Article

Hourly Water Level Forecasting in an Hydroelectric Basin Using Spatial Interpolation and Artificial Intelligence

Mauro Tucci

Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Italy;
mauro.tucci@unipi.it

Abstract: In this work a new hydroelectric basin modelling approach is described and applied to the Pontecosi basin, Italy. Several types of data sources were used to learn the model: a number of weather stations, satellite observations, Reanalysis dataset and basin data. With the goal of predicting the water level of the basin, the model was composed by three cascade modules. Firstly, different spatial interpolation methods, such as Kriging, Radial Basis Function and Natural Neighbours, were compared and applied to interpolate the weather stations data nearby the basin area to infer the main environmental variables (air temperature, air humidity, precipitation and wind speed) in the basin area. Then, using these variables as inputs, a neural network was trained to predict the mean soil moisture concentration over the area, also to improve the low availability due to satellite orbits. Finally, a non-linear auto regressive exogenous input (NARX) model was trained to simulate the basin level with different prediction horizons, using the data from the previous modules and past basin data (water level, discharge flow rate, turbine flow rate). Accurate predictions of the basin water level were achieved within 1 to 6 hours ahead, with mean absolute errors (MAE) between 2cm and 10cm respectively.

Keywords: hydroelectric basin modelling; spatial interpolation; neural networks; Kriging

1. Introduction

Water is one of most important resources of the world. Monitoring rivers, basins and seas is a very important challenge for many different applications, from agricultural utilization to electric energy production [1]. Wrong management can lead to natural calamities such as floods and dry rivers [2]. From an energy production point of view, basin level and turbine water flow are deeply linked, and weather conditions can influence not only the basin level but also the plant operations [3]. Traditional physical models often struggle with parameters evaluation, and in any case a large measurement campaign must be conducted in the geographic area of study. Machine learning and black box modelling can overcome this problem, and for this reason a large number of studies in the literature related to water level forecasting and monitoring focus on artificial intelligence methods [4].

A monthly [5–9] or daily [11–14] average water level time series is usually considered for lakes and basins, while a prediction within hours [16,17] or even minutes [18] is often necessary in the case floods. To predict water levels, most of these works use past measurements, although from different sources, of the water level itself, in an auto-regressive fashion [5,7,11,12,14,16,17]. Some works also consider rainfall [13,18], and temperature as well [6]. Other than using local sensors, water level can be monitored by satellite altimeters [14] or radars [8,14]. Some of the cited works use feed-forward neural networks to predict the water level [7,16]. Very good prediction performance is reported in [16], with errors between 0.06m and 0.12m for 1 to 2 hours-ahead respectively, using water level data from multiple stations, indicating that adding more data sources can be beneficial. Other works use support vector machines (SVM) [5,11] or adaptive neuro-fuzzy inference system (ANFIS) [6,12]. In the case of hourly (or less) forecast horizons, NARX neural networks are better suited [17,18]. Some works [9] focus on clustering algorithms [10] and monitoring [8], while recently developed deep learning methods [15] have also been applied [14]. Other methods, such as boosted decision trees and Bayesian linear regression were reported to be successful for daily water level prediction in a hydroelectric basin using also rainfall data as input [13]. Successful applications of machine learning approaches for time series forecasting in the energy sector can be found also in the case of electricity price prediction [19], where Kalman filer and Echo State Networks are used, electrical load prediction [20], where a variant of the K-Nearest Neighbours algorithm is proposed, as well as prediction of the power generation from photovoltaic plants [21], and wind plants [22,23].

In this work, we focus on the Garfagnana valley hydroelectrical system, which is composed by several basins and power plants connected to each other. Managing the water resource all over the valley is a complex task, and knowing in advance the level of the basins would be of great convenience. Moreover, the need of an hydrological model is justified by the environmental challenges that nowadays are constantly on the spotlight: knowing the status of the water resource can be crucial to avoid natural calamities and to maintain the environmental flow.

The main objective of this work is to develop an accurate hourly forecasting model of the water level in a hydroelectric basin, exploiting a much larger number of data sources with respect to most of the literature.

In fact, a first contribution of this work is that a considerably large number of variables was used to create the model with respect to many other works related to water level forecasting. The data include several weather variables (temperature, humidity, precipitation and wind speed) from 14 weather stations (WS) scattered throughout the geographical area of the basin (northwestern area of Tuscany), satellite measurements of the soil moisture concentration (SMC), Reanalysis data of net solar radiance and snow depth, and hourly values of basin level, turbine flow rate and discharge flow rate. Moreover, data was collected for almost 19 consecutive month (631) days, starting June 6, 2017, and ending February 27, 2019.

A second contribution of this work is related to the articulate data preprocessing. In fact, some of the weather stations are several kilometers far from the basin. For this reason, the weather stations variables were spatially interpolated, considering the distance between each WS and the basin, to obtain an accurate average value in the basin area. In particular, three different spatial interpolation algorithms were tested (Kriging, Radial Basis Function, Natural Neighbour) and compared. Another problem is related to the low availability of the satellite SMC maps, which are obtained twice a week. To improve SMC availability, a neural network surrogate model was trained to predict SMC, using the spatially interpolated weather variables as inputs and the satellite SMC values as output.

As a final contribution, a state of the art NARX model was developed using all of the previously mentioned variables as inputs to predict the water level h hours ahead, for different values of the time horizon h. Hourly prediction of water levels, as well as NARX models, are most common in the case of flood prediction, but it can be certainly beneficial for the management of the hydroelectric basin for several operational reasons, especially in the case of several interconnected basins as in the case of the Garfagnana valley. As an example, an important activity in the hydroelectric basin is predictive maintenance [24–26], and an hourly prediction of the basin status can be conveniently used to perform condition monitoring and fault detection.

In section 2 the main characteristics of the hydrological system are presented, the input data of is presented and described in detail, the various models are formulated and their characteristics are highlighted. In section 3 the results of the methods are presented and discussed. Section 4 reports the conclusions and future developments.

2. Materials and Methods

2.1. Garfagnana Hydrological System

The Garfagnana hydrological system is the most important on the Tuscany region, the main river is Serchio (111 *km*), whose source is located in Monte Sillano (1864 *m*), and its basin has a total catchment surface of 1500 Km^2. An overview of the Serchio basin is shown in Fig. (1).

