Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 1659–1668

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Machine Learning for Hydropower Scheduling: State of the Art and Future Research Directions

Chiara Bordin [1,2*], Hans Ivar Skjelbred [2], Jiehong Kong [2], Zhirong Yang [3]

[1] *UiT, The Arctic University of Norway, Dept of Computer Science, Tromsø, (Norway)*
[2] *SINTEF Energy Research, Dept of Energy Systems, Trondheim, (Norway)*
[3] *Norwegian University of Science and Technology, Dept of Computer Science, Trondheim (Norway)*

Abstract

This paper investigates and discusses the current and future role of machine learning (ML) within the hydropower sector. An overview of the main applications of ML in the field of hydropower operations is presented to show the most common topics that have been addressed in the scientific literature in the last years. The objective is to provide recommendations for novel research directions that can be taken in the near future to cover those areas that have not been studied so far. The key contribution of this paper lies in a critical investigation of the state of the art of ML applications in hydropower scheduling. In light of the established literature available in the last years, this study identifies and discusses new roles that can be covered by ML, coupled with cyber-physical systems (CPSs), with a particular focus on short-term hydropower scheduling (STHS) challenges.

Keywords: Artificial Neural Networks; Digitalisation; Hydropower Scheduling; Machine Learning

1 Introduction

The last years have seen a wide digital transformation in many areas of the energy field, where new concepts, methods, and models are receiving more attention, not only in the scientific communities but also within the industrial sector. Among these concepts, machine learning (ML), cyber-physical systems (CPSs), internet of things (IoT), and big data analytics are covering an important role due to a massive availability of data to be exploited. Digitalisation and massive data availability open the doors to novel ways of addressing many of the current energy domain challenges. They provide instruments and digital approaches that will improve or even drastically change the currently established methodologies for analyses, simulations, and optimisation in the energy field. Meanwhile, the global fleet of hydropower assets is adapting to a new era of digitalised systems and processes from design and construction to operation and maintenance. Smart digital control systems can improve the performance of hydro units and extend their lifetimes. Operation and maintenance can be optimised, and costs can be reduced using advanced performance monitoring analytics. However, a significant challenge is how to efficiently collect and analyse data to make full use of, and derive benefit from, the information in the data. ML techniques can be adopted to identify functional relationships between variables by analysing large amounts of possibly disparate big data (streaming measurements, batch data from measurement campaigns, metadata, etc.) and extracting favourable information.

This paper aims at investigating the role of ML within the hydropower sector. The objective is to provide an overview of the main applications of ML in the field of hydropower operations and to define potential research directions that can be taken in the near future to cover those areas that have not been studied so far. The key contribution of this paper lies in a critical investigation of the state of the art of ML applications in hydropower scheduling. In light of the established literature available

* Corresponding author: Chiara Bordin
E-mail address: chiara.bordin@uit.no

10.1016/j.procs.2020.09.190

in the last years, the objective is to identify and discuss new roles that can be covered by ML, coupled with CPSs, with a particular focus on short-term hydropower scheduling (STHS) challenges.

The remainder of this paper is organised as follows. Section 2 will give a brief introduction to hydropower scheduling problems. Section 3 will discuss the state of the art of ML applications in the hydropower sector. Section 4 will present the current challenges in STHS. Section 5 will illustrate future research directions by discussing how ML techniques, coupled with CPS approaches, can contribute to the STHS models in terms of the improvement in input data quality and the innovation in operating methods. Finally, section 6 will draw the conclusions.

2 Notes on hydropower scheduling

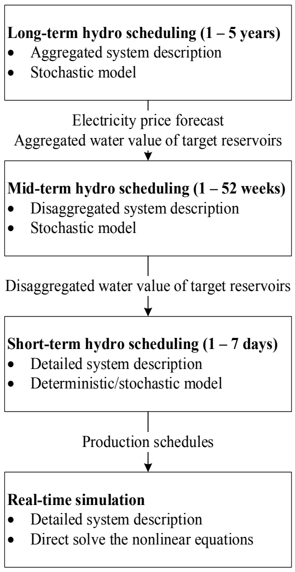

Figure 1: Decomposition of the hydropower scheduling problem

Hydropower plants can broadly be classified as reservoir-based and run-of-river type [1]. In a cascaded watercourse, there is a combination of storage reservoirs and run-of-river plants [2]. Each reservoir is either connected to a gate or associated to a hydropower plant composed of different or identical hydro units. The hydro units can be generating or pumping units. The main source of inflow for hydropower plants comes from precipitation, streams, and melting snow. According to the usual terminology adopted in literature, this paper will identify "streamflow" forecasting as a way to predict long-term monthly hydraulic measures that include precipitation, evaporation, temperature, and the like, while the term "inflow" forecasting will identify the predictions within short-term hydropower scheduling that include the forecasted natural inflow, i.e., precipitations and melting snow to the reservoirs.

Since the stored water in reservoirs can be used in later periods, the decision to generate energy for the present or save water for future use is coupled in time. In the meantime, since the released outflow of a plant will affect the power production of all the plants downstream, the operation of the cascaded hydraulic objects (i.e., reservoirs, gates, plants, hydro units) is coupled in space. The complex spatial-temporal coupling makes the hydro scheduling problem more difficult to solve than the thermal scheduling counterpart [3]. As a competitive source of renewable energy, optimal generation scheduling of hydro units has an important role in the electricity markets. Hydropower scheduling aims at generating maximum energy by utilising the available water potential. Depending on the characteristics of the power system, data availability, and computational resources, different methods are candidates for deciding the optimal hydropower scheduling policy. However, it is not possible to include all the details of the system into one single large-scale optimisation model.

Normally the hydropower scheduling problem is decomposed into different scheduling levels extending from aggregated long-term, disaggregated mid-term, detailed short-term, to real-time simulation (Figure 1). Each problem is modelled by the appropriate mathematical formulation and solved by dedicated solution techniques [4]. The method used in long-term scheduling demands an aggregation of the hydro system. On the other hand, short-term optimisation requires detailed information. These two requirements are incompatible, and, therefore, a mid-term scheduling process is needed to establish a link between long-term planning and short-term optimisation. A nonlinear generation scheduling simulator is used to verify the short-term optimisation results and to modify the final operational plan.

