Modeling the climate dependency of the run-of-river based hydro power generation using machine learning techniques: an application to French, Portuguese and Spanish cases

Valentina Sessa, Edi Assoumou, Mireille Bossy

Modeling the climate dependency of the run-of-river based hydro power generation using machine learning techniques: an application to French, Portuguese and Spanish cases.

Valentina Sessa[1], Edi Assoumou[1], and Mireille Bossy[2]

[1]MINES ParisTech, Centre de Mathématiques Appliquées (CMA), Sophia Antipolis, France.
[2]Université Côte d'Azur, Inria, France.

Abstract

A big challenge of sustainable power systems is to integrate climate variability into the operational and long term planning processes. In this paper, we focus on the run-of-river based hydro power generation. In particular, we deal with the modeling of this form of power production based on weather variables. Translating time series of meteorological data (precipitations, snowfall and air temperature) into time series of run-of-river based hydro power generation is not an easy task as it is necessary to capture the complex relationship between the availability of water and the generation of electricity. Indeed, this kind of hydro power generation is limited by the flow of the river in which the power plants are located. Moreover, the water flow is a nonlinear function of the weather variables and the physical characteristics of the river basins. Finally, the impact of the weather variables on the runoff may occur with a certain delay, whose determination depends on physically based phenomena (e.g., melting snow–local temperature).

This work aims at formalizing an efficient technique for the prediction of the run-of-river based hydro power generation. Several well-established regression algorithms based on machine learning are used and compared in terms of correlation coefficient, adjusted coefficient of determination, mean absolute and mean square percentage errors. We consider three case studies: France, Portugal and Spain. Results indicate that the models based on ensemble of trees and neural networks exhibit the best performance for evaluating both the short term and the long term hydro power generation.

1 Introduction

Hydro power is the world's most dominant (86%) source of renewable electrical energy. Installed hydro power capacity continues to grow quickly with the aim at decreasing carbon-based or nuclear power generation. This is in line with the

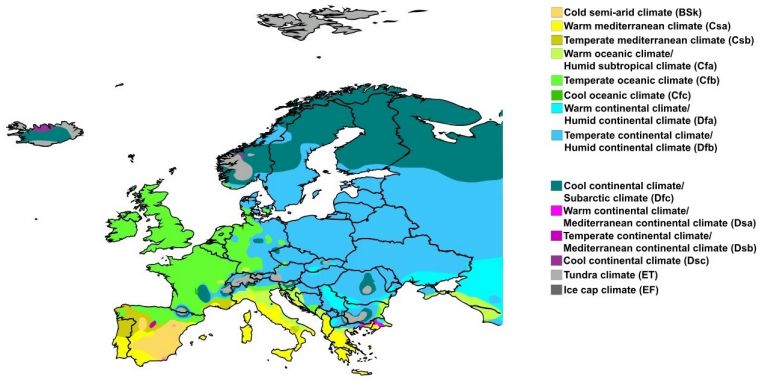


Figure 1: Europe map of Köppen climate classification.

objective of the European Community strategy, which called for a mandatory target of a 20% share of renewable energies by 2020.

Hydro power is either produced in run-of-river plants with low hydraulic heads or from water stored in accumulation lakes with hydraulic heads up to several hundred meters, possibly with recirculation of water between lower and higher level reservoirs in so-called pump-storage systems. Among these existing technologies, we focus on the run-of-river based one, which is the most affected by meteorology. This exists alongside rivers and does not contain large reservoir to store or regulate the flow of the adjacent river. Typically, it generates electricity according to the water flow. This latter is defined by seasonal patterns of precipitations, evaporation, drainage, and other characteristics, which all depend on the geography and weather peculiarity of a location [1]. Although the seasonal patterns of wet and dry seasons are relatively predictable, they are not guaranteed and can change from one year to another [2]. An assessment of climate change impacts on hydroelectric generation in different climate regions requires an in-depth analysis of individual case studies. Given the dominance of local conditions, generalizations are difficult, sometimes even for small regions. Another difficulty is the determination of the temporal relation between the hydro power generation and meteorological variables. In fact, the impact of the weather variables on the water flow, and on the corresponding power production, may occur with a certain delay, whose determination depends on physically based phenomena. For instance, the melting process of snow at high altitude requires a certain amount of time which depends on the local air temperature. Therefore, the increment of the water flow due to the snow fallen during the winter period may occur only after many months with an increase of the temperature. Due to climate changes, such delay is not easy to be predicted.

In this paper, we consider the challenging problem of predicting the daily total national run-of-river hydro power generation based on the impact of weather variables such as precipitation, snow fall and air temperature of some climate regions. We consider climate regions based on the Europe map of Köppen climate classification [3] shown in Figure 1 and we analyze three case studies: France, Portugal, and Spain. This goal will be addressed by using well-established machine learning (ML) techniques. ML has been gaining more and more importance in many areas of science, finance and industry [4]. Typically it is used

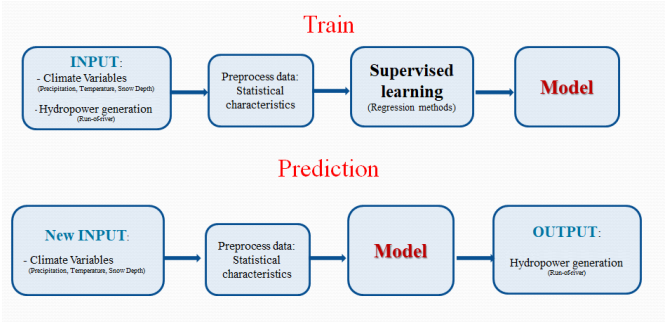
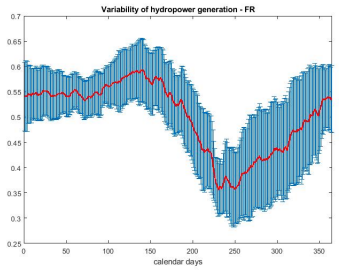


