Predicting the behaviour of water distribution networks with machine learning models

Pedro Matos[1], Sérgio Matos[1,2], and A. Andrade-Campos[3]

[1] Departamento de Eletrónica, Telecomunicações e Informática, Universidade de Aveiro, Campus Universitário de Santiago, Aveiro, Portugal
[2] Instituto de Engenharia Electrónica e Informática de Aveiro, Campus Universitário de Santiago, Aveiro, Portugal
[3] Departamento de Engenhatia Mecânica, Campus Universitário de Santiago, Aveiro, Portugal
{pedroguilhermematos,aleixomatos,gilac}@ua.pt

Abstract. Water supply systems are indispensable infrastructures in any modern society, considering that a modern house is expected to have running water all the time. Water supply systems must pump water to meet their clients demands and face large cost-efficiency problems related to pumping operations. This work presents and analyses a possible solution to this problem using machine learning to both forecast water demands and simulate the consequent behaviour of the network which enables the optimisation of the energy cost. The study was conducted using data from real water demands from the central region of Portugal and previously modelled networks such as Richmond's network [12]. The results indicate that Artificial Neural Networks (ANNs) and Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) are capable of achieving good performance in forecasting water demands, and that it is possible to create a model that mimics the behaviour of a water supply network of reasonable size using ANNs.

Keywords: Water supply systems · machine learning · water demand forecasting · modelling · simulation · Richmond's network.

1 Introduction

Transporting water from place to place has its own associated costs. Between 90% and 95% of the electricity that its purchased by water supply systems is used for pumping [3]. The current operational strategy does not prioritise energy costs. It is possible to improve the efficiency of this process and reduce the cost associated with electrical consumption for pumping by as much as 20% to 25% [2], [10]. A possible solution to this problem is to take advantage of the electric tariff structure by scheduling the pumping operations to hours where the tariff is lower [8]. Matching the pumping operations with the lowest electric tariff is no easy task. Several constraints must be respected, the most important of which is that the clients water demands must always be satisfied.

1.1 A conceptual decision support system

A decision support system that can forecast the water demands, simulate the behaviour of the network, calculate the system energy use for a given pumping schedule, and eventually find the optimal pumping schedule is a useful tool to help water supply companies to achieve the goal of reducing the energy consumption costs. Such system would also allow the companies to examine the future state of the network which in itself is a very useful feature. Taking this into consideration, it is easy to conceptualize a decision support system that helps the company responsible for managing the water distribution network. Such system can be composed of three modules:

- **Water demand prediction module** - The pumping schedule will depend on the water demand, *i. e.*, the amount of water that is going to be needed at each moment of the day will directly affect the number of pumping operations and the time of those operations. This module must predict, for a given observation window with a certain time-step, the water demand during the day.
- **Network simulation module** - Based on the current state of the network and on the pumping operations that are expected to be executed to satisfy the predicted water demands, this module predicts and simulates the network behaviour and calculates the energy consumption to execute those operations.
- **Optimisation** - The main goal of the module is to minimise the cost of energy consumption by matching the schedule of the pumping operations with the best electric tariff structure while satisfying water demands. This module is not in the scope of this work.

In order to work, the system must be supplied with water demands that were observed in the past, and the water demand prediction module will then create a prediction for a certain time window in the future (for example the next 24 h). With the predicted water demands it is then possible to know how much water must be supplied and/or pumped to keep the clients satisfied and, consequently, the amount of water that is going to be drained from the network. A second set of information is sent, related with the current state of the network, tank levels, pump operations that must be performed in order to satisfy the water demands, etc. This information combined with the predicted water demands is passed to the water simulation module. The module outputs the consequent state of the network and how much energy would be spent in order to achieve it. Taking this into consideration the optimization model tries to find a pumping schedule that minimizes the energy costs. With this new pumping schedule, a new simulation is formed, and the process is repeated until the optimal schedule is found. When that happens the optimal schedule and the consequent network simulated state can be sent to the management company, allowing them to take an informed decision on how to perform the pumping operations.