Uncertainty has to be considered in long-term and mid-term hydro scheduling because it cannot be assumed that the input parameters for electricity prices and natural inflow to reservoirs will be known for the entire planning horizon that may stretch over the years. For STHS models covering a relatively short period of time, it is reasonable to assume that the prices and inflow are deterministic. However, the short-term variability of prices increases due to intermittent renewables to the power system, and the variability of inflow also increases due to climate change. Many new methods have been proposed in the modelling of the STHS problem to give robust schedules in the face of these variations of input parameters [5].

3 ML in hydropower - state of the art

ML is the field of study that gives computers the ability to learn without being explicitly programmed [6]. A computer program is said to learn from experience *E* with respect to some task *T* and some performance measure *P*, if its performance on *T*, as measured by *P*, improves with experience *E* [7]. As shown in Table 1, different ML methodologies have been applied to the field of hydropower scheduling, mainly for the purpose of dataset forecasting. Among these, the most widely used techniques found in the literature are Linear Regression (LR), Support Vector Machine (SVM), Support Vector Regression (SVR), Clustering, Fuzzy clustering, Artificial Neural Networks (ANNs). Specifically, LR, SVM, and SVR fall under the umbrella of the so-called "supervised learning", while Clustering and Fuzzy clustering fall under the umbrella of the so-called "unsupervised learning". In addition, ANN is part of the more advanced "deep learning" approach.

Supervised learning refers to an ML approach in which both input variables and output variables are available, and an algorithm is used to learn how to map functions from the inputs to the outputs. The objective of the learning algorithm is to get an approximation of the mapping function that is good enough to allow predictions of output variables when new input data arise. It is called supervised learning because it uses a training dataset to train the algorithm. The algorithm is then able to treat a new dataset according to the training experience.

Table 1. An overview of ML applications in the hydropower sector

ML Method	Objectives	Input	Time Scale	Ref
LR	Annual streamflow trend assessment	Annual flow for 32 years	Annual	[8]
LR	Hydropower generation projections	Historic hydropower generation, precipitation, and runoff	Annual	[9]
LR	Hydropower cascade, simulation, control and optimisation	Inflow data with 15 minutes time resolution, turbine discharge data with 1-minute time resolution. A 5-day period is considered	Minute	[10]
SVM	Streamflow forecasting	Streamflow data for 31 years	Monthly	[11]
SVM	Flow forecasting	Daily flow, precipitation, evaporation, streamflow data for 25 years	Daily	[12]
SVM	Streamflow forecasting	Streamflow data for 100 years	Monthly	[13]
SVM	Streamflow forecasting	Streamflow data for 45 years	Monthly	[14]
SVM	Streamflow forecasting	Streamflow records and rainfall records for 48 years	Monthly	[15]
SVM	Streamflow time series forecasting	Streamflow historical data for 133 years	Monthly	[16]
SVR	Hydropower consumption forecasting	Hydropower consumption data for 24 years	Seasonal	[17]
SVR	Streamflow forecasting	Streamflow time series for 40 years	Monthly	[18]
Clustering	Hydro energy inflow forecasting	Monthly streamflow from 54 hydropower reservoirs; climate data	Seasonal	[19]
Clustering	Forecast of 3038 inflow scenarios	Historical weather realisations	Daily	[20]
Clustering	Streamflow forecasting	Periodic hydrologic series	Monthly	[21]
Fuzzy Cluster	Streamflow forecasting	68 historical streamflow datasets	Monthly	[22]
ANN	Optimal production patterns	Price and inflow	Daily	[23]
ANN	Reservoir operation scheme	Inflow time series	Daily	[24]
ANN	River streamflow forecasting,	Daily precipitations, daily min and max temperatures	Daily	[25]
ANN	Reservoir runoff forecasting	Daily runoff data for 7 years	Daily	[26]
ANN	Reservoir inflow	Rainfall	Monthly	[27]
ANN	Reservoir inflow	Precipitation, evaporation, temperature, inflow for 30 years	Monthly	[28]
ANN	Reservoir Inflow	Tropical rainfall, inflow data	Daily	[29]
ANN	Reservoir storage	Inflow, turbine release, reservoir storage, evaporation losses for 29 years	Monthly	[30]

The LR establishes a relationship between the dependent variable and one or more independent input variables using a best fit straight line (also known as a regression line). While simple LR has only one input variable, multivariate LR has multiple input variables. Examples of LR applications in the hydropower field can be found in [8], [9], and [10]. In [8], the authors propose a new approach for trend assessment that takes into account the long-term periodicity of annual flows. The objective is to assess the current and future water availability for hydropower generation by using LR together with a novel linear moving window approach. The observed annual flow for 32 years is used as input. The authors conclude that the proposed approach represents a more reliable technique for long-term flow predictions. The LR regression is also used in [9] to project changes in annual and regional hydropower generation in multiple power marketing areas. For this purpose, the authors conduct a series of regression analyses based on the 1989-2008 annual time series of temperature, precipitation, runoff, and hydropower generation. The study shows that the change of annual runoff results in a proportional change of regional hydropower generation, and hence confirms the need for more detailed studies on hydropower operation and climate change. Another application of LR can be found in [10], where the authors aim at simulating a hydropower cascade using historical data of inflow.

Flow and streamflow forecasting is also addressed in the literature through SVM by looking for suitable ways to improve the performance of such an established technique. The SVM is an ML technique used in classification problems. A classification problem requires the prediction of a category or a class Y from some inputs X. The main idea behind the SVM algorithm is to plot each data as a point within an n-dimensional space (n is the number of features), where the value of each feature is the value of a particular coordinate. Classification is then performed by finding the hyper-plane that differentiates the two classes in the best possible way. The best hyperplane is the one whose distance between the nearest data point is maximised.