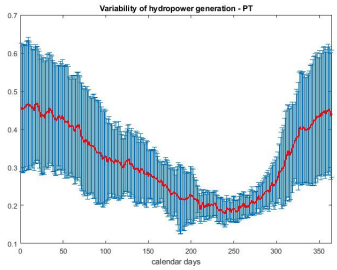
Figure 2: Machine learning workflow.

to predict an outcome based on a set of features. Clearly, in the case of the present paper, the outcome is the run-of-river hydro power generation and the features are the climate variables. The workflow of ML procedure is given in Figure 2. The procedure starts by training a so-called (supervised) learner with a set of data including the observed outcome and feature measurements. This leads to build a prediction model, which enables predicting the unobserved outcome based on a different set of input features. A good learner is one that accurately predicts such an outcome. In the statistical literature, the features are often called the predictors or the inputs, whereas the outcomes are called the responses or the outputs. Along this paper, we will make use of all these terms. Based on an evaluative metric such as the correlation coefficient, the adjusted coefficient of determination, the mean absolute and mean square percentage errors, we will apply and compare five ML algorithms with the aim at determining a model of highest accuracy.

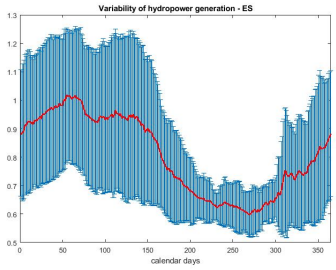
It has been shown that machine learning methods and neural networks are well-suited to the domain of wind speed and wind power prediction [5] and also for solar radiation and solar production [6]. On the other hand, ML techniques have been applied for the run-off forecast, see [7] and references therein, but at the best of our knowledge few attention has been dedicated in the literature to the prediction of run-of-river based hydro power generation from meteorological data. The reason for this lack could be due to the fact that, while the spatio-temporal relation between wind speed-wind power generation (solar radiation-solar power) is local [8], the one between weather variables and river run-off and hydro power generation is way more complex, as we mentioned above. It also interesting to look at the percentages of hydro, wind and solar electric power production generated over the total one in 2017 in the three countries under consideration. In France, we have $10.1\%, 4.5\%,$ and 1.7%, respectively, (it was $12\%, 3.9\%,$ and 1.6% in 2016), in Portugal it is $12.8\%, 20.6\%,$ and 1.7%, respectively, (it was $28.1\%, 20.7\%,$ and 1.4% in 2016), finally, in Spain we have $7.5\%, 19\%,$ and 3.2% ($14.5\%, 19\%,$ and 3.1% in 2016). As highlighted by these data, differently from wind and solar production, there exists a big variability of hydro power generation from a year to another. This behavior, which is mainly due to climate changes, makes the prediction very challenging, but decisive for the optimal power planning [9, 10]. To give a better insight of this phenomena, in Figure 3 we plot the standard deviation around the calendar mean of the observed capacity factor of the hydro power generation (to be



(a) France



(b) Portugal



(c) Spain

Figure 3: Calendar mean and standard deviation of the observed capacity factor computed over 22 years (1982-2004).

defined later) over 22 years.

This work is carried on within the CLIM2POWER project [11], whose overall goal is to provide improved guidance to power systems' stakeholders by combining high resolution weather variables and enhanced energy system model. The predicted values of hydro power production computed in this paper as well as the variability analysis of the prediction error will be used as input for stochastic versions of energy system models (TIMES [12]) and unit commitment models for understanding the impact of the climate variability on the energy systems.

The paper is organized as follows. In Sections 2 and 3, we present the data collection and the ML algorithms used in this paper. Details about the design of the performed experiments and the evaluation criteria are given in Sections 4 and 5, respectively. We dedicate Section 6 to the presentation of the main results. Section 7 concludes this paper providing final remarks and future research ideas.

2 Data collection

Meteorological data include the time series of precipitations, air temperature and snow depth. The historical meteorological dataset is extracted by the ECEM project [13] and is based on the ERA-Interim Reanalysis. The precipitation bias adjustment is carried out by calibrating the parameters of the gamma distribution of ERA-Interim based on the E-OBS (http://eca.knmi.nl/download/ensembles/ensembles.php) dataset (version 12.0, from 1979 to 2016) of grid-

ded observed precipitation. Air temperature bias-adjusted datasets is measured about $2\mathrm{m}$ height from the surface. Snow depth is provided as the average over a relevant time period. This variable is calculated by ERA-Interim and is not bias adjusted. Datasets are on a standard $0.5°$ latitude/longitude grid. The meteorological data at daily level are aggregated considering the same clustering procedure used in the ECEM demonstrator [13].

The historical data for hydro power generation are again from the ECEM project. In particular in this paper, we consider the capacity factor, which is defined as the unitless ratio between the actual power and the installed capacity, assumed to be unchanged in the time period considered in this document. Note that the lack of hydro power generation historical data is a serious issue. Only starting from 1 January 2015, energy demand and generation data were systematically collected at hourly time resolution for almost all countries in Europe, see [14]. This period is however very short to build models aiming at reproducing climate variability effects. Therefore, we rely on the huge effort made by the ECEM project team, for gathering, cleaning and homogenizing different datasets and we use those hydro energy data for our work.

3 ML Algorithms

We use five regression methods: Linear Regressor (LR) [4], Support Vector Machine (SVM) [15], Boosted Ensemble of Trees (BT) [16], Random Forests (RF) [17] and Artificial Neural Networks (ANN) [4]. The first four regression methods are implemented in the Statistics and Machine Learning Toolbox 11.4 [18], while the ANN is in the Deep Learning Toolbox 12.0 [19] in Matlab R2018b. In the following, we give a few details of the algorithms cited above.

The most simple algorithm is the linear regression, which consists of finding the best-fitting straight line through the points of input and response variables. The best-fitting line is called a regression line. The most common type of linear regression is obtained by minimizing a loss function which is the squared error between the observed values and the linear combination of the inputs.