1.2 Objectives

A decision support system like the one previously mentioned requires a very precise water demand prediction module. All the processes of simulating the behaviour of the network and optimizing the pumping schedule are based on the capability of predicting water demands. The current literature related to this topic demonstrates several solutions capable of achieving very good results. Since water demand forecasting is such an important part of the system, it is essential that the best possible results can be achieved. Therefore, the first objective of this work is to develop and improve a state-of-the-art solution for this problem.

Concerning the water simulation module, the most common solution is to use the industry standard: the EPANET software [1]. The EPANET software has some problems concerning the computational effort required to both run and calibrate the software. Additionally, the calibration process requires that an operator adjust the roughness of the pipes, water leaks, etc. until the model is able to correspond to the real-life scenario. This makes its use somewhat impractical. A solution that does not require a calibration process and it is more computational efficient is required. The literature points out that it is possible to use machine learning to meta-model a water distribution, eliminating the need to use a simulation model like EPANET. Thus, the second objective of this work is to implement and analyse a machine learning technique to model a water distribution network.

2 State-of-the-art

2.1 Machine learning in forecasting water demands

Antunes et al. [2] explore and analyse several machine learning techniques for short-term water demand forecast such as, KNN, SVR, Random Forest Regression and Artificial Neural Networks. The results of these techniques are compared to traditional mathematical models. Several model benchmarks are presented on different datasets with the intention of analysing their forecasting performance. The data is provided by two water companies, one on the northern region of Portugal and the other on the central region, dating between September 2012 to July 2013 and September 2015 to December 2016, respectively. The data from the two regions has both quality and quantity errors, so only selected points of the networks are being considered. Even so, it was possible to identify some problems regarding outliers, here defined as values that were not in the range between the average and margin of 3 times the standard deviation, and absence of data. The absent values are replaced with the global average. A second iteration of this process is done to reduce the impact of the errors of the first iteration, after the normalization of values. This study concluded that small ANN with the LBFGS learning algorithm and the ReLU activation function have the best results respecting the criteria of 24 h forecast window and using approximately

2 weeks of previous water demand observations.

Guancheng et al. [6] explore the potential of deep learning in water demand forecasting and compare its performance with conventional ANN models. In this study, a Gated Recurrent Unit Network (GRUN) was developed, along with an ANN. Both models were used for 15 min online forecasts where the model receives data updates making each prediction on real observed data, and for 24 h forecasts with 15 min timesteps where the output of the previous moment forecast was used as input for the next moment until 96 values were forecasted. A correction model was implemented with the goal of decreasing the accumulated error that occurs between forecasts. This model takes the output of 96 values of the 24 h forecast and approximates them to the real values. Two more model variants can then be considered: ANN-correction and GRUN-correction. The correction model consists of an Artificial Neural Network model with one dense layer trained with the predicted values obtained from the ANN model or ANN model, and sample labels are the observed values obtained from the field. This model is only applied in the 24 h forecast. The GRUN model demonstrates better performances in both 15 min forecast and 24 h. The application of the correction model improved the prediction accuracy of the models, making the GRUN-correction model the one with the best performance overall. In [6] it is concluded that the deep learning-based method that was proposed achieves an accurate and reliable water demand prediction for the 15 min and 24 h. The GRUN predicts more accurately and is more stable than the conventional ANN model.

2.2 Machine learning in hydraulic simulation

Rao and Alvarruiz [9] present an entirely new approach for using machine learning in the water distribution systems focusing on operational control rather than on planning or design exercises as all other applications known until that date. This study describes the development of an ANN that takes as input:

- Control variables representing pump settings,
- Valves settings,
- Water demands for a certain time period,
- Storage tank levels,

And outputs:

- Pump power consumption for a certain time period,
- Pressures in specific network nodes,
- Flow in specific network pipes,
- Storage tank levels.

The model was applied to a hypothetical network, the Any Town network [11] since it is simple, well-documented and extensively modelled previously. The results are compared to those of the EPANET model on the same network. A high degree of accuracy and a 10-fold reduction in the computational time required by conventional simulation models was reported in [9].