In [11], a modified SVM framework is proposed to improve the predictability of the inflow to a reservoir in two months, using climate data from the prior period. A combination of SVM and genetic algorithms is proposed to improve the uncertainty assessment and the overall performance of the prediction processes. A combination of SVM and genetic algorithms is also illustrated in [12], where the authors apply a genetic algorithm to optimise the parameters of SVM for daily flow forecasting. Four model structures with different input vectors are developed and discussed to show the improved prediction accuracy of the proposed SVM model. An improved SVM model is proposed in [13] as well. In this work, the authors aim at developing a methodology to improve the performance of SVM in predicting monthly streamflow by introducing an adaptive insensitive factor. A case study is proposed to test and show the feasibility of the new model. A modified SVM model for monthly streamflow forecasting is investigated in [14]. In particular, the authors show that combining empirical mode decomposition and SVM can provide a superior alternative to the basic SVM technique used as stand-alone. A combination of SVM with different methods of time series decomposition can be found in [15], where the authors aim at a better estimation of the streamflow. The results show that models, coupled with decomposition techniques, perform better than the single stand-alone models. Monthly streamflow time series prediction is performed with SVM in [16], where authors aim at overcoming the variety of frequency components that natural runoff often contains. The proposed method is compared with traditional methods, such as ANN to show the improved forecasting accuracy.

The SVM concepts can be generalised to become applicable to regression problems [31]. The method is, therefore, called Support Vector Regression SVR. Even though SVR is less popular than SVM, it has been proven to be an effective tool in real-value function estimation. As discussed in [31], one of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Moreover, SVR has excellent generalisation capability, with high prediction accuracy. An example of an application to the field of hydropower can be found in [17], where SVR is used instead of SVM. The objective of the study is hydropower consumption forecasting, and results show that the proposed approach is promising for complex time series forecasting with seasonality. A hybrid SVR framework is also proposed in [18] for the streamflow forecast. The proposed method is compared with a stochastic autoregressive integrated moving average (ARIMA) streamflow forecast model. The average error of the proposed model is reduced to less than one-tenth in contrast to the state-of-the-art method.

Beyond supervised learning, also the so-called "unsupervised learning" has been used within the hydropower sector. As opposed to "supervised learning", the "unsupervised learning" refers to an ML approach in which there are only input data and no corresponding output variables. The goal of unsupervised learning is to learn more about the data and to model the underlying structure in the data. Therefore, in unsupervised learning, there is not a training phase as the algorithm is supposed to discover and present the structure that lies behind the data. Clustering falls under the unsupervised learning umbrella and aims at discovering the inherent groupings in the data. The *K*-means clustering is the simplest form of unsupervised learning. It identifies *k* number of centroids, and then allocates every data point to the nearest cluster.

Examples of applications of such techniques to the hydropower sector can be found in [19], [20], and [21]. A clustering methodology is used in [19] for streamflow forecasting. Given a large number of interconnected hydropower reservoirs, the authors propose the alternative to consider the concept of large clusters of hydropower reservoirs. Therefore, they optimise the equivalent energy of each cluster and finally optimise the individual energy production of the reservoirs within each cluster. A stochastic short-term hydropower planning with inflow scenario trees is proposed in [20]. In this case, the authors use the clustering approach to partition data into clusters and assign initial values to the scenario tree nodes. Clustering techniques are used in [21] and applied in the monthly streamflow generation model developed for the hydroelectric system. The objective is to alleviate the computation effort in the mid-term operation planning model.

Fuzzy clustering (also referred to as soft clustering) is a form of clustering in which each data point belongs to more than one cluster. An example of such an approach can be found in [22], where the authors suggest a fuzzy prediction model based on fuzzy clustering as an alternative for the streamflow forecasting. The method uses the fuzzy *c*-means clustering technique to group data patterns and fuzzy clustering to classify prediction patterns.

ANNs, especially deep neural networks, have widely been used in these years. An ANN usually comprises a number of neurons organised in multiple layers. Each neuron implements a linear transformation followed by a nonlinear activation, where the output of neurons (except the last layer) becomes the input of neurons in the next layer. With many neurons and layers, deep neural networks are capable of representing very complex transformation function from input to output [32]. Given sufficient input-output training pairs, ANN can be well fitted and identify the relationship between input and output variables with respect to the data distribution. Such a data-driven modelling method tends to provide more objective and more accurate performance compared with conventional domain-specific modelling in numerous applications.

ANN has also been employed in hydropower scheduling. For a single-unit hydropower plant, ANN has been used to map the input data for price and inflow directly to optimal production patterns [23]. To perform the mapping, the ANN needs to obtain the optimal production pattern from an STHS optimisation tool. The method applied in [23] does not guarantee to find the global optimum. An ANN-based general reservoir operation scheme is presented in [24]. The operation scheme can be added to daily hydrologic routing models for simulating the releases from dams, in regional and global-scale studies. The ANN technique is also used in [25] for modelling rainfall-runoff and overcome the challenges related to the non-linearities of such models. The objective is to predict river streamflow and provide water resource management with a tool that can support them in operating the reservoir properly, especially in the case of extreme events such as flooding and drought. A daily reservoir runoff forecasting method using ANN is also presented in [26] in combination with particle swarm optimisation. The latter is employed to select the ANN optimal parameters, and the ANN is then used for the prediction after the training process. Results show that combining ANN with particle swarm optimisation was improving the accuracy of the forecast. A dynamic neural network approach is used in [27] for monthly reservoir inflow forecasting. Results showed the suitability of the approach but also the necessities for further improvements by fitting the model for de-seasonalised series as the inflow series exhibits monthly seasonality. The reservoir inflow modelling is studied in [28] as well, by using ANN and hydrometeorological parameters. The ANN model results reveal that there is a positive relationship between the actual and forecasted reservoir inflow with a fairly high value of correlation coefficient for all the selected locations in the studied area. This shows that the model is appropriate for reservoir inflow modelling. A novel wavelet-artificial neural network hybrid model (WA-ANN) for short-term daily inflow forecasting is proposed in [29]. The study shows how the wavelet transformation, coupled with ANN, can improve the forecasting results. Finally, in [30], an ANN-based model for forecasting reservoir storage for hydropower dam operation is successfully applied. Results show the ability of ANN to perform well for such objectives and their versatility in reservoir management modelling.

Table 2 shows an overview of papers aimed at comparing different ML techniques for hydropower applications. In [33], a comparison between ANN and multivariate LR is presented to estimate rainfall and its impact on hydropower generation. The model test results indicate that the ANN produces more accurate results compared to LR, which can be attributed to the fact that ANN performs tasks that a linear program is unable to do.