In SMV regression, the goal is to find a function that has at most ϵ deviation from the target points for all the training data and at the same time is as flat as possible. SVM regression uses a type of loss function called 'insensitive' which was proposed by Vapnik [20]. This function defines a ϵ-tube so that if the predicted value is within the tube the loss is zero, while if the predicted point is outside the tube, the loss function is the magnitude of the difference between the predicted value and the radius ϵ of the tube. The optimization problem derived in [20] is solved by considering its dual formulation. Nonlinearities are then added to the SVM algorithm by mapping the training patterns onto a high-dimensional feature space using some fixed nonlinear functions (kernels). It is well known that SVM regression performance (estimation accuracy) depends on a good setting of hyper parameters, which are the regularization constant used in the definition of the objective function, the width ϵ of the insensitive zone and the kernel parameters.

The RF algorithm is based on an ensemble of decision trees. Random vectors are used for growing each tree in the ensemble. A tree is grown by considering a random selection of training set. Then each tree depends on the values of a random vector sampled independently and with the same distribution for all trees

in the forest. In [17], it has been proved that a significant accuracy improvement is gained when randomness is introduced for the parameter selection in ensemble of trees.

The BT algorithm is also based on ensemble of decision trees. The difference is that the predictors are not trained independently but in an iterative manner: every training instance gets a weight assigned that is adapted in every iteration. In every iteration, a new predictor is added to the ensemble and afterwards the prediction quality of the ensemble is tested on the training set instances.

The ANN model simulates the characteristics of the human neural network to deal with distributed parallel information. A classical ANN architecture consists of input, hidden, and output layers with node activation functions. The activation function used in the Matlab toolbox is the sigmoid function in the hidden layer and a linear function in the output layer. Careful attention must be put on the building of the model, as too complex ANN will easily overfit the training data. The most used technique for estimating the ANN model's parameter is the Levenberg-Marquardt learning method.

The tuning of the hyper parameters for all the algorithms is implemented by using the optimization procedure offered by the Matlab toolboxes and the trial and error approach.

3.1 Hybrid method

Recently, it has been shown that the ensemble of machine learning techniques may improve the prediction accuracy [6, 21]. The idea goes as follows: one first uses several ML algorithms to obtain the predicted response, then some combination of the algorithms' outputs is built. In this paper, we simply apply a weighted linear combination of the two best methods for each country. Similarly to [21], we consider the weights derived from each model's mean absolute percentage error (MAPE), to be defined in Section 5, over the validation set and we consider the output of the hybrid method to be

$$\hat{y}_{hyb} = w_1 \hat{y}_{M1} + w_2 \hat{y}_{M2},$$

where $w_i = \frac{MAPE_i^{-1}}{MAPE_1^{-1}+MAPE_2^{-1}}$, $i = 1, 2$ and $\hat{y}_{M1}$ and $\hat{y}_{M2}$ being the outputs of the two ML algorithms with the best accuracy. This means that $\hat{y}_{hyb}$ is obtained by giving more importance, that is a bigger weight, to the algorithm's output with a smaller MAPE.

4 Experiment design

The experiments aim at formalizing an ML model of highest accuracy for the prediction of the capacity factor of the run-of-river-based hydro power generation at daily level.

The first step in the ML workflow is the training phase. Let us indicate with $T_{\text{train}} = \{t_1, ..., t_j, ..., t_N\}$ a given daily spaced time interval, where t_1 and t_N are respectively the initial and final date in the ISO8601 format 'YYYY-MM-DD' and t_j is the j-th day of this interval. We assume that over the training period we collect the data corresponding to

- Month, Day in T_{train}
- Air temperature, i.e., the time series $AT = [AT_{t_1}, ..., AT_{t_N}]$
- Precipitation, i.e., the time series $P = [P_{t_1}, ..., P_{t_N}]$
- Snow depth, i.e., the time series $SD = [SD_{t_1}, ..., SD_{t_N}]$
- Capacity factor of hydro power generation, i.e., the time series $y = [y_{t_1}, ..., y_{t_N}]$.

Note that the time series of air temperature, precipitations and snow depth correspond to the regions in [13]. For each one of the three case studies, we selected different regions' data. For instance, in the case of France, we considered time series corresponding to the fourteen national regions, plus one German region and the two Swiss ones. This choice was suggested by the fact that those neighbor regions have a similar climate, as shown in Figure 1. Moreover, we have taken into consideration the location of the run-of-river based hydro power plants and the corresponding basins [22]. Similarly, by crossing the climate map with the information in [23], in the case of Portugal we collect data relative to the two national regions and four neighbor Spanish ones. Finally, for Spain, besides the eleven national regions, we also select part of France and the two Portuguese regions, see also [24].

As we explained above, the effects of the weather data on hydro power generation occur with a certain delay. In order to counting that, we enrich the list of inputs by considering that the hydro power generation at a day t_i is influenced by

- the air temperature at the preceding k_1-th day with respect to t_i, where k_1 is computed by considering the lag that maximizes the sample Pearson correlation [25] between the time series of the hydro power generation y and of the air temperature AT, say $\rho(y, AT)$;
- the precipitation at the preceding k_2-th day with respect to t_i, where k_2 is computed similarly to k_1 by considering $\rho(y, P)$;
- the sum of precipitation in the last $k_2 + 1$ days with respect to t_i, with k_2 defined above;
- the sum of precipitation in the last k_3 days with respect to t_i, where k_3 is the lag which maximizes $\rho(y, sumP)$, with $sumP$ being the moving sum of the precipitations;
- the snow depth at the preceding k_4-th day with respect to t_i, where k_4 is computed by evaluating the lag which maximizes $\rho(y, SD)$.