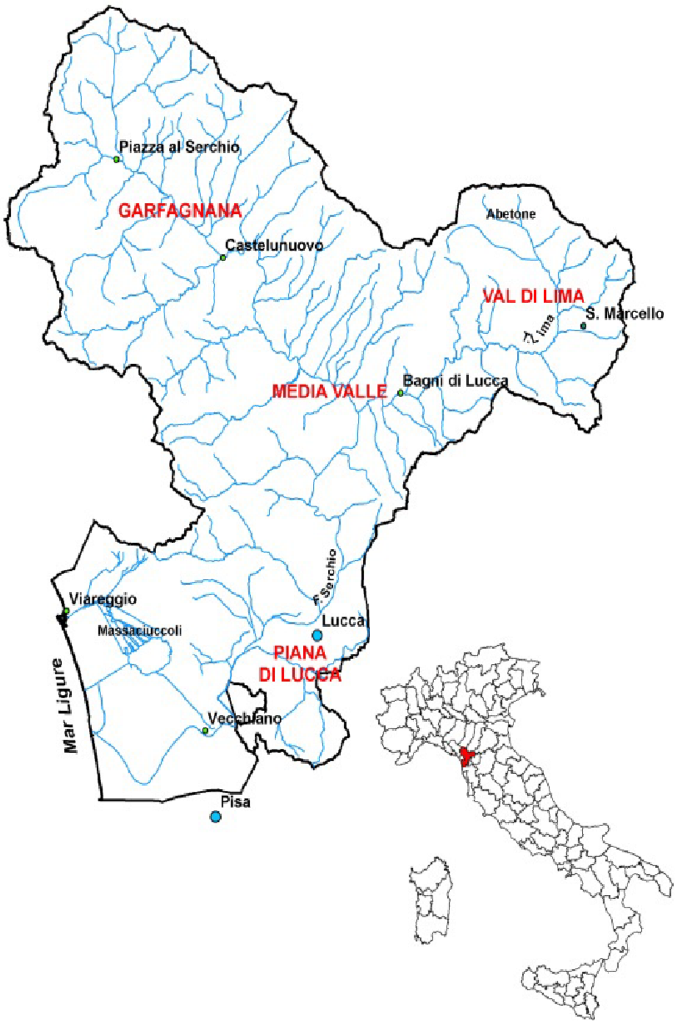

Figure 1. Overview of Serchio basin.

Serchio goes through the Garfagnana valley until it reaches the Tirreno sea. The presence of large amount of hydro resource and its various distribution on the Garfagnana valley has influenced the realization of a complex system of river, basins and power plants, as shown in Fig. (2). The production of a specific power plant depends from the production of all the other power plants upstream. For example, if an upstream plant does not use the turbine for a certain period, the downstream basin will receive less water than usual. On the other side the upstream basin will increase its water level. Optimizing the hydro resource all over the Garfagnana valley is one of the goals of this research. This kind of optimization is very complex because it requires a high level of coordination between all power plants of the area. The other main goal is to prevent the risk of flooding: the management of the hydro basin/plant must be handled in a safe way for the surrounding area. Knowing in advance the increase of water levels can help in taking the right decision and avoiding flooding. From a production point-of-view, having different basins at different altitudes

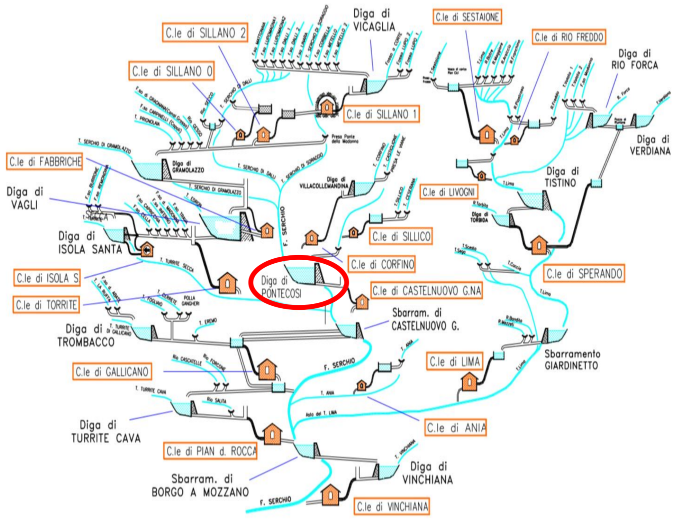

Figure 2. Hydro system in the area of Garfagnana.

requires a deep study for optimizing the global hydro resource. Being able to know the basin status in advance can strongly help the optimization process.

This work focuses on the Pontecosi basin which is located in the upper area of Garfagnana. It is an artificial lake with a 30 meters dam which serves as hydro-tank for the Castelnuovo hydroelectric power plant, located close to the lake. Castelnuovo power plant is the third biggest of the area; that justifies the importance of an accurate study of the hydro resource.

Figure 3. Pontecosi basin.

From an environmental safety point of view the basin status is represented by the water level of the basin: upper and lower bounds are set by environmental laws. Going over upper bound can cause floods; on the other side, a minimum level of water must be preserved to ensure the environmental flow (EF). The water level (neat head) in a hydroelectric basin depends on several variables: the flow used by the turbine, for example, is a human activity that influences the net head value. Higher is the turbine generation, higher is the amount of water withdrawn from the basin. Another aspect that influences the net head value is the EF, a minimum water outflow imposed by environmental laws to ensure a proper life quality of the river ecosystem. The EF must be guaranteed in all conditions by the owner of the plant.

2.2. *Datasets Characteristics*

2.2.1. Reanalysis Dataset

Reanalysis dataset [27] were initially created to improve the performance of the Numerical Weather Prediction (NWP) models. NWP models need data about the initial state of weather variables: Reanalysis data can provide it. Reanalysis datasets are generated by several data assimilation schemes and models. Data assimilation is a class of techniques which mixes different data sources, in order to obtain an output with less uncertainty than the original data. In this case data assimilation uses machine learning (ML) algorithms instead of physical laws: this is mainly pushed by the complexity of the laws that rule all the weather variables.

Reanalysis dataset contains a large set of weather variables, extended on a large datetime range. The decision to use also this kind of data source is related to the possibility of having more inputs to the basin status model, hence increasing its performance.

The dataset used in this work is the ERA-Interim [27] which is released by European Centre for Medium-Range Weather Forecasts (ECMWF). This dataset is based on a 2006 Integrated Forecast System (IFS). The main properties are: 4-D variational analysis, 80 km of horizontal resolution and 60 vertical levels (from the surface to 0.1 hPa). Similarly to satellite based observations, Reanalysis dataset is provided in a gridded way: a matrix is associated with a specified position (latitude and longitude) and time. In comparison with satellite data, Reanalysis has a lower resolution (80Km vs 5 Km): the accuracy of Reanalysis data on a small area around Pontecosi basin is lower with respect to satellite data. Considering this, the variables selected from the Reanalysis dataset were:

- net solar radiance,
- snow depth.

These variables are less influencing the basin status, with respect to rain and soil moisture concentration (SMC), so the lower accuracy of these variables can be accepted.

2.2.2. Weather Stations Data

Weather stations (WS) are the most direct way to measure a weather variable. Usually, a WS is a complex system of sensors including not only the sensing elements, but also a data acquisition and transmission system. High quality weather stations are composed by expensive products, so usually they are installed in key points of the area to be observed. In this project, we considered the following four weather variables acquired from weather stations:

- mean daily temperature (°C),
- mean daily air humidity (%),
- total daily precipitation (mm),
- mean daily wind speed (km/h).