Table 2. An overview of papers aimed at comparing different ML techniques for hydropower applications

AI Methods Compared	**Objectives**	**Input**	**Time Scale**	**Ref**
Multivariate LR, ANN	Rainfall estimation	Rainfall	Monthly	[33]
ARIMA, ANN, ANFIS, Genetic programming, SVM	Discharge time series	River flow discharging	Monthly	[34]
ANN, SVR, Multivariate R	Streamflow forecasting	Weather and climate inputs	Daily	[35]
ANN, Fuzzy clustering	Streamflow forecasting	Streamflow time series	Monthly	[36]
ANN, SVM	Streamflow forecasting	Streamflow data for 38 years	Monthly	[37]
ANN, SVM, ANFIS	Inflow forecasting	Inflow, precipitation, humidity, temperature	Daily	[38]

A comparison of several artificial intelligence methods for forecasting monthly discharge time series is presented in [34]. In this study, the authors focus on ARIMA, ANN, Adaptive Neural-based Fuzzy Inference System (ANFIS), Genetic programming, and SVM. The results obtained in this study indicate that artificial intelligence methods are powerful tools to model the discharge time series and can give good prediction performance than traditional time series approaches. The results indicate that the best performance can be obtained by Adaptive Neural-based Fuzzy Inference System, genetic programming, and SVM in terms of different evaluation criteria during the training and validation phases. In [35], three ML methods are compared for daily streamflow forecasting, ANN, SVR, and multivariate LR. The study shows that, in terms of forecast scores, the nonlinear models generally outperform multivariate LR, and ANN tends to slightly outperform the other nonlinear models. Two ML techniques are addressed in [36] for streamflow forecasting, i.e., ANN and fuzzy clustering. The results show a generally better performance of the ANN for the case studied. Finally, in [37], the performances of SVM and ANN models are compared in predicting monthly streamflow. According to the results, SVM models for all different input combinations provide better prediction results in comparison of the ANN models for monthly streamflow prediction. Three different data-driven models for reservoir inflow prediction are compared in [38], with a particular focus on ANN, SVM, and ANFIS. The predictive accuracy of the data-driven models is then discussed. The study reveals that there is no significant difference among the data-driven models, but more attention should be paid to the choice of the data-driven model in winter and summer for getting a more accurate inflow forecast.

Looking at the overview provided by Table 1 and Table 2, it is possible to note that even though a variety of ML methodologies has been adopted, there is a very narrow selection of objectives and choices of applications. In fact, it seems that the scientific community mainly focuses just on inflow and streamflow forecasting (together with weather and climate inputs), with very few digressions such as those proposed by [9] and [17]. However, in the latter two, the objective is not on data generation for hydropower scheduling, but on broader topics of hydropower generation and consumption projections.

Other applications of ML in the hydropower field that can be found in the literature are related to the use of ML for the dam's drought estimation. However, these are not part of the overview proposed in Table 1, as they lie on the board between hydropower and hydrology, and they are not of interest for the purpose of hydropower scheduling that represents the focus of this paper. It is also worth mentioning that price forecasting and load forecasting play an important role in renewable energy generation in general, and hydropower scheduling in particular. A review of ML methods applied to electricity price forecasting can be found in [39], [40], and [41], while an overview of works dedicated to load forecasting is available in [42], [43], and [44]. However, price and load forecasting are a very wide subject that covers all the energy sectors, from smart grids to renewables, to buildings, and many others. Hence a thorough review of the ML techniques for price and load forecasting would be outside the scope of this paper that wants to focus specifically on ML application for hydropower scheduling.

In conclusion, the literature review shows that ML in hydropower is mainly used for inflow forecasting with very rare variations from this main topic. However, many other challenges arise in STHS that are worthy of being investigated. This will be further discussed in the following section in order to better identify and discuss other roles that can be covered by ML and new potential applications of such techniques within the field of STHS.

4 Challenges in short-term hydropower scheduling

As mentioned in Section 2, STHS considers complex watercourses and technical details of the production system together with various strategic, regulatory, and market constraints. The scheduling horizon is normally from one day to two weeks with a 15-minute or hourly time resolution. The former practice in STHS was to optimally determine the water release of reservoirs and to attain the power generation schedules of the available hydro resources by minimising the operating cost. After the deregulation and market competition was introduced in many countries in the 1990s, STHS was also employed to support spot bids in the day-ahead market and to provide final schedules after the market clearing process [45].

Nowadays, with the rapid development of wind and solar technology, non-dispatchable renewable energy sources play a notable part in the power production mix of many countries. Because of its storability, flexibility, and controllability, hydropower is of critical importance in ensuring system safety. The application of STHS to the power system integrated with non-dispatchable renewable energy brings about new business opportunities as well as operational challenges. Participation in both energy and capacity markets highlights the need for the precise calculation for energy conversion and available capacity of each unit. A large amount of production from wind and solar power that varies within a short period also necessitates a quick response in real-time intraday and balancing markets [46].

Mathematically, the STHS problem is formulated as a large-scale, discrete, nonlinear, and non-convex problem. A wide range of optimisation techniques has been proposed to address this complex problem. In [46] the authors point out that compared to the widespread plant-based STHS problem formulations where the hydro-turbine generator units in a hydropower plant are aggregated as one equivalent unit, it is more critical to model each unit separately in order reflect the complexity of real-world daily operations and to match the requirements from the hydro producers. It demands a more accurate and detailed representation of the hydropower generation, considering the impact of head variation, hydraulic losses, efficiency curves, and restricted operational zones on the power produced by each unit.

On the other hand, the quality of the output of a decision support tool is not only a matter of good mathematical modelling but also a matter of the quality and precision of the input datasets that are provided into the model. They play a significant role in representing the STHS problem, where many technical configurations of the hydraulic system, together with electricity price and inflow forecasting mentioned in Section 3, should be taken into account. These input parameters affect the final operational decisions. When the input parameters provided into the model are approximated or inaccurately estimated, the quality of the decision-making process becomes inferior due to the gap existing between the parameter estimation and the actual measurements.