Depending on the country, not all the above listed datasets are relevant for the prediction of the hydro power generation. Then, in order to choose if a certain time series is used as input in the ML algorithms, we compute the correlation ρ between this time series and that of the response over the training period. Then, this dataset is added to the list of predictors if $|\rho|$ is bigger than a certain threshold $\bar{\rho}$. This choice was implemented as we observed that adding inputs whose correlation with the response is lower that a chosen threshold does not improve the prediction in terms of the evaluation criteria to be presented

below. Moreover, it is well-known that predictors generated as linear combinations of input variables may improve the accuracy of the learner. Then, in this paper, we also add to the input the meteorological data aggregated considering the national average.

For efficient learning, all input features as well as the output data are normalized by subtracting the mean and dividing by the standard deviation [4], which are computed over the training period. Then, the predicted hydro generation is re-transformed using the same normalization parameters.

Once the predictors are selected, these are used for training a learner. The way of learning depends on the ML algorithm selected. We use the ML algorithms presented in Section 3, and we generate several models for determining the one which provides the highest accuracy.

5 Model evaluation

In this section, we introduce the criteria selected for evaluating the prediction accuracy of the ML algorithms.

We are now in the second part of the ML workflow in Figure 2. Once a model has been built, we can use it for the prediction of the response by considering a new dataset of features. Such features are of the same type of the inputs described above, but corresponding to the time interval chosen for the prediction. For instance, if the time series of the air temperature over the training period was used for the generation of the model, now the time series of the air temperature over the new time interval will be used for the prediction. We also set the lags k_i, $i = 1, .., 4$ to the values computed in the training phase. The main difference here is that the input list does not include the time series of the hydro power generation, which instead will be the final output of this second phase.

As in this section we wish to measure the prediction accuracy of the ML algorithms, we will perform the second phase of the ML workflow over a time interval in which the time series of the response is actually known. We call $T_{\text{test}} = \{\tau_1, ..., \tau_M\}$ this daily spaced time interval and we indicate with $\bar{y} = [\bar{y}_{\tau_1}, ..., \bar{y}_{\tau_M}]$ the time series of the observed capacity factor over this testing period. From now on, we will use the term 'modeled' instead of 'predicted' output for the results of the ML process, which we indicate as $\hat{y} = [\hat{y}_{\tau_1}, ..., \hat{y}_{\tau_M}]$. It is important to highlight that the testing period is distinct from T_{train} and that $\bar{y}$ is not used as input to the model, but it will be used only for reason of comparison.

For the performance evaluation of the regression models used in this report, we consider the following measures:

- **Correlation coefficient (R)**

$$R = \frac{cov(\bar{y}, \hat{y})}{\sigma_{\bar{y}} \sigma_{\hat{y}}},$$

  where cov is the covariance, and $\sigma_{\bar{y}}$ and $\sigma_{\hat{y}}$ are the standard deviation of $\bar{y}$ and $\hat{y}$, respectively. It is a measure of the strength and direction of the linear relationship between the observed and the modeled variables.

- **Adjusted R-squared** ($\bar{R}^2$)

$$\bar{R}^2 = 1 - (1 - R^2)\frac{M-1}{M-m-1},$$

where M is the number of observations, m is the number of predictors and R^2 is the determination coefficient, that is the square of the correlation coefficient. It compares the explanatory power of regression models that contain different numbers of predictors. The adjusted R-squared is a modified version of R^2 that has been adjusted for the number of predictors in the model.

- **Mean Absolute Percentage Error** (MAPE)

$$MAPE = \frac{100}{M}\sum_{i=1}^{M}\left|\frac{\hat{y}_i - \bar{y}_i}{\bar{y}_i}\right|$$

- **Mean Squared Percentage Error** (MSPE)

$$MSPE = \frac{100}{M}\sum_{i=1}^{M}\left(\frac{\hat{y}_i - \bar{y}_i}{\bar{y}_i}\right)^2$$

MAPE and MSPE give a measure of the residual value between the observed and the modeled responses in percentage terms.

6 Results

In this section, we analyze the performance of the ML algorithms presented in Section 3 on two sets of experiments. First, we evaluate their accuracy for the computation of the one-year-ahead daily capacity factor. Then, the best trained models are used in the second experiments for estimating the response with a lead time bigger than one year.

6.1 Modeling the one-year-ahead hydro power generation

Let us start with the first analysis. For each one of the case studies, we set the training period T^i_{train} such that $t_1 = 1982\text{-}01\text{-}01$ and $t_N = (2003+i)\text{-}12\text{-}31$, whereas for the testing period T^i_{test} we set $\tau_1 = (2004+i)\text{-}01\text{-}01$ and $\tau_M = (2004+i)\text{-}12\text{-}31$, with $i \in \{1, 2, ..., 12\}$. For each i and for each ML algorithm, we generate a model which is used for computing the hydro power generation over the period T^i_{test}. The values of the modeled response are compared with the observed ones. Hence for each algorithm, we collect twelve values of each evaluation coefficient. In Table 1, we present the worst and the average performance of the ML algorithms. From this table, we can see that the algorithms based on ensemble of trees, both RF and BT and also the artificial neural networks perform quite well in most of the evaluative criteria. Moreover, the hybrid method introduced in Section 3.1 seems to obtain a slightly better accuracy than the RF algorithm in all the three case studies.

Table 1: Performance evaluation of the ML algorithms for the computation of the one-year-ahead run-of-river based hydro power generation for France (FR), Portugal (PT) and Spain (ES). We highlight in orange the two ML algorithms with the best performace and in yellow the results of the hybrid method obtained by their combination.

