity + Reservoir area + Dam Height + Type	-13844.15	-4,437.75

Table C2.

Yearly model performance summary.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	3.44E-01	9.52E-03	3.51E+02	3.61E+01	2.92E-120
Discharge	7.18E-04	3.94E-05	2.94E+03	1.82E+01	1.44E-70
Capacity	-6.40E-04	5.87E-05	4.89E+02	-1.09E+01	6.14E-25

Table C3.

Monthly model performance summary.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	3.53E-01	1.02E-02	3.16E+02	3.47E+01	6.08E-110
Discharge	3.68E-04	5.67E-06	5.68E+04	6.50E+01	2.00E-16
Capacity	-6.42E-04	5.61E-05	4.75E+02	-1.14E+01	5.83E-27
Reservoir Area	6.25E-04	5.14E-05	2.78E+03	1.22E+01	3.34E-33