For instance, turbine efficiency is a measure of the relation between the mechanical energy produced and the potential energy of the water discharged. The efficiency of the turbine is usually represented as a function of the net head and water discharge from the generating unit, usually known as the "hill curve". It is usually provided by the turbine manufacturer and obtained by data extrapolation from a reduced scale model. Turbine manufacturers use empirical relations to correct the model efficiency to better reflect the prototype efficiency. Despite this correction, the real performance of the turbines can be affected by the plant's constructive and operational characteristics not considered by the manufacturer. Besides, changes in turbine efficiency happen with deterioration, damage, or problems with the equipment over time [47]. If the turbine efficiency is still formulated based on the original data, ignoring the altered conditions from the initial design stage, inconsistency will occur between the mathematical function and practical observations.

Except for turbine efficiency, other input parameters such as generator efficiency, head loss in shared penstocks, and tailraces loss are also based on common knowledge or theoretical data. Nowadays, hydropower producers are experiencing a technological revolution that provides an unprecedented number of data, either through the expansion of existing equipment or the construction of new plants. Digitalisation is being integrated into the modernisation programs for existing hydropower assets. The increasing amount of data from on-site measurements creates the possibility to investigate the real correlations between production, water discharge, losses, effective water head, and efficiency. It will lead to more physically correct input parameters and instructions to the STHS models.

Authors in [48] take the first attempt to provide a more realistic mathematical model by estimating all the essential parameters in the pumped hydro system. While in [49], the ANN methodology is used to improve the identification of the optimal operation curve under conditions of the limited amount of training data. The focus here is the identification of the optimal turbine rotational speed as a function of a flow rate. Moreover, experiences of ML applications for fault detection and maintenance of hydropower plants can be found in [50] and [51]. These works are all aimed at addressing the challenge of parameter definition, which is becoming more and more relevant in recent years. Recent developments in literature are already pointing out that existing models within current literature, produce a high error in calculating stored energy since some critical parameters are ignored.

To sum up, the significant challenges in STHS can be summarised as follows.

1) How to find mathematically precise and computationally solvable problem formulations for the unit-based STHS that can achieve the balance between theoretical perfection and operational feasibility;
2) How to efficiently acquire, measure, analyse, and interpret data in order to make full use of, and derive benefit from, the information in the data. Parametric ML techniques may be used to appropriately parameterise and fit these input parameters;
3) How to continuously update the input parameters quality and problem formulation by using reinforcement learning-type methods to close the mismatch between output results of the optimisation model and the actual, and measured, performance.

5 Recommendations for future research directions

In the previous section, the challenges related to the quality of the technical input dataset for the STHS models were identified and discussed. Given such premises, three main research questions arise for the scientific community:

- Is it possible to enhance the performance of hydropower scheduling models by improving the quality of technical input parameters that are currently approximated (i.e., turbine efficiency, head losses, and the like)?
- Which technical input parameters mostly affect the solutions and are worthy of being better estimated?
- Can ML, coupled with CPSs, meet the need for better technical parameter estimation for hydropower scheduling, and which ML techniques will perform better for this purpose?

CPSs [52] are systems that interconnect and integrate a physical space or process with computational software to perform tasks that require a mix of cyber and physical components. The physical processes are monitored and controlled by embedded computers and networks, usually with feedback loops where physical processes affect computations and vice versa [53]. CPSs are going to play an important role in the context of hydropower. In fact, nowadays, many monitoring systems are installed in hydropower plants, through which it is possible to collect technical information from various components. There is an increasing availability of data coming from the physical hydropower space through sensors, devices, and smart meters. This massive data available must be manipulated and utilised within a cyber space in order to support and enhance the various decision-making processes and system operations. From this point of view, hydropower systems integrate the physical space (the hydropower plant) and the cyber space (sensors, ICT, and advanced technologies), and thus exhibit characteristics typical of CPS. Within the hydropower sector, the CPS is good for energy efficiency, energy resource management, and energy monitoring and control, thus making the systems "Cyber-Physical Energy Systems" (CPESs) [54].

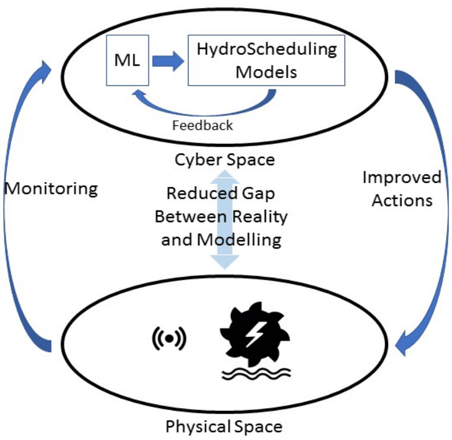

Figure 2: A Cyber-Physical System for technical parameters estimation in STHS

The combination of ML and CPES can enhance the hydropower field, especially for monitoring related tasks. Indeed, the information coming from sensors and devices installed in the hydropower physical space can be used to train ML algorithms and perform better estimations for many parameters that are going to be part of the datasets of an optimisation model. The cyber space would, therefore, be represented by the ML algorithms, which would be deeply interconnected with the physical hydropower space through the sensors systems. They interact through retroactive approaches by constantly improving the parameter estimation, reducing the gap between prediction and reality, and enhancing the output of hydropower scheduling optimisation models. Figure 2 illustrates the concept. The physical space containing hydropower plants and monitoring devices is connected to the cyber-space that includes the short-term models for hydropower and ML algorithms. Thus, the connection between physical space and cyber-space defines a CPES. Through the monitoring devices, a large dataset related to the status of technical parameters within the hydropower plant will be available. Therefore, the monitoring task is the key task that, departing from the physical space, will bring data into the cyber-space. In this way, ML algorithms will be able to create more precise input datasets to feed the STHS models. The optimisation models will then provide optimal decisions in terms of hydropower scheduling and will thus be connected again with the physical space through improved actions. The real events and technical parameters values observed within the physical space will be continuously compared with the forecast parameters used to feed the short-term hydropower models. The gap between reality and modelling can, therefore, be reduced by continuously providing feedback to the ML algorithms in a retroactive approach aimed at improving the estimation of the technical parameters.