Country	ML	min R	avg R	min $\bar{R}^2$	avg $\bar{R}^2$	max MAPE	avg MAPE	max MSPE	avg MSPE
FR	RF	0.6461	0.8672	0.4077	0.7538	11.7623	8.4720	2.7961	1.5269
	BT	0.6113	0.8412	0.3632	0.7094	13.7290	9.0403	3.3508	1.8065
	LR	0.5963	0.7071	0.3448	0.4993	18.2396	12.9408	6.8176	3.3627
	SVM	0.6137	0.8442	0.3662	0.7142	11.6795	8.6811	2.9344	1.6266
	ANN	0.3294	0.6623	0.0935	0.4559	18.4999	13.7608	6.0919	3.6061
	hyb	0.6394	0.8707	0.3990	0.7595	11.5576	8.2000	2.6991	1.4431
PT	RF	0.7498	0.9057	0.5340	0.8164	19.6537	12.4886	5.4753	2.5109
	BT	0.7093	0.8889	0.4711	0.7856	21.0215	13.4691	6.5617	3.1715
	LR	0.6591	0.8478	0.3980	0.7080	28.0094	17.5775	11.0646	4.7761
	SVM	0.6044	0.8392	0.3245	0.6945	28.8882	17.4627	11.4810	4.6905
	ANN	0.6445	0.8690	0.3896	0.7575	21.3948	12.9696	7.1421	3.6641
	hyb	0.7058	0.9007	0.4898	0.8175	19.5963	12.5053	5.3985	2.5696
ES	RF	0.7498	0.9578	0.4990	0.9102	6.5761	4.0459	0.6980	0.3106
	BT	0.6567	0.9417	0.3490	0.8795	7.3091	4.7837	0.8845	0.4417
	LR	0.5329	0.8601	0.1804	0.7199	13.1970	10.4631	2.4494	1.5959
	SVM	0.5417	0.8593	0.1913	0.7165	16.0668	10.6341	3.3280	1.6646
	ANN	0.6267	0.9205	0.3049	0.8351	7.5446	6.3769	0.9904	0.6999
	hyb	0.7211	0.9552	0.5119	0.9161	6.6079	4.1791	0.7041	0.3148

In Figures 4, 5, and 6, we report the results obtained by the RF algorithm over the twelve testing periods. In particular, Figures 4(a), 5(a), and 6(a) show the comparison of the observed and the modeled time series of the hydro power generation capacity factor. Coherently with the evaluation criterion values, the modeled response is quite close to the observed data for all the cases considered. It is worthy to highlight that only testing data are shown in those figures. This means that once the leaner has been trained, the resulting model is fed only with a new set of meteorological input for the computation of the response.

Let us define the daily relative error at the calendar day j as follows

$$\left| \frac{\hat{y}_j - \bar{y}_j}{\bar{y}_j} \right|, \; j \in \{1, ..., 365\},$$

where $\hat{y}$ and $\bar{y}$ are the modeled and the observed response, respectively. We compute the calendar mean and the standard deviation of this error over T^i_{test} for all i and we report these values in Figures 4(b), 5(b), and 6(b). We can notice that the prediction is typically more accurate in the first part of the year for France, whereas for the case of Portugal and Spain the variability of the prediction error is smaller during the summer. By comparing these results with Figure 3, as expected, we notice that the level of the daily relative error is well

adjusted to the dynamics of the variability in the distribution of the data. In particular, in France graphs, we clearly have two distinct periods in the year: from January to the end of July the relative error is uniformly distributed between 3% and 7%, and from August to December the error increases to be uniformly distributed around 12%. This simple pattern is visible also in Figure 3(a) for same period and level of the standard deviation. In Portugal, the error in the daily prediction along the year is relatively uniformly distributed as well around 12% with a short period apart from mid-August to end of October where the magnitude of the error is slightly smaller around 9%. Again this is in phase with the pattern of the variability level that we find in the observations, see Figure 3(b). The situation is similar for Spain, which is characterized by the same pattern of the variability. The level of the relative error is slightly lower in the period of the lowest level of production. This is what we observe also for Portugal, but it is in contrast with the situation in France. It should be noticed that the overall level of error for Spain is globally better than for Portugal and France. However, this could be also due to the fact that the capacity factor values for Spain have been renormalized differently.

For sake of completeness, let us consider that the modeled capacity factor is given by the calendar mean of the observed response computed over the training period. This is the red line in Figure 3. We compute the relative error between this value and the observed response. The calendar mean and the standard deviation computed over the twelve testing years are depicted in Figure 7. As expected, these results show that the use of the calendar mean for the prediction leads to a much less accurate model.

6.2 Relevance for long term modeling

The second set of experiments aims at showing how the best trained models behave in terms of MAPE and adjusted R-squared values, when they are used for computing the hydro power generation with a lead time bigger than one year. In particular, in Figure 8, we present the performance the RF algorithm. First, we show the variation of the evaluation measures, when an RF model, which is built over a fixed training period, is used over several distinct testing periods. For the results in Figure 8(a), we consider the training period T_{train} such that $t_1 = 1982\text{-}01\text{-}01$ and $t_N = 2004\text{-}12\text{-}31$ and we generate a model by using the RF algorithm. This model is then used for the prediction over the intervals T^i_{test} with $\tau_1 = (2004+i)\text{-}01\text{-}01$ and $\tau_M = (2004+i)\text{-}12\text{-}31$ and $i = 2, ..., 12$. These results show that the trained model keeps the accuracy also for computing a response with a bigger lead time. This can be seen by comparing this figure with Table 1 in which MAPE and $\bar{R}^2$ are computed by performing a prediction with a lead team equal to one year. This evaluation view also allows us to point out countries or areas for which this modeling approach could be more sensitive. In our study, this is the case for Portugal where we clearly see that some years are quite difficult to be modeled, such as $2007, 2008$ and 2015. These years correspond to recent repeating events of very low level of rainfall that strongly contrast with long term historical data.

Now let 2016 be the year chosen for the testing. Then for T_{test} we have $\tau_1 = 2016\text{-}12\text{-}01$ and $\tau_M = 2016\text{-}12\text{-}31$. The training period is T^i_{train} such that $t_1 = 1982\text{-}01\text{-}01$ and $t_N = (2016-i)\text{-}12\text{-}31$, with $i = 2, ..., 12$ being the lead time. So, we retrain the RF model with less and less input information and we evaluate

the accuracy of the prediction over the same year. The results are depicted in Figure 8(b) and show that the elimination of more recent meteorological and energy data for the training set does not impact much the quality of prediction.