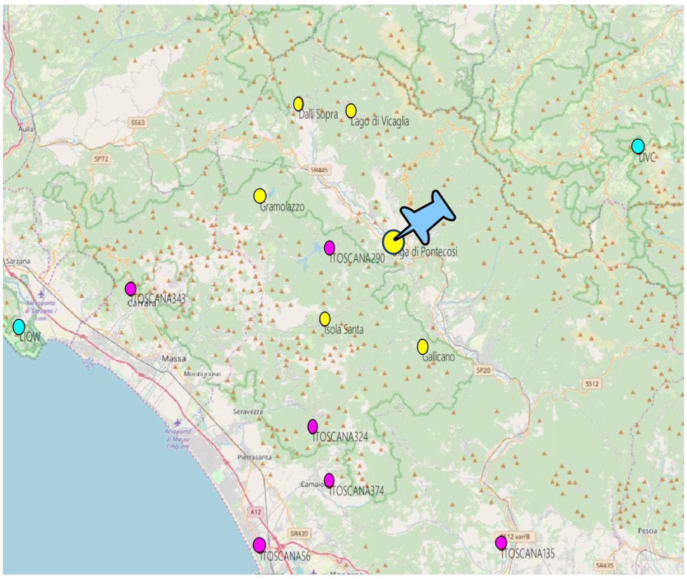

Figure 4. Weather stations positions. Yellow points are the plant owner stations, magenta points are Wunderground stations and cyan points are Aeronautica Militare stations

A total of 14 weather stations were used in this project, which are located in a large area nearby the Garfagnana Valley, as shown in Fig. 4. The chosen weather stations belong to three different providers:

- Wunderground Personal Weather Stations (PWS) is a platform which offers weather observation datasets. PWS used in this work are located close to the Garfagnana area. The 6 weather stations are marked with a magenta circles in Fig. 4, and they provide all 4 variables with worst-case availability of 90%. Wunderground dataset is the core of the input data of this project due to the high availability and proximity to the basin;
- Plant Owner Weather Stations. Plant owner made a weather measurements campaign in the Garfagnana valley. This dataset contains temperatures and air humidity from 6 weather stations located very close to Pontecosi area (yellow points), so they have high relevance with respect to the other providers;
- Aeronautica Militare (AM) Syrep stations are located far from the other stations, however additional two AM stations were selected to cover lack of availability of other providers (cyan points int the map). In addition, AM is the only authority which can certificate weather observations in Italy, so the data quality is generally very high.

2.2.3. Satellite data

Nowadays, satellite observations are widely used in many fields, from the environmental study to military use. Satellite images are gridded data: it means that values are stored in a matrix, which is geo-referenced through a raster object. For longitudes between 80°North to 80°South the raster object is usually based on a coordinate system called Universal Transverse Mercator (UTM), while for polar zones, Universal Polar Stereographic (UPS) projection is used. UTM system divides the Earth into 1200 zones: each of them has a size of 6°in longitude and 8°in latitude. Pontecosi zone is named 32N. Inside the zone, each point is defined within a grid. The images used in this project have a 8074x9885 resolution. To select a smaller area than the original, latitude and longitude can be converted in UTM coordinates: so a smaller matrix can be extracted. In this project, satellite data images consist of soil moisture concentration (SMC) maps of the area. More in-depth, SMC values are mapped on a wide area of Tuscany, and a small area around Pontecosi can be extracted as shown in Fig. 5. Satellite SMC images are available only twice a week. Averaging the values in the image a single scalar SMC value can be obtained, that represents the average soil moisture concentration around the basin.

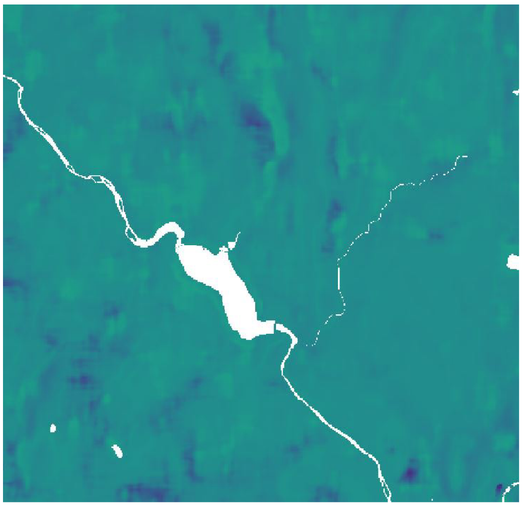

Figure 5. Soil Moisture Concentration satellite map around Pontecosi

2.2.4. Basin Data

Pontecosi basin data were necessary to generate a basin model. Variables included in this dataset are:

- basin level,
- turbine flow rate,
- discharge flow rate.

Each variable is given with a hourly time-step. It is important to note that these data contain several discharge flow events, which are major events during which the discharge flow outlet is activated to prevent flooding. The amount of water released by the basin during this kind of event is huge and it deeply influences the basin level.

2.3. Spatial Interpolation of Weather Stations Data

Point-wise data provided by Weather Stations (WS) cannot efficiently describe a weather variable in a large area. On the other hand, simply calculating an average value of the variables measured by several WS can lead to large estimation errors. In these cases the recommended practice is to use a spatial interpolator [28], which is a mathematical tool that can express a relationship between the spatial observations of a variable to predict its value in a different location. There is a large variety of spatial interpolators, and in this work we considered three main algorithms:

- Kriging [29]
- Radial Basis Function [30]
- Natural Neighbour [31]

The interpolated variable can be predicted in a specific new location as shown in Fig. (6), most precisely, the interpolated values can be averaged over a specific area around the basin.

2.3.1. Kriging interpolation

The term "Kriging" comes from Danie Krige, an engineer who first developed this particular method. Several variants of the the Kriging algorithm exist: in this work Ordinary Kriging is used. In the following, the n measurement sites are defined as "sample points".

The Kriging prediction $\hat{z}(\mathbf{x}_0)$ of a scalar quantity $z(\mathbf{x}_0)$ at an unobserved location $\mathbf{x}_0$ is a linear combination of the observed values $z(\mathbf{x}_i)$ at sample points $\mathbf{x}_i$ with scalar weights $w_i, i = 1 \dots n$:

$$\hat{z}(\mathbf{x}_0) = \sum_{i=1}^{n} w_i z(\mathbf{x}_i). \tag{1}$$

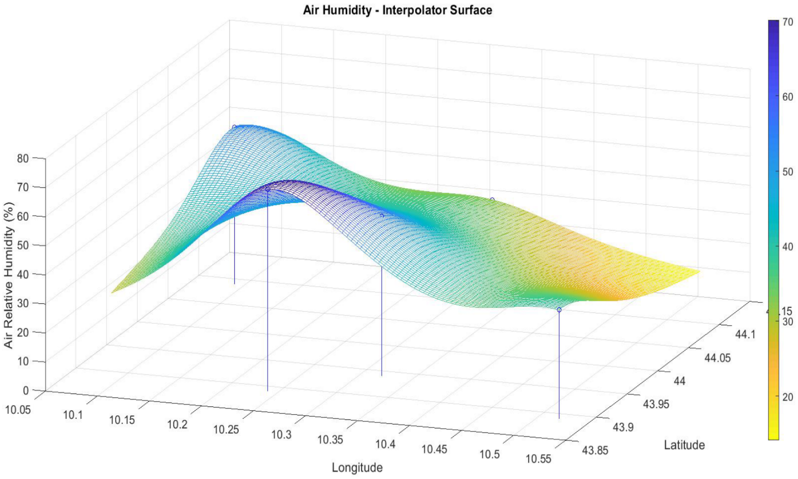

Figure 6. Kriging interpolation of air humidity over the area of interest in a r day. Vertical lines localize the weather stations.