In addition, predictive maintenance becomes increasingly appealing with the advent of digital sensing technologies coupled with ML. Condition monitoring and diagnostics can detect component failures or deterioration of equipment [55]. Prognosis is used to predict future development and when failures will occur. Therefore, maintenance can be scheduled based on the observed condition and prediction. Such approaches significantly reduce unnecessary interruptions of production and increase maintenance effectiveness. Pioneering projects [56] and case studies [57], have been conducted to develop models and algorithms for condition monitoring and generation of early faults warnings. In addition, recent preliminary studies aimed at comparing machine learning techniques in prognostic maintenance of hydropower plants have been proposed [58]. Therefore, it is of strategic interest to establish an intelligent system that combines STHS and maintenance optimisation utilising new digital solutions and technologies, such as big data analytics, data mining, ML, CPS, and advanced optimisation algorithms. The combination of safe operation and predictive maintenance will have a positive impact in the form of reduced operational costs, increased reliability of power supply, and enhanced asset management.

At present, no matter how sophisticated the optimisation tools are, the hydropower operators must manually set up the executive commands before running the optimisation models. Usually, the choice of commands is based on the operators' personal experience or the model developers' general suggestion. Though various solution methods or heuristics have been developed and are free to use [59], limited by human analytic competence, the operators prefer to choose the commands they are familiar with or directly adopt the default setting. These commands are only set up once before optimisation and valid for all the hydraulic objects and the entire scheduling horizon. The manual and static setup of commands delimits the power of the optimisation tools. The full value of the optimisation tools (e.g., increased profit or reduced computational time) could be exploited if the commands are dynamically determined according to the particular operating and market conditions of the hydro systems. However, the optimal selection of commands for the optimisation model is impossible to be done by hand due to thousands of complex constraints and millions of coupling variables. To address this problem, ML can be employed to efficiently and automatically select the optimal commands for a given hydro system and the end user's preference. This type of research will pioneer a brand-new decision-making process for hydro scheduling. It broadens the traditional comprehension of optimisation from the general model formulation to the tailored operational options.

One of the challenges in ML is interpretability. Although the prediction made by complex ML systems is often more accurate than conventional domain-specific approaches, it is hard to interpret or explain why the ML systems make such predictions [60]. To answer the "why" question, one needs to identify the major causal input variables. As mentioned in Section 4, STHS is a large-scale complex problem with numerous interconnections. If ML is applied as a black box solution, as in [23], whether the result predicted by the machine can be acceptable is indistinct. This could be a promising research direction in the future.

Last but not least, as to the ML techniques, ensemble learning (EL) is a branch of ML and has been successfully used in many recent ML applications [61], [62]. EL combines a number of base ML models, where the base models can be obtained by using, for example, different training datasets or different ML algorithms. In decision making, the combination can be done by, for example, bagging that applies majority votes of decisions from base models, and boosting which involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models misclassified. In estimation theory, EL can effectively average out the estimation bias (e.g., by bagging) and reduce the estimation variance (e.g., by boosting) when there is significant diversity among the base models. These can bring improvement over the base models in terms of generalization performance, which is the key indicator of machine learning. EL improves ML results by combining several models. This approach brings a better predictive performance compared to a single model. Just like ML, EL can be applied to the hydropower CPES, to successfully extract knowledge from the data gained from the hydropower physical space through sensors, devices, and smart meters. Surprisingly, to the best of our knowledge, EL has not been applied in hydropower scheduling yet, which could be another future research direction.

6 Conclusions

Many hydro producers are in the process of installing new sensors in their hydropower stations, collecting and storing SCADA (Supervisory Control And Data Acquisition, a system for control of the power plant) and sensor data, and making the data available for new types of analyses. The rapid development of ML techniques combined with advances in optimisation algorithms creates new possibilities for increasing profitability for hydropower producers. From the research point of view, successful applications of ML in STHS will make a closer connection between the practical operations and problem formulation, and prepare for autonomy in the hydropower industry by defining the need for and availability of autonomous systems today, and into the future. In this paper, a review of the state of the art of ML applications for the hydropower sector has been presented. The review shows that even though a variety of ML methodologies have been applied to hydro data, the investigated objectives and applications only represent a narrow selection of opportunities. The scientific community has mainly focused on inflow and streamflow forecasting. Additional challenges in the hydropower field have been discussed, and the importance of technical inputs parameter estimation has been highlighted. Recommendations for future research directions have been drawn based on the current scientific literature, the challenges in hydropower modelling, and a large amount of data that is available through monitoring systems. In particular, more effort should be put in investigating the opportunity of implementing ML techniques, coupled with CPSs, to estimate technical parameters for a better performance of hydropower scheduling models. In addition, EL has been identified as a promising methodology to apply to the hydropower scheduling, within the CPES, given that it has not been used for such purposes yet.

7 Acknowledgments

This paper is partially supported by an Innovation Project for the Industrial Sector granted by the Research Council of Norway (Grant award no. 309936).