This second set of results are quite promising as they suggest that the proposed approach could be applied also for long term prediction.

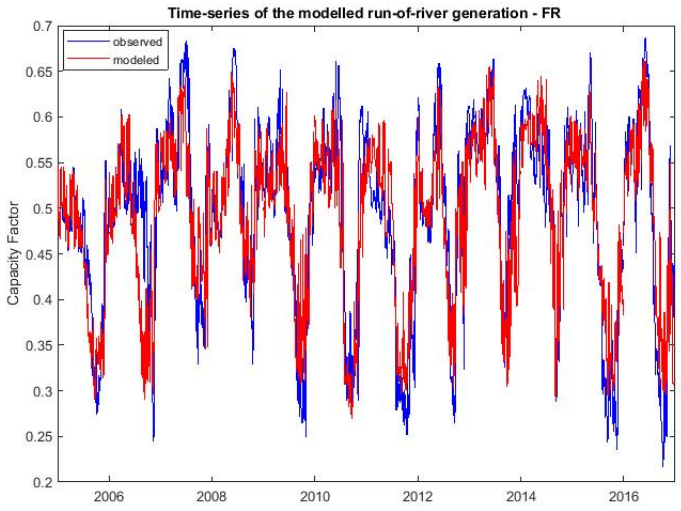
7 Conclusion and future works

This paper investigates the potential of using machine learning (ML) for predicting hydro power generation from meteorological data. We compared the performance of five ML algorithms and selected the models with the highest accuracy. The experiments showed that the algorithms based on ensemble of trees and the artificial neural networks perform quite well in most of the evaluative criteria. We also showed that the obtained models are quite stable and can be applied also for the long term prediction of the hydro power generation.

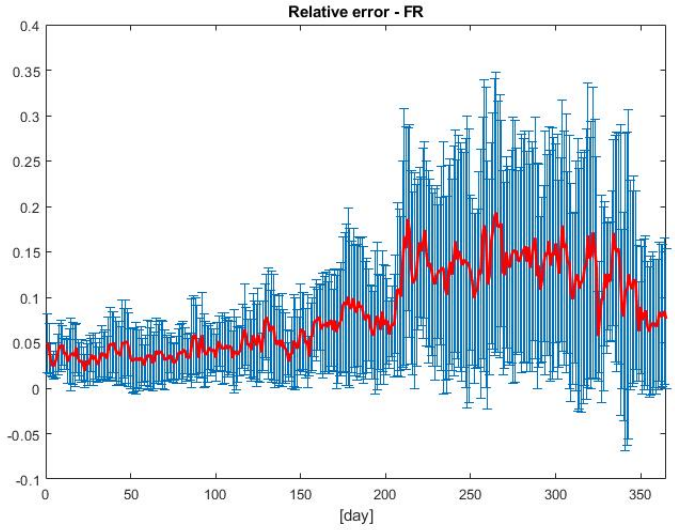
Future works will be dedicated to the extension of the proposed approach to other European countries. We also plan to evaluate the impact of the uncertainty of the weather data on the model results. The idea is to measure the variability of the predicted response for different meteorological data and enrich the prediction with dedicated variability scenarios.

The prediction of the hydro power production along with the estimation of the variability of the prediction error will be used within the CLIM2POWER project for the generation of appropriate power production scenarios for the optimal integration of renewable resources into existing power systems.

We will also study possible improvements in the ML prediction of hydro power production based on a two-step methodology. First, we use the information on the weather variables for predicting the river discharge of some selected basins. Then, in the second step, we pass from the predicted river discharge to the hydro power generation. The reason for this two-step approach is supported by the more direct physical connection between the rainfall and the river flow and the fact that the hydro power generation is typically a linear function of this river discharge. The application of this new strategy to the Portuguese case have shown promising results.

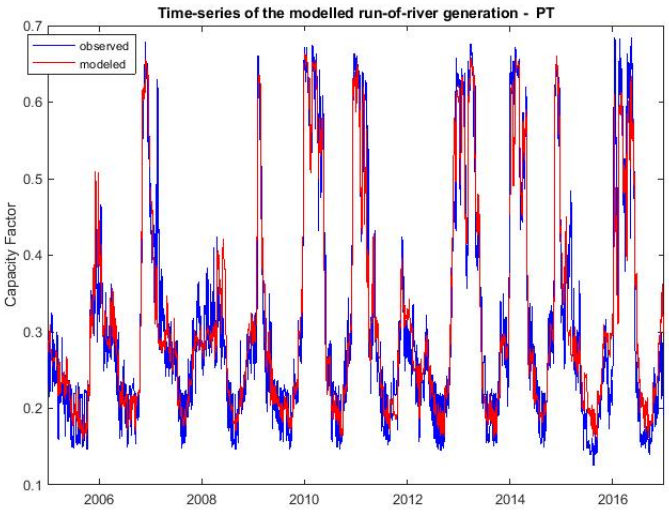


(a)

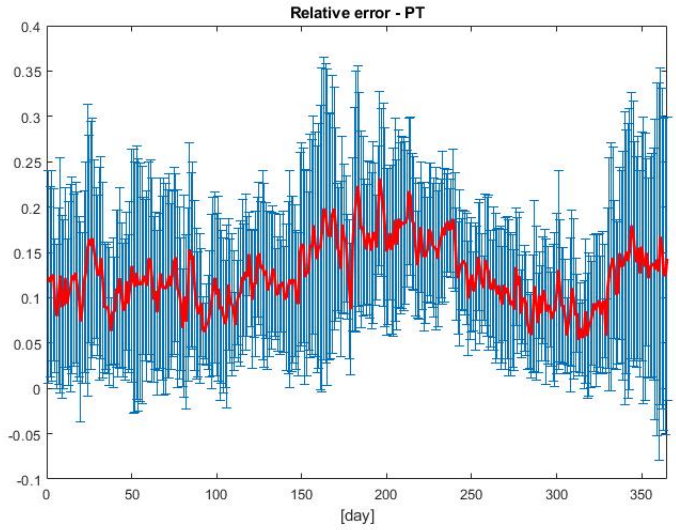


(b)

Figure 4: RF models for the hydro power prediction in France: (a) Time series of the observed and modeled capacity factor. (b) Calendar mean and standard deviation of the prediction error computed over the twelve years of testing.