The spatial variable $\mathbf{x}_i$ represents a vector in a three-dimensional coordinate system. To obtain an unbiased predictor the weights should sum to one and they are determined by minimizing the variance of the prediction error:

$$w_i = \underset{w_i}{\arg\min} \{E(\hat{z}(\mathbf{x}_0) - z(\mathbf{x}_0))^2\}, \tag{2}$$

where $E()$ denotes the expectation operator. The variance is minimized by the use of the so called semivariogram $\gamma(\mathbf{x}_i, \mathbf{x}_j)$, which is needed to calculate the variance in (2) and can be fitted using the historical dataset of measurements of the quantity $z(\mathbf{x}_i)$ at the sample points. Based on the homogeneity of samples in the area where the random variable $z(\mathbf{x}_i)$ is distributed, its first and second moments are usually assumed to be stationary, which means:

- all random variables have the same mean, that can be estimated by the arithmetic mean of sampled values;
- the correlation between two random variables solely depends on the spatial distance $h = ||\mathbf{x}_i - \mathbf{x}_j||$ between them and is independent of their location.

Under these assumptions the semivariogram is defined as:

$$\gamma(\mathbf{x}_i, \mathbf{x}_j) = \gamma(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} (z(\mathbf{x}_i) - z(\mathbf{x}_j))^2, \tag{3}$$

where $N(h)$ is the set of pairs of observations i, j such that $|\mathbf{x}_i - \mathbf{x}_j| = h$, and $|N(h)|$ is the number of pairs in the set. The solution of equation (2) is obtained using Lagrange multipliers as

$$\begin{bmatrix} \hat{\mathbf{W}} \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_1, \mathbf{x}_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{x}_n, \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_n, \mathbf{x}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(\mathbf{x}_1, \mathbf{x}_0) \\ \vdots \\ \gamma(\mathbf{x}_n, \mathbf{x}_0) \\ 1 \end{bmatrix}, \tag{4}$$

where $\hat{\mathbf{W}} = [w_1, w_2, \ldots, w_n]^T$ is the vector of estimated weigths and μ is the mean value. Commonly used fitting functions of $\gamma(h)$ are shown in table 1.

Table 1. Fitting functions of the semivariogram

Name	Function
Exponential	$\gamma(h) = C_0 + C_1(1 - e^{-\frac{h}{\theta}})$
Gaussian	$\gamma(h) = C_0 + C_1(1 - e^{-\frac{h^2}{\theta^2}})$
Spherical	$\gamma(h) = C_0 + C_1(\frac{3}{2}(-\frac{h}{\theta}) - \frac{1}{2}(\frac{-h}{\theta})^3)$

After building the empirical semivariogram, using the available dataset and equation (3), the next step is to choose the function that will fit it best. As can be seen in table, function expressions contains two unknown parameters C_0 and C_1, the distance h and a shape parameter $\theta > 0$. The shape parameter θ is an hyper-parameter that can be freely selected by the user: different θ values will cause different C_0 and C_1 values, which are determined, after selecting θ, with a least square problem resolution.

2.3.2. Radial Basis Function interpolation

Another commonly used spatial interpolation method relies on the Radial Basis Function (RBF). The predicted variable at an unknown location $\mathbf{x}_0$ is calculated by the RBF interpolator as:

$$\hat{z}(\mathbf{x}_0) = \sum_{i=1}^{n} w_i \phi(||\mathbf{x}_0 - \mathbf{x}_i||), \tag{5}$$

where the base RBF function $\phi(h)$ is defined as:

$$\phi(h) = e^{-\frac{h^2}{\theta^2}}. \tag{6}$$

After selecting the hyper-parameter θ, the weights $w_i, i = 1 \dots n$ are determined solving the least squares problem:

$$w_i = \arg\min_{w_i} \sum_{k=1}^{n} (\hat{z}(\mathbf{x}_k) - z(\mathbf{x}_k))^2. \tag{7}$$

2.3.3. Natural Neighbour Interpolation

Natural neighbour interpolation is a method developed by Robin Sibson[31], which works with a similar principle to Kriging. The basic equation is:

$$\hat{z}(\mathbf{x}_0) = \sum_{i=1}^{n} w_i z(\mathbf{x}_i). \tag{8}$$

The method differs in the way the weights are calculated, which is based on Voronoi tessellation of the discrete set of spatial points $\mathbf{x}_i, i = 1 \dots n$, also called centroids. Voronoi tessellation is a partition of the space into n regions (Voronoi cells), where cell i consists of all points of the space closer to $\mathbf{x}_i$ than to any other centroid. We define as $S_n(\mathbf{x}_i)$ the space region corresonding to cell i in the tessellation with n centroids, and with $|S_n(\mathbf{x}_i)|$ the size of cell i, which can be the volume or the area of the space region depending on the space being tree-dimensional or two-dimensional respectively. Given the new location $\mathbf{x}_0$, a new tessellation is computed using $n + 1$ centroids, adding $\mathbf{x}_0$ to the previous set of n centroids. The Sibson weight w_i is calculated as the ratio of the size of the intersection between the new cell $S_{n+1}(\mathbf{x}_0)$ and the old cell $S_n(\mathbf{x}_i)$ with respect to the size of the new cell:

$$w_i = \frac{|S_n(\mathbf{x}_i) \cap S_{n+1}(\mathbf{x}_0)|}{|S_{n+1}(\mathbf{x}_0)|} \tag{9}$$

2.4. Satellite Soil Moisture Modelling

An important step of this project is the SMC modelling. SMC represents the volumetric water concentration and it is expressed as the ratio between the volume of water and and the total volume of considered soil:

$$SMC = \frac{V_{water}}{V_{soil}}. \tag{10}$$

Typical SMC values goes from 0.0 (completely dry) to 0.6 (fully wet). Values above 0.6 are usually associated with rivers, seas, etc., so they are not related to the soil and have to be discarded. In this project, satellite SMC images are available only twice a week: this depends mainly on the satellite orbit around the Earth. Days without observations have to be “imputed”, which means assigning a reasonable value to the missing observations. To accomplish this goal, an ML model was realized , based on the satellite observations of SMC in the time range of the archived data. The inputs of the SMC model are the output values of the interpolation step, described in the previous section, over the entire Pontecosi area, i.e. the following variables:

- mean daily temperature on the area;
- mean daily air humidity on the area;
- total daily precipitation on the area;
- mean daily wind speed on the area.

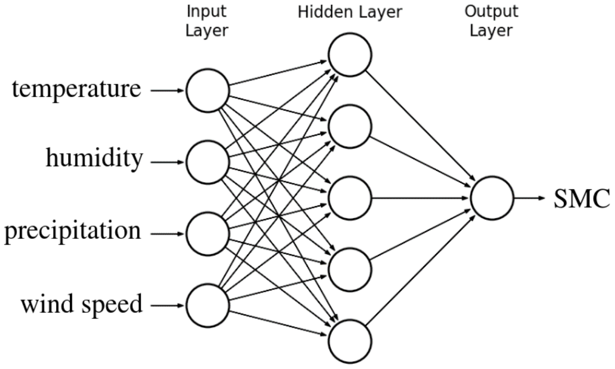

Figure 7. Neural network to predict SMC values.

This was decided thanks to the high availability of this data and the intimate relationship between them and SMC. The target data of the neural network is the average SMC over the basin area obtained from the satellite images. A one hidden layer feed-forward neural network [32] was selected to model SMC as a function of the four weather variables, as shown in Fig. 7. The hidden layer uses sigmoidal neurons and the training function employs the Bayesian regularization algorithm [33]. The number of neurons in the hidden layer was chosen through a model selection procedure: a 3-Fold Cross Validation was performed.