References

[1] V. K. Singh and S. K. Singal, “Operation of hydro power plants-a review,” *Renew. Sustain. Energy Rev.*, vol. 69, pp. 610–619, 2017.
[2] M. T. L. Barros, F. T. C. Tsai, S. Yang, J. E. G. Lopes, and W. W. G. Yeh, “Optimization of large-scale hydropower system operations,” *J. Water Resour. Plan. Manag.*, vol. 129, no. 3, pp. 178–188, 2003.
[3] E. C. Finardi and E. L. da Silva, “Solving the hydro unit commitment problem via dual decomposition and sequential quadratic programming,” *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 835–844, 2006.
[4] M. M. Belsnes, *Optimal Utilization of the Norwegian Hydropower System*. Fakultet for informasjonsteknologi, matematikk og elektroteknikk, 2008.
[5] E. K. Aasgård, “Stochastic programming for short-term hydropower scheduling and bidding,” 2019.
[6] A. Munoz, “Machine Learning and Optimization,” *Courant Inst. Math. Sci.*, pp. 1–2, 2014.
[7] T. M. Mitchell, “Key Ideas in Machine Learning,” *Mach. Learn.*, pp. 1–11, 2017.
[8] M. Stojković, S. Kostić, S. Prohaska, J. Plavšić, and V. Tripković, “A New Approach for Trend Assessment of Annual Streamflows: a Case Study of Hydropower Plants in Serbia,” *Water Resour. Manag.*, vol. 31, no. 4, pp. 1089–1103, 2017.
[9] S. C. Kao *et al.*, “Projecting changes in annual hydropower generation using regional runoff data: An assessment of the United States federal hydropower plants,” *Energy*, vol. 80, pp. 239–250, 2015.
[10] A. Hamann, “Simulation of a Hydropower Cascade Using Historical Data and State Estimation,” pp. 1–19.
[11] P. H. Li, H. H. Kwon, L. Sun, U. Lall, and J. J. Kao, “A modified support vector machine based prediction model on streamflow at the Shihmen Reservoir, Taiwan,” *Int. J. Climatol.*, vol. 30, no. 8, pp. 1256–1268, 2010.
[12] J. Wang, L. Liu, and J. Chen, “Combination of Genetic Algorithm and Support Vector Machine for Daily Flow Forecasting,” *2008 Fourth Int. Conf. Nat. Comput.*, pp. 31–35, 2008.
[13] J. Guo, J. Zhou, H. Qin, Q. Zou, and Q. Li, “Monthly streamflow forecasting based on improved support vector machine model,” *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13073–13081, 2011.
[14] S. Huang, J. Chang, Q. Huang, and Y. Chen, “Monthly streamflow prediction using modified EMD-based support vector machine,” *J. Hydrol.*, vol. 511, pp. 764–775, 2014.
[15] S. Zhu, J. Zhou, L. Ye, and C. Meng, “Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River, China,” *Environ. Earth Sci.*, vol. 75, no. 6, pp. 1–12, 2016.
[16] Z. kai Feng *et al.*, “Monthly runoff time series prediction by variational mode decomposition and support vector machine based on quantum-behaved particle swarm optimization,” *J. Hydrol.*, vol. 583, no. October 2019, p. 124627, 2020.
[17] S. Wang, L. Yu, L. Tang, and S. Wang, “A novel seasonal decomposition based least squares support vector regression ensemble learning approach for hydropower consumption forecasting in China,” *Energy*, vol. 36, no. 11, pp. 6542–6554, 2011.
[18] X. Luo, X. Yuan, S. Zhu, Z. Xu, L. Meng, and J. Peng, “A hybrid support vector regression framework for streamflow forecast,” *J. Hydrol.*, vol. 568, no. November 2018, pp. 184–193, 2019.
[19] C. H. R. Lima and U. Lall, “Climate informed long term seasonal forecasts of hydroenergy inflow for the Brazilian hydropower system,” *J. Hydrol.*, vol. 381, no. 1–2, pp. 65–75, 2010.
[20] S. Séguin, S. E. Fleten, P. Côté, A. Pichler, and C. Audet, “Stochastic short-term hydropower planning with inflow scenario trees,” *Eur. J. Oper. Res.*, vol. 259, no. 3, pp. 1156–1168, 2017.
[21] D. L. D. D. Jardim, M. E. P. MacEira, and D. M. Falc??o, “Stochastic streamflow model for hydroelectric systems using clustering techniques,” *2001 IEEE Porto Power Tech Proc.*, vol. 3, pp. 244–249, 2001.
[22] M. H. Magalhães, R. Ballini, R. Gonçalves, and F. Gomide, “Predictive fuzzy clustering model for natural streamflow forecasting,” *IEEE Int. Conf. Fuzzy Syst.*, vol. 3, pp. 1349–1354, 2004.
[23] B. V. Matheussen, O.-C. Granmo, and J. Sharma, “Hydropower Optimization Using Deep Learning,” in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2019, pp. 110–122.
[24] N. Ehsani, B. M. Fekete, C. J. Vörösmarty, and Z. D. Tessler, “A neural network based general reservoir operation scheme,” *Stoch. Environ. Res. Risk Assess.*, vol. 30, no. 4, pp. 1151–1166, 2016.
[25] A. Tayebiyan, T. A. Mohammad, and A. H. Ghazali, “SCIENCE & TECHNOLOGY Artificial Neural Network for Modelling Rainfall-Runoff,” vol. 24, no. 2, pp. 319–330, 2016.
[26] C. Cheng, W. Niu, Z. Feng, J. Shen, and K. Chau, “Daily Reservoir Runoff Forecasting Method Using Artificial Neural Network Based on Quantum-behaved Particle Swarm Optimization,” *Water*, vol. 7, no. 8, pp. 4232–4246, 2015.