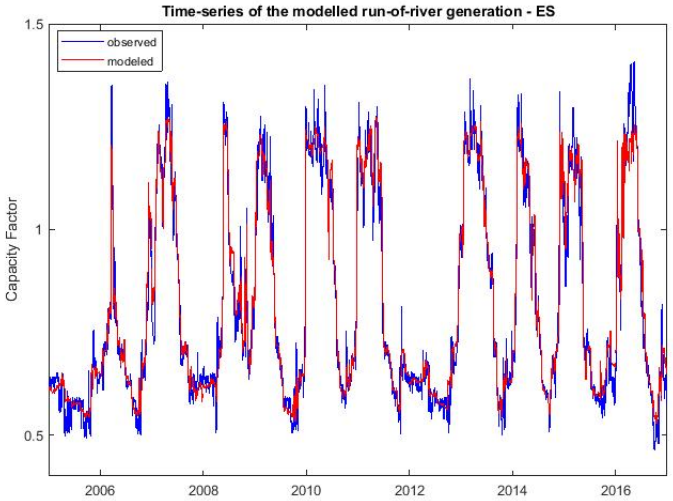


(a)

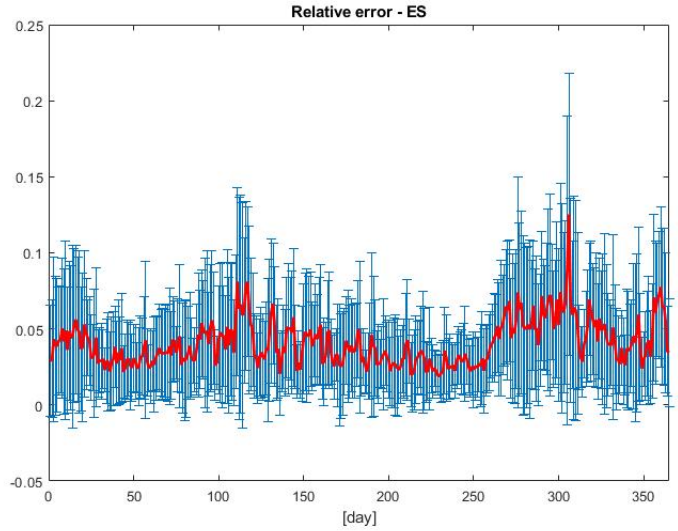


(b)

Figure 5: RF models for the hydro power prediction in Portugal: (a) Time series of the observed and modeled capacity factor. (b) Calendar mean and standard deviation of the prediction error computed over the twelve years of testing.

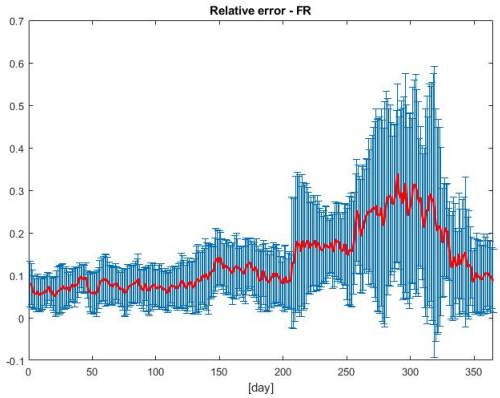


(a)

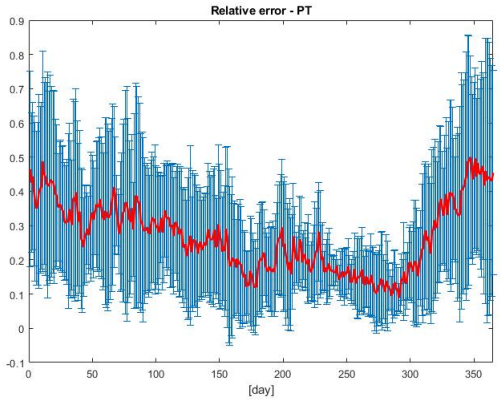


(b)

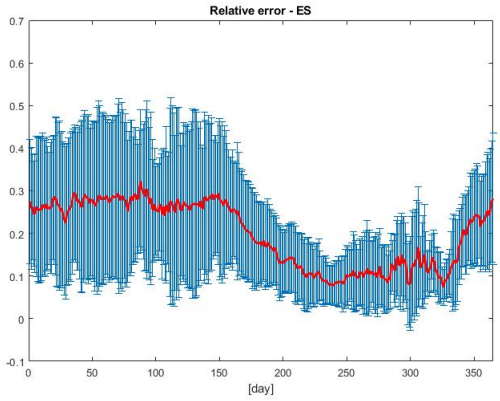
Figure 6: RF models for the hydro power prediction in Spain: (a) Time series of the observed and modeled capacity factor. In theory, the capacity factor assumes values in the range 0 and 1. In ECEM data, the capacity factor is obtained by considering a fixed installed capacity along the years. This is the reason for having a capacity factor bigger than one in this case. (b) Calendar mean and standard deviation of the prediction error computed over the twelve years of testing.