2.5. Basin Modelling

The core of this work is to predict the hydroelectric basin water level, that strongly depends not only on recent past levels but also on past turbine production and past weather events. According to this, a NARX model was created and tested: the possibility to have as inputs past level values makes the NARX choice suitable to this application.

NARX network is one of most common architectures used to model non-linear processes [34]. The basic principle is that the predicted output of the process $\hat{y}(t)$ depends on past values both of the input $\mathbf{u}(t)$ (that could also be multivariate) and the scalar output itself $y(t)$:

$$\begin{aligned}\hat{y}(t) = F(&y(t-h_y), y(t-h_y-1), y(t-h_y-2)\ldots, y(t-h_y-N_y),\\ &\mathbf{u}(t-h_u), \mathbf{u}(t-h_u-1), \mathbf{u}(t-h_u-2), \ldots, \mathbf{u}(t-h_u-N_u)),\end{aligned} \tag{11}$$

where $F()$ is the non-linear function that describe the process (usually modelled with a neural network), N_u and N_y are the number of delayed inputs and outputs respectively, while h_u and h_y represent the forecast horizons with respect to the input and output respectively. Equation (11) describes the so called open loop-NARX, while if the predictions $\hat{y}(t)$ themselves are used in feedback in place of the measured outputs $y(t)$ we have the so called closed-loop NARX:

$$\begin{aligned}\hat{y}(t) = F(&\hat{y}(t-h_y), \hat{y}(t-h_y-1), \hat{y}(t-h_y-2), \ldots, \hat{y}(t-h_y-N_y),\\ &\mathbf{u}(t-h_u), \mathbf{u}(t-h_u-1), \mathbf{u}(t-h_u-2), \ldots, \mathbf{u}(t-h_u-N_u)),\end{aligned} \tag{12}$$

The neural network is usually trained using the open-loop scheme, while the closed-loop approach can be used at the prediction step to achieve a forecast horizon larger than $\max(h_u, h_y)$, as far as the exogenous input $\mathbf{u}(t)$ is known. However the closed-loop scheme can propagate errors worsening the performance.

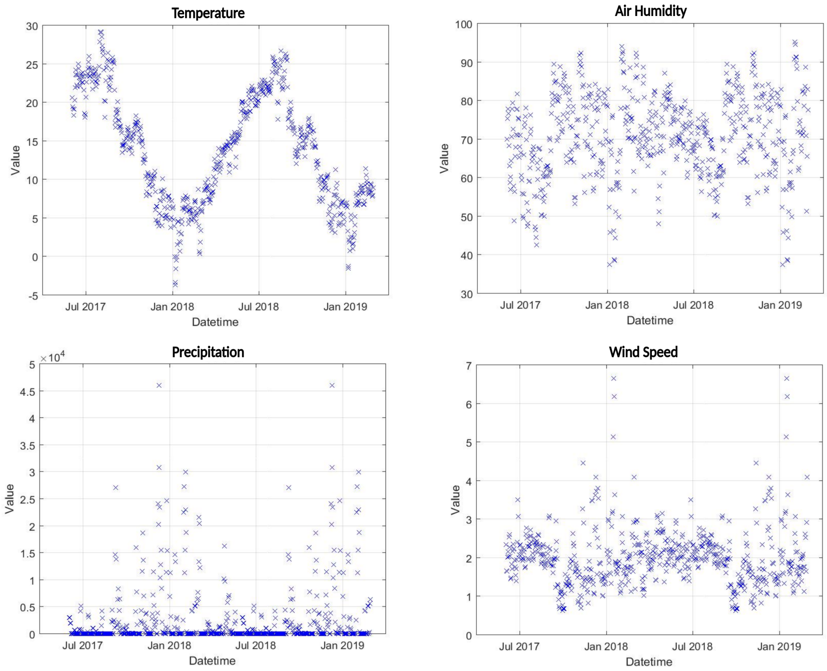

Figure 8. Daily values of the weather variables after averagig the Kriging interpolation over the area of interest.

3. Results and Discussion

3.1. Interpolation of Daily Weather Stations Data Results

The three kinds of interpolators previously described were tested on the weather stations dataset. In particular they were optimized and tested on each of the four daily variables separately.

RBF and Kriging require one hyperparameter, that has to be known in advance, while Natural Neighbour does not require any. The hyperparameter chosen should be the one that generalizes best the spatial distribution of the variable. The procedure used to select hyperparameters is described in the following steps:

1. hold out one station and consider it as the test station;
2. perform a leave one out cross validation on the remaining stations and determine the hyperparameter;

3. test the interpolator on the test station, taking as input all the remaining stations, and calculate the corresponding test error;
4. repeat the procedure until all the stations had been the test station once;
5. calculate averaged test error among all test stations.

This procedure was applied for each interpolator and variable, using the mean absolute error (MAE) as loss measure, defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{13}$$

where N is the number of test observations, y_i is the variable and $\hat{y}_i$ is the predicted value. The interpolator with lowest average MAE over all the test stations was chosen as the winning interpolator. This procedure was selected thanks to its robustness: each interpolator has been tested on every station, ensuring that the resulting model is optimized to perform well in each station of the dataset. Kriging, with Gaussian fitting function, was the winning interpolator for each variable. As shown in Table 2, which reports the results of the above procedure, prediction errors of the Kriging interpolator are small and acceptable. The second best performance is from Natural Neighbour interpolator, while RBF shows the worst performance, especially for the prediction of air humidity and wind speed.

Table 2. Averaged MAE of the interpolation methods.

Interpolator	Temperature	Air Humidity	Precipitation	Wind Speed
Kriging	2.17°C	8.76%	0.23 mm	1.68 Km/h
RBF	2.81°C	18.99%	0.25 mm	3.51 Km/h
Natural Neighbour	2.25°C	11.12%	0.28 mm	2.09 Km/h

3.2. Satellite Soil Moisture Modelling Results

The Kriging interpolator was used to generate daily values of the four weather variables over a 100 × 100 grid over the Pontecosi area, as shown in Figure 6. In particular, the interpolated values were generated also for the days when the satellite SMC values are available. All the spatially interpolated values were then averaged to provide a single daily value for each variable, which are shown in Figure 8 over all the dataset time-span. The four variables are then used as inputs to the NN that shall predict the SMC as output.

The result of a 3 fold cross-validation is shown in Figure 9. The optimal number of neurons in the hidden layer results to be 4, and the corresponding MAE values are around 0.035. Considering that all observed SMC values are between 0.1 and 0.35 the NN model can be considered accurate. Figure 10 shows the correlation matrix between all the variables, including SMC. In particular, SMC is most correlated with temperature and air humidity, while a low correlation can be observed with the wind speed.

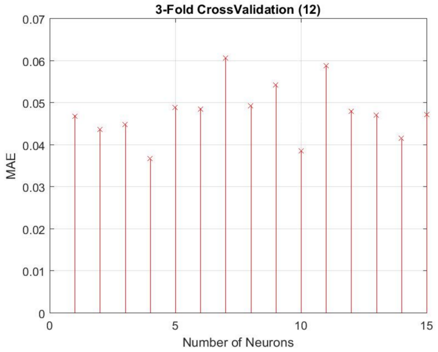

Figure 9. Selection of the number of neurons of the NN to predict SMC.