[27] W. M. N. Dilini and D. Attygalle, "Ensemble Forecast for monthly Reservoir Inflow ; A Dynamic Neural Network Approach."
[28] A. W. Salami, A. A. Mohammed, J. A. Adeyemo, and O. K. Olanlokun, "Modeling Reservoir Inflow for Hydropower Dams Using Artificial Neural Network," *Niger. J. Technol.*, vol. 34, no. 1, pp. 28–36, 2015.
[29] C. A. G. Santos, P. K. M. M. Freire, R. M. da Silva, and S. A. Akrami, "Hybrid wavelet neural network approach for daily inflow forecasting using Tropical Rainfall Measuring Mission data," *J. Hydrol. Eng.*, vol. 24, no. 2, 2019.
[30] T. S. Abdulkadir, A. W. Salami, F. B. Sule, and J. A. Adeyemo, "Neural Network Based Model for Forecasting Reservoir Storage for Hydropower Dam Operation," *Int. J. Eng. Res. Gen. Sci. Vol.*, vol. 3, no. 5, pp. 639–647, 2015.
[31] M. Awad and R. Khanna, "Support vector regression," in *Efficient Learning Machines*, Springer, 2015, pp. 67–80.
[32] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. S. Dickstein, "On the expressive power of deep neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 2847–2854.
[33] T. Soares Dos Santos, D. Mendes, and R. Rodrigues Torres, "Artificial neural networks and multiple linear regression model using principal components to estimate rainfall over South America," *Nonlinear Process. Geophys.*, vol. 23, no. 1, pp. 13–20, 2016.
[34] W. C. Wang, K. W. Chau, C. T. Cheng, and L. Qiu, "A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series," *J. Hydrol.*, vol. 374, no. 3–4, pp. 294–306, 2009.
[35] K. Rasouli, W. W. Hsieh, and A. J. Cannon, "Daily streamflow forecasting by machine learning methods with weather and climate inputs," *J. Hydrol.*, vol. 414–415, pp. 284–293, 2012.
[36] I. Luna, S. Soares, M. H. Magalh??es, and R. Ballini, "Streamflow forecasting using neural networks and fuzzy clustering techniques," *Proc. Int. Jt. Conf. Neural Networks*, vol. 4, pp. 2631–2636, 2005.
[37] R. Muhammad, X. Yuan, O. Kisi, and Y. Yuan, "Streamflow Forecasting Using Artificial Neural Network and Support Vector Machine Models," pp. 286–294.
[38] X. Zhang, H. Wang, A. Peng, W. Wang, B. Li, and X. Huang, "Quantifying the Uncertainties in Data-Driven Models for Reservoir Inflow Prediction," *Water Resour. Manag.*, 2020.
[39] R. Weron, "Electricity price forecasting: review of state-of-the-art with a look into the future," *Int. J. Forecast.*, vol. 30, no. 4, pp. 1030–1081, 2014.
[40] F. Ziel and R. Steinert, "Probabilistic mid- and long-term electricity price forecasting," *Renew. Sustain. Energy Rev.*, vol. 94, no. March 2017, pp. 251–266, 2018.
[41] J. Nowotarski and R. Weron, "Recent advances in electricity price forecasting: A review of probabilistic forecasting," *Renew. Sustain. Energy Rev.*, vol. 81, no. March 2017, pp. 1548–1568, 2018.
[42] A. Almalaq and G. Edwards, "A review of deep learning methods applied on load forecasting," *Proc. - 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2017*, vol. 2018-Janua, pp. 511–516, 2018.
[43] A. Baliyan, K. Gaurav, and S. Kumar Mishra, "A review of short term load forecasting using artificial neural network models," *Procedia Comput. Sci.*, vol. 48, no. C, pp. 121–125, 2015.
[44] M. Q. Raza and A. Khosravi, "A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings," *Renew. Sustain. Energy Rev.*, vol. 50, pp. 1352–1372, 2015.
[45] O. B. Fosso, A. Gjelsvik, A. Haugstad, B. Mo, and I. Wangensteen, "Generation scheduling in a deregulated system. The Norwegian case," *IEEE Trans. power Syst.*, vol. 14, no. 1, pp. 75–81, 1999.
[46] J. Kong, H. I. Skjelbred, and O. B. Fosso, "An overview on formulations and optimization methods for the unit-based short-term hydro scheduling problem," *Electr. Power Syst. Res.*, vol. 178, p. 106027, 2020.
[47] M. K. P. Ã and R. P. Saini, "A review on silt erosion in hydro turbines," vol. 12, pp. 1974–1987, 2008.
[48] N. Mousavi, G. Kothapalli, D. Habibi, M. Khiadani, and C. K. Das, "An improved mathematical model for a pumped hydro storage system considering electrical, mechanical, and hydraulic losses," *Appl. Energy*, vol. 247, no. March, pp. 228–236, 2019.
[49] D. Borkowski, "Identification of the optimal control characteristics of a small hydropower plant using artificial neural networks and the support vector machines method," *J. Hydraul. Res.*, vol. 1686, 2018.
[50] T. M. Welte, J. Foros, M. H. Nielsen, and M. Adsten, "MonitorX - Experience from a Norwegian-Swedish research project on industry 4.0 and digitalization applied to fault detection and maintenance of hydropower plants," 2018.
[51] X. Zheng, Z. Gui, and Y. Wang, "Support vector machine model based on glowworm swarm optimization in application of vibrant fault diagnosis for hydro-turbine generating unit," *Proc. 2017 IEEE 3rd Inf. Technol. Mechatronics Eng. Conf. ITOEC 2017*, vol. 2017-Janua, pp. 238–241, 2017.
[52] R. L. Hartung, A. Hakansson, and E. Moradian, "A prescription for cyber physical systems," *Procedia Comput. Sci.*, vol. 60, no. 1, pp. 1552–1558, 2015.
[53] E. A. Lee, "Cyber physical systems: Design challenges," *Proc. - 11th IEEE Symp. Object/Component/Service-Oriented Real-Time Distrib. Comput. ISORC 2008*, pp. 363–369, 2008.
[54] T. B. Rasmussen, G. Yang, A. H. Nielsen, and Z. Dong, "A review of cyber-physical energy system security assessment," *2017 IEEE Manchester PowerTech, Powertech 2017*, 2017.
[55] L. Selak, P. Butala, and A. Sluga, "Condition monitoring and fault diagnostics for hydropower plants," *Comput. Ind.*, vol. 65, no. 6, pp. 924–936, 2014.
[56] T. M. Welte, J. Foros, M. H. Nielsen, and M. Adsten, "MonitorX--Experience from a Norwegian-Swedish research project on industry 4.0 and digitalization applied to fault detection and maintenance of hydropower plants," *Hydro 2018-Progress through partnership, Gdansk, Poland, 15-17 Oct. 2018*, 2018.
[57] A. Åsnes, A. Willersrud, F. Kretz, and L. Imsland, "Predictive maintenance and life cycle estimation for hydro power plants with real-time analytics," *Predict. Maint. life cycle Estim. hydro power plants with real-time Anal.*, 2018.
[58] N. Khan, L. Hasan, Y. Afridi, and A. Rashid, "Comparison of Machine Learning Techniques in Prognostic Maintenance of Hydro Power Plant Subsystem," vol. 17, no. 2, pp. 31–37, 2019.
[59] H. I. Skjelbred, J. Kong, and O. B. Fosso, "Dynamic incorporation of nonlinearity into MILP formulation for short-term hydro scheduling," *Int. J. Electr. Power Energy Syst.*, vol. 116, p. 105530, 2020.
[60] C. Molnar, *Interpretable machine learning*. Lulu. com, 2019.
[61] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[62] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in neural information processing systems*, 2018, pp. 6638–6648.