(a) France

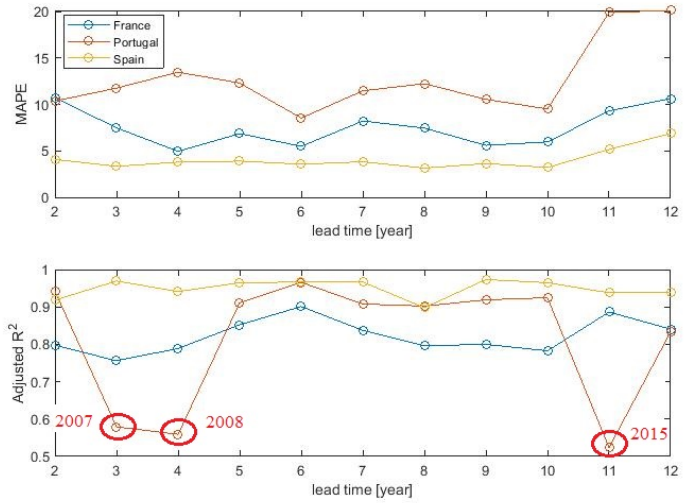


(b) Portugal

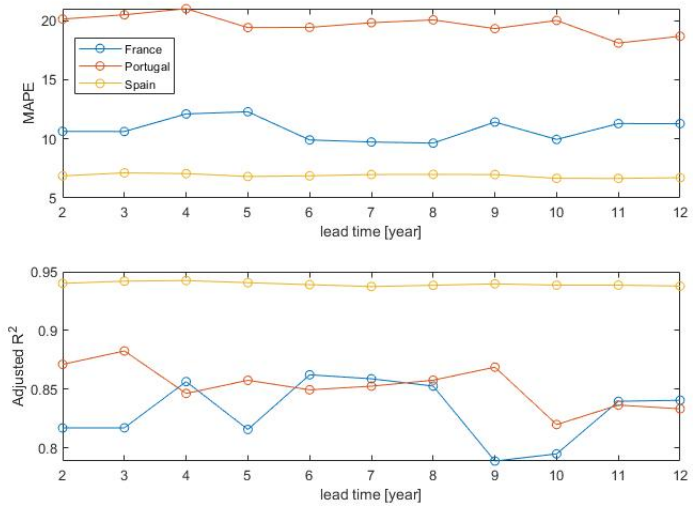


(c) Spain

Figure 7: Calendar mean and standard deviation of the prediction error computed over the twelve years of testing. Here the modeled response is the calendar mean of the observed data over the training period.



(a)



(b)

Figure 8: Performance of the RF algorithm for the prediction of hydro power generation with a lead time bigger than one year. (a) Evaluation errors obtained by considering a model trained over a fixed period for computing the response over several different testing periods. (a) Evaluation errors for modeling the response over the year 2006 obtained by training the RF algorithm with different input data.

References

[1] B. Stoll, J. Andrade, S. Cohen, G. Brinkman, C. Brancucci Martinez-Anido, Hydropower modeling challenges, Tech. Rep. WFGX. 1040, National Re-

newable Energy Laboratory, www.nrel.gov/publications (2017).

[2] B. Hamududu, A. Killingtveit, Assessing climate change impacts on global hydropower, Energies 5 (2012) 305–322.

[3] W. Köppen, The thermal zones of the earth according to the duration of hot, moderate and cold periods and to the impact of the heat on the organic world, Meteorologische Zeitschrift 20 (2011) 351–360.

[4] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag New York, 2009.

[5] N. A. Treiber, J. Heinermann, O. Kramer, Computational Sustainability, Studies in Computational Intelligence, Vol. 645, Springer International Publishing Switzerland, 2016, Ch. Wind Power Prediction with Machine Learning.

[6] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: A review, Renewable Energy 105 (2017) 569–582.

[7] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, M. Herrnegger, Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci. 22 (2018) 6005–6022.

[8] P. Drobinski, Wind and solar renewable energy potential resources estimation, Encyclopedia of Life Support Systems (EOLSS) (2012).

[9] L. Gaudard, F. Romerio, The future of hydropower in Europe: Interconnting climate, markets and policies, Enviromental science and policy 37 (2014) 172–181.

[10] B. Schaefli, Projecting hydropower production under future climates: a guide for decision-makers and modelers to interpret and design climate change impact assessments, WIREs Water 2 (2015) 271–289.

[11] [online] Available at: https://clim2power.com [Access 12 Dec. 2018].

[12] V. Krakowski, E. Assoumou, V. Mazauric, N. Maïzi, Feasible path toward 40-100% renewable energy shares for power supply in france by 2050: A prospective analysis, Appl. Energy. 171 (2016) 501–522.

[13] European Climatic Energy Mixes (ECEM) website, http://ecem.climate.copernicus.eu/ (accessed Jan, 2019) (2016).

[14] [online] Available at: https://www.entsoe.eu/db-query/miscellaneous/net-generating-capacity [Access 12 Dec. 2018].

[15] A. J. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and Computing 14 (2004) 199–222.

[16] J. H. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (2001) 1189–1232.

[17] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[18] MATLAB and Statistics and Machine Learning Toolbox release 2018b, the MathWorks, Inc., Natick, Massachusetts, United States.

[19] MATLAB and Deep Learning Toolbox, the MathWorks, Inc., Natick, Massachusetts, United States.

[20] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Advances in neural information processing systems (1997) 155–161.

[21] Y. Gala, A. Fernández, J. Díaz, J. R. Dorronsoro, Hybrid machine learning forecasting of solar radiation values, Neurocomputing 176 (2016) 48–59.

[22] [online] Available at: https://www.edf.fr/groupe-edf/nos-energies/carte-de-nos-implantations-industrielles-en-france [Access 12 Dec. 2018].

[23] [online] Available at: https://a-nossa-energia.edp.pt/centros_produtores/mapa_centrosProdutores.php [Access 12 Dec. 2018].

[24] [online] Available at: http://www.unesa.net/investigar/sabereinvestigar/mapas/centraleshidroelectricas.htm [Access 12 Dec. 2018].

[25] K. Pearson, Notes on regression and inheritance in the case of two parents, in: Proceedings of the Royal Society of London, Vol. 58, 1895, p. 240–242.