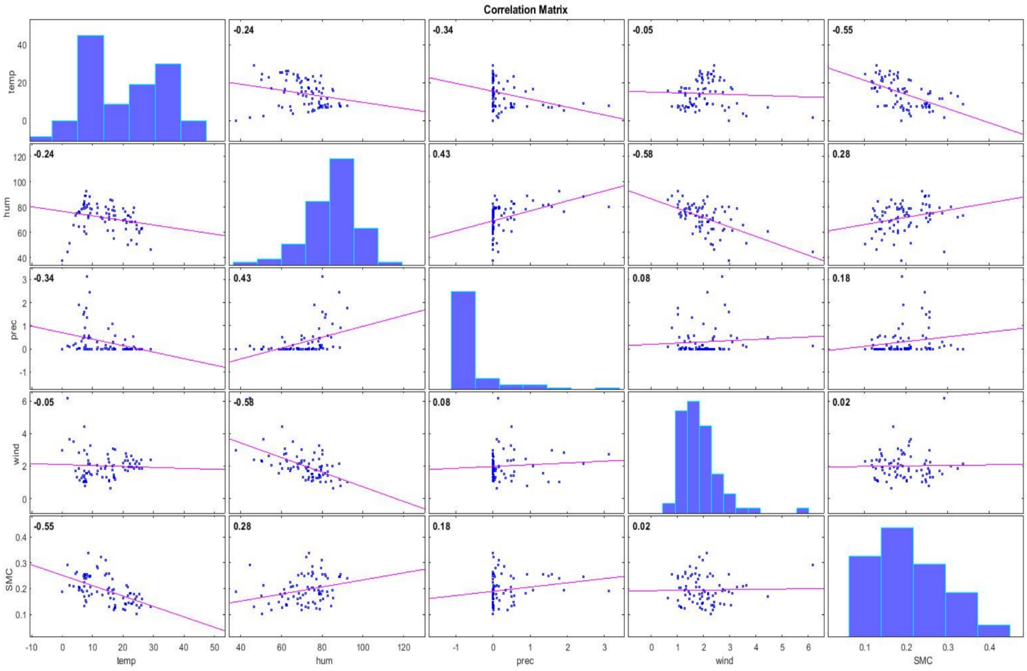

Figure 10. Correlation matrix.

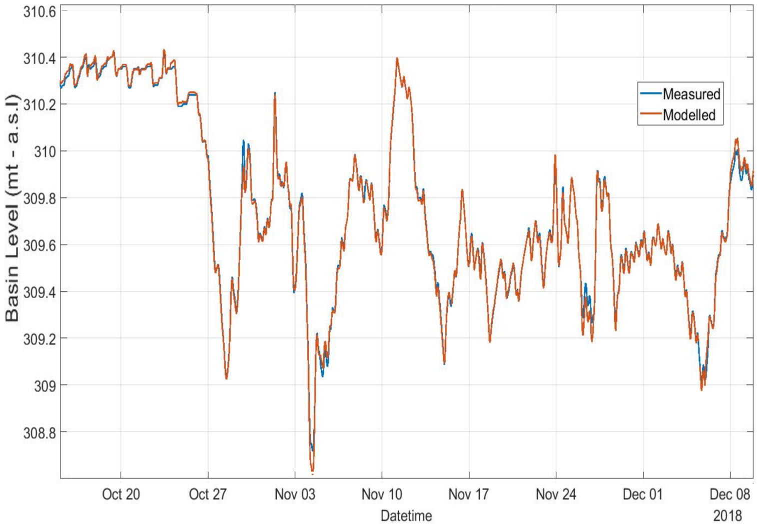

Figure 11. Predicted level during test period, comparison between NARX model and measured for $h = 1$ hour ahead. MAE is $2cm$.

3.3. Basin modelling results 355

The basin water level was modelled using the NARX model shown in equation (11). Considering that the basin data (water level, turbine flow rate and discharge flow rate) are provided hourly, we consider an hourly time-step in the NARX model. The exogenous input $\mathbf{u}(t) \in \mathbb{R}^9$ is composed by two flow rates and seven environmental variables: four of them are from Kriging interpolation of weather stations (temperature, air humidity, precipitation and wind speed), two of them are from Reanalysis dataset (snow depth, net solar radiance), and the last is SMC as predicted by the trained NN model. All the seven weather variables have a daily temporal resolution, and during day d we assume constant intraday hourly values equal to the last known values of day $d-1$. Regarding the output water level $y(t)$ it is measured in meters above sea level (a.s.l), and it represents also the tenth input to the NARX model as depicted in the complete scheme of the system shown in

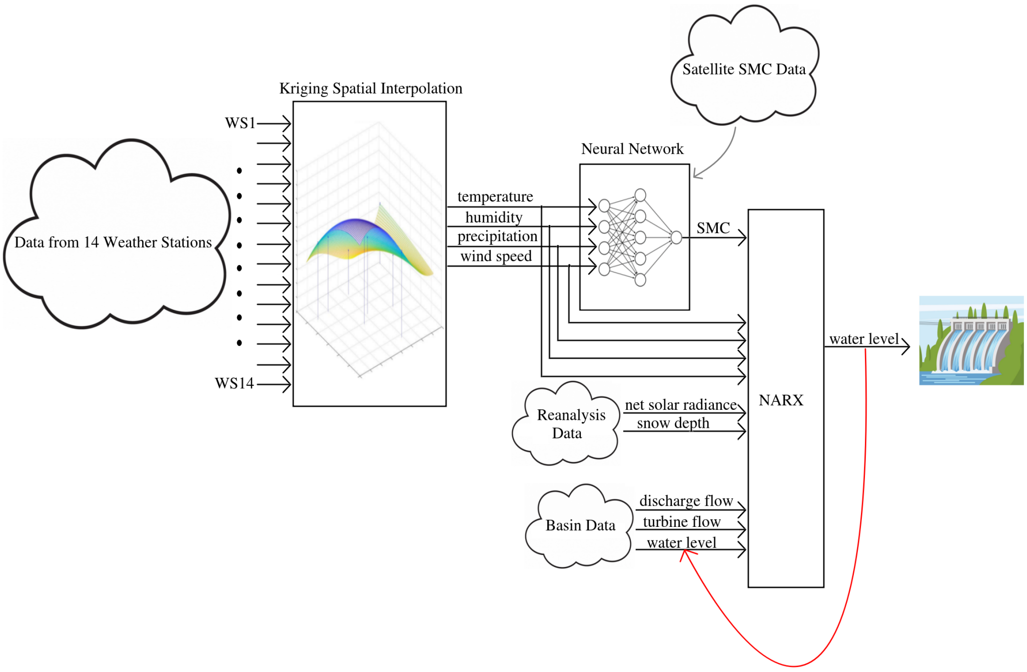

Figure 12. Complete schematic of the forecasting model.

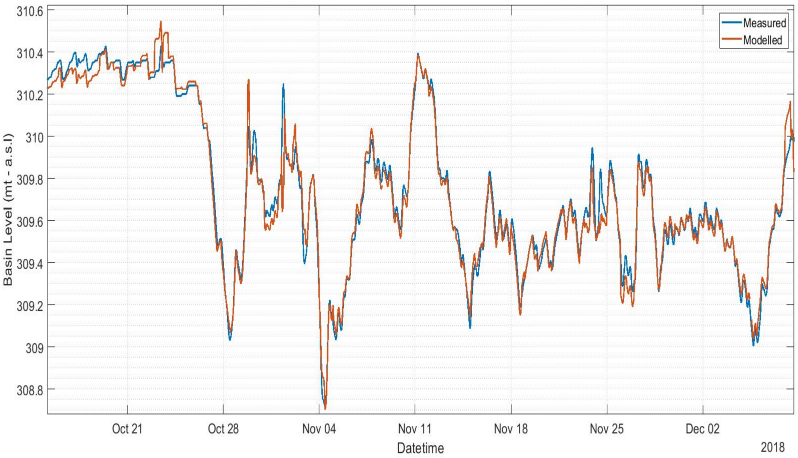

Figure 13. Predicted level during test period, comparison between NARX model and measured for $h = 6$ hours ahead. MAE is $10cm$.

Fig. 12. The time delays were set to $N_u = 0$, $N_y = 0$, while model was trained for different values of the forecast horizon $h = h_u = h_y = \{1, 3, 6, 12, 24\}$. The resulting NARX equation can be then written as:

$$\hat{y}(t) = F(y(t-h), \mathbf{u}(t-h)). \quad (14)$$

The NARX training period includes the first 500 days of data, and the test set consists in the successive 131 days. We used an expanding window approach: the test predictions during day d were obtained including day $d-1$ in the training set. Figures 11 and 13 show a comparison of the measured level against the predicted using an horizon $h = 1$ and $h = 6$ respectively. The mean absolute error MAE in the test set as a function of the forecast horizon is shown in Figure 14. The plant operator assumed as acceptable an error lower than $0.1m$, which is reached with a maximum of 6 hours of time horizon. Therefore our model produces acceptable results for $h \leq 6$, while with $h = 1$ the average error is within $0.02m$.

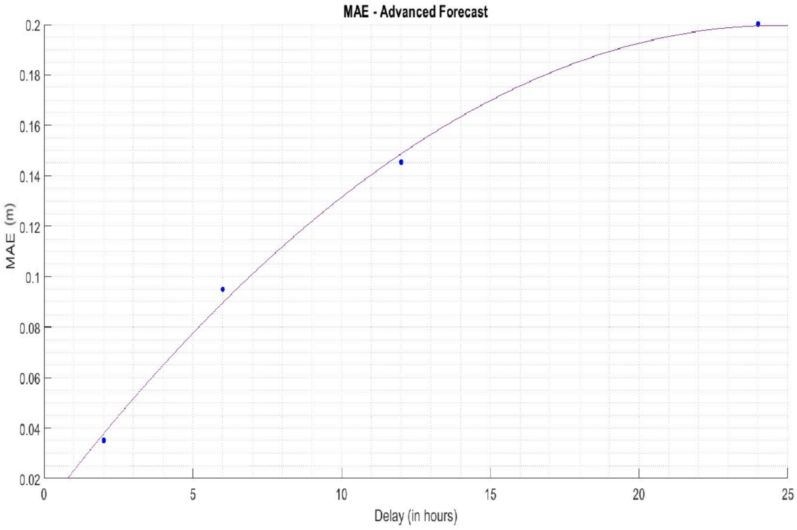

Figure 14. Prediction Error as a function of the forecast horizon.

4. Conclusions

The objective of this work was to develop a forecasting algorithm to aid a decision support system for the hydrological basin. The complete model is composed of several modules, the last of which is the basin model. To improve the performance, different data sources were used. Weather station data were managed through the use of spatial interpolators: this was necessary because of the large distances between weather stations and basin location. Different types of spatial interpolators were tested and compared, and Kriging was the best one. Soil moisture concentration values in days not available from satellite data were predicted using a neural network. Finally, a NARX model was created to predict the basin water level using past weather and basin data, tacking advantage of the previous modules. Predictions at different time horizons were simulated: acceptable results were achieved within 1 to 6 hour ahead. In particular, the basin level can be known up to 6 hours ahead with 10cm accuracy. For larger values of h the model produced larger errors, especially in the case of unpredictable discharge events. If these events are known in advance they can be integrated in the model improving its performance. The methodologies presented in this paper allowed to improve the input data quality and to model accurately an hydroelectric basin. As a future development, the author intend to integrate the model presented in this work with the condition monitoring model previously proposed in [24], also applied to photovoltaic plants [35], in order to improve the performance of the model based on SCADA (Supervisory Control And Data Acquisition) data [37] and self-organizing maps [36]. As a second future development we shall investigate the application of sensitivity analysis techniques, such as [38,39] to estimate the influence of the input variables to the performance of the model.

Funding: This research was funded by Hydrocontroller Regional Project

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable

Acknowledgments: The author wishes to acknowledge company i-EM (www.i-em.eu) for the support given in providing the data and testing the methods.

Conflicts of Interest: The author declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NARX	Non-linear Auto Regressive eXogenous input
RBF	Radial Basis Function
EF	Environmental Flow
NWP	Numerical Weather Prediction
ML	Machine Learning
ECMWF	European Centre for Medium-Range Weather Forecasts
IFS	Integrated Forecast System
PWS	Personal Weather Stations
AM	Areonautica Militare
UTM	Universal Transverse Mercator
UPS	Universal Polar Stereographic
SMC	Soil Moisture Concentration
WS	Weather Stations
MAE	Mean Absolute Error
NN	Neural Network
SCADA	Supervisory Control And Data Acquisition

References

1. Boretti, A.; Rosa, L. Reassessing the projections of the world water development report. *NPJ Clean Water* **2019**, *2*, 1-6.
2. Munawar, H.S.; Hammad, A.W.A.; Waller, S.T. Remote Sensing Methods for Flood Prediction: A Review. *Sensors* **2022**, *22*, 960. https://doi.org/10.3390/s22030960
3. Gaudard, L.; Gilli, M.; Romerio, F. Climate change impacts on hydropower management. *Water resources management* **2013**, *27*, 5143-5156.
4. Zhu, S.; Lu, H.; Ptak, M.; Dai, J.; Ji, Q. Lake water-level fluctuation forecasting using machine learning models: a systematic review. *Environmental Science and Pollution Research* **2020**, *27*, 44807-44819.
5. Khan, M. S.; Coulibaly, P. Application of support vector machine in lake water level prediction. *Journal of Hydrologic Engineering* **2006**, *11*, 199-205.
6. Ehteram, M.; Ferdowsi, A.; Faramarzpour, M.,; Al-Janabi, A. M. S.; Al-Ansari, N.; Bokde, N. D.; Yaseen, Z. M. Hybridization of artificial intelligence models with nature inspired optimization algorithms for lake water level prediction and uncertainty analysis. *Alexandria Engineering Journal* **2021**, *60*, 2193-2208.
7. Azad, A.S.; Sokkalingam, R.; Daud, H.; Adhikary, S.K.; Khurshid, H.; Mazlan, S.N.A.; Rabbani, M.B.A. Water Level Prediction through Hybrid SARIMA and ANN Models Based on Time Series Analysis: Red Hills Reservoir Case Study. *Sustainability* **2022**, *14*, 1843. https://doi.org/10.3390/su14031843
8. Bogning, S.; Frappart, F.; Blarel, F.; Niño, F.; Mahé, G.; Bricquet, J.-P.; Seyler, F.; Onguéné, R.; Etamé, J.; Paiz, M.-C.; Braun, J.-J. Monitoring Water Levels and Discharges Using Radar Altimetry in an Ungauged River Basin: The Case of the Ogooué. *Remote Sens.* **2018**, *10*, 350. https://doi.org/10.3390/rs10020350
9. Becker, M.; Da Silva, J.S.; Calmant, S.; Robinet, V.; Linguet, L.; Seyler, F. Water Level Fluctuations in the Congo Basin Derived from ENVISAT Satellite Altimetry. *Remote Sens.* **2014**, *6*, 9340-9358. https://doi.org/10.3390/rs6109340
10. Ferraro, P.; Crisostomi, E.; Tucci, M.; Raugi, M. Comparison and clustering analysis of the daily electrical load in eight European countries. *Electric Power Systems Research* **2016**, *141*, 114-123.
11. Cao, Y.; Yin, K.; Zhou, C.; Ahmed, B. Establishment of landslide groundwater level prediction model based on GA-SVM and influencing factor analysis. *Sensors* **2020**, *Sensors*, 845.
12. Seo, Y.; Kim, S.; Kisi, O.; Singh, V. P. Daily water level forecasting using wavelet decomposition and artificial intelligence techniques. *Journal of Hydrology* **2015**, *520*, 224-243.
13. Sapitang, M.; M. Ridwan, W.; Faizal Kushiar, K.; Najah Ahmed, A.; El-Shafie, A. Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy. *Sustainability* **2020**, *12*, 6121.
14. Baek, S.-S.; Pyo, J.; Chun, J.A. Prediction of Water Level and Water Quality Using a CNN-LSTM Combined Deep Learning Approach. *Water* **2020**, *12*, 3399. https://doi.org/10.3390/w12123399
15. Barmada, S.; Fontana, N.; Sani, L.; Thomopulos, D.; Tucci, M. Deep learning and reduced models for fast optimization in electromagnetics. *IEEE Transactions on Magnetics* **2020**, *56*, 1-4.
16. Sung, J. Y.; Lee, J.; Chung, I. M.; Heo, J. H. Hourly water level forecasting at tributary affected by main river condition. *Water* **2017**, *9*, 644.
17. Faruq, A.; Abdullah, S. S.; Marto, A.; Abu Bakar, M. A.; Mohd Hussein, S. F.; Che Razali, C. M. The use of radial basis function and non-linear autoregressive exogenous neural networks to forecast multi-step ahead of time flood water level. *International Journal of Advances in Intelligent Informatics* **2019**, *5*, 1-10.
18. Chang, F. J.; Chen, P. A.; Lu, Y. R.; Huang, E.; Chang, K. Y. Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control. *Journal of Hydrology* **2014**, *517*, 836-846.

19. Crisostomi, E.; Gallicchio, C.; Micheli, A.; Raugi, M.; Tucci, M. Prediction of the Italian electricity price for smart grid applications. *Neurocomputing* **2015**, *170*, 286-295.
20. Tucci, M.; Crisostomi, E.; Giunta, G.; Raugi, M. A multi-objective method for short-term load forecasting in European countries. *IEEE Transactions on Power Systems* **2015**, *31*, 3537-3547.
21. Gigoni, L.; Betti, A.; Crisostomi, E.; Franco, A.; Tucci, M.; Bizzarri, F.; Mucci, D. Day-ahead hourly forecasting of power generation from photovoltaic plants. *IEEE Transactions on Sustainable Energy* **2017**, *9*, 831-842.
22. Bai, L.; Crisostomi, E.; Raugi, M.; Tucci, M. Wind turbine power curve estimation based on earth mover distance and artificial neural networks. *IET Renewable Power Generation* **2019**, *13*, 2939-2946.
23. Bai, L.; Crisostomi, E.; Raugi, M.; Tucci, M. Wind power forecast using wind forecasts at different altitudes in convolutional neural networks. *In Proceedings of the 2019 IEEE Power & Energy Society General Meeting (PESGM), Atlanta, Georgia, USA, August 2019*.
24. Betti, A.; Crisostomi, E.; Paolinelli, G.; Piazzi, A.; Ruffini, F.; Tucci, M. Condition monitoring and predictive maintenance methodologies for hydropower plants equipment. *Renewable Energy* **2021**, *171*, 246-253.
25. Piazzi, A.; Tucci, M.; Ruffini, F.; Crisostomi, E. One year Operation of an Innovative Condition Monitoring Technique in Four Hydropower Plants.*In Proceedings of the 2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), Delft, The Netherlands, 26-28 October 2020*.
26. Betti, A.; Crisostomi, E.; Paolinelli, G.; Piazzi, A.; Ruffini, F.; Tucci, M. Condition monitoring and early diagnostics methodologies for hydropower plants. *arXiv preprint* **2019** *arXiv preprint*.
27. Dee, D. P.; Uppala, S. M.; Simmons, A. J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Vitart, F. The ERA-Interim Reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society* **2011**, *137*, 553-597.
28. Li, J.; Heap, A. D. Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software* **2014**, *53*, 173-189.
29. Oliver, M. A.; Webster, R. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System* **1990**, *4*, 313-332.
30. Wright, G. B. Radial basis function interpolation: numerical and analytical developments. Ph.D Thesis. University of Colorado at Boulder, Boulder, Colorado, USA, 2003.
31. Sibson, R. A brief description of natural neighbor interpolation (Chapter 2). In *Interpreting Multivariate Data*; Barnett V., Eds; John Wiley: Chichester, 1981; pp. 21–36.
32. Gurney, K. *An introduction to neural networks*; CRC press: London, UK, 2018.
33. MacKay, D. J. C. Bayesian interpolation. *Neural computation* **1992**, *4*, 415–447.
34. Lin, T.; Horne, B. G.; Tino, P.; Giles, C. L. Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Neural Networks* **1996**, *7*, 1329-1338.
35. Betti, A.; Tucci, M.; Crisostomi, E.; Piazzi, A.; Barmada, S.; Thomopulos, D. Fault prediction and early-detection in large pv power plants based on self-organizing maps. *Sensors* **2021**, *21*, 1687.
36. Tucci, M.; Raugi, M. Adaptive FIR neural model for centroid learning in self-organizing maps. *IEEE transactions on neural networks* **2010**, *21*, 948-960.
37. Gigoni, L.; Betti, A.; Tucci, M.; Crisostomi, E. A scalable predictive maintenance model for detecting wind turbine component failures based on SCADA data.*In Proceedings of the 2019 IEEE Power & Energy Society General Meeting (PESGM), Atlanta, Georgia, USA, August 2019*.
38. Barmada, S.; Musolino, A.; Raugi, M.; Tucci, M. Analysis of power lines uncertain parameter influence on power line communications. *IEEE transactions on power delivery* **2007**, *22*, 2163-2171.
39. Barmada, S.; Musolino, A.; Rizzo, R.; Tucci, M. Multi-resolution based sensitivity analysis of complex non-linear circuits. *IET circuits, devices & systems* **2012**, *6*, 176-186.