Salomons et al. [10] applied the same technique to real world data from Haifa-A, Israel. The goal of this study was to reduce the operational costs of water distribution system by finding the near-optimal control process, that is the one that matches with the best energy tariff structure. It explains that the typical operating regime of the storage tanks depends on their water levels. When the level goes below a certain margin the pump is switched on and vice-versa. This process does not consider the energy consumption to be a high priority and no special attention is given to the energy tariff structure. The problem is formulated by an objective function that minimizes the overall cost of delivering the required amount of water in a given period. A 24 h operating horizon was adopted since water distribution systems operate in a daily cycle, and a time step of 1 hr was chosen considering the impact of the computational burden. A method that is referred as a DRAGAN (Dynamic, Real-time, Adaptive Genetic Algorithm Artificial Neural Network) was developed, an ANN combined with a Genetic Algorithm (GA). The ANN replicates a conventional hydraulic simulation model, this is significantly more computationally efficient than using a simulation model directly. The cost of each potential solution proposed by the GA is estimated by the ANN. In order to capture the domain knowledge of a conventional hydraulic simulation model, the ANN is used to map a multivariate space (inputs) to another (outputs). It can be regarded as a input/output model. That said, to train the ANN, a set of input/output pairs was generated by EPANET. To forecast the water demands a method, which consists of a combination of Fourier series and time-series analysis is used. Salomons et al. [10] reports a potential cost reduction of about 25%.

3 Methodology

3.1 Developing the water forecasting module

Since it is not possible to draw a definitive conclusion on the best technique, two methods were implemented in order to assess their performance on short-term forecasting. The data used in this study was provided by a water utility in the northern region of Portugal. It represents data collected from a point in a water distribution network designated as WD4. The observed water demands date between September 2012 and July 2013.

3.2 Handing the data

All the data is processed beforehand, adopting the methodology of [2]. Outliers are identified by singling out values that are above or below 3 times the standard deviation. Values that correspond to the upper and lower limit of the allowed range values are then assigned to those outliers. A second iteration of this process is executed to mitigate the impact of the errors that were previously found.

In order to train the models, it is necessary to create a dataset of input features associated with a given output, in this case, a water demand at time t, denoted as D_t. This dataset must be constructed from a stationary time series of water demands $[D_{t-\gamma},\ ...,\ D_{t-2},\ D_{t-1}]$, where γ is the number of previous recorded water demands. Taking in consideration the procedure described in [2], which results in a good model performance, all the data is pre-processed removing its outliers. Then, the data is arranged so that to predict a water demand at time t, D_t, the model is given the values of the demands that took place an operational horizon (24 h) prior to t, which means that D_t is predicted using $[D_{t-24},\ D_{t-24\ *\ 2}, D_{t-24\ *\ 3},\ ...,\ D_{t-\ 24n}]$, where n is the number of demands of prior days that are being considered, and, consequently, the number of features that are being used to predict D_t.

3.3 Artificial Neural Network

The data used in [2] is similar to the data used in this work since it originates from the same water consumption point, WD4. Thus, it is expected that the topology of the artificial neural network (ANN) [7] that is used in that study can achieve the same results using this work's data. With that in mind, the implementation and training of the ANN followed the criteria that achieved the best results in [2], which was an ANN using the "identity" activation function, two hidden layers with 8 and 25 nodes, respectively, and LBFGS [4] as its learning algorithm.

3.4 Gated Recurrent Unit Network

The gated recurrent unit network (GRUN) model was developed based on the architecture presented in [6]. It consists of three GRUN layers that process data at different time periods. The GRUN layers put the water demand data through an nonlinear transformation that produces a memory state for past water demand, establishing dependencies among demands at different time periods. The output of the GRUN layers is then concatenated in a merge layer which is connected to a set of regular fully connected layers, that all except the output layers which uses a linear activation function, use the ReLU activation function.

For this model some adaptation to the way the data is handled had to be made. The data arrangement is similar to 3.2. However, the feature input vector is divided into three sub-vectors: recent, near and distant. So, for example, if the feature vector is $[D_{t-24}, D_{t-24*2}, ..., D_{t-24*15}]$, the recent sub-vector will be $[D_{t-24}, ..., D_{t-24*4}]$, the near sub-vector $[D_{t-24*5}, ..., D_{t-24*9}]$, and the distant sub-vector $[D_{t-24*10}, ..., D_{t-24*14}]$. Like the name suggest, the recent sub-vector corresponds to the demands that are the closest to the value that is being predicted, D_t, unlike the distant sub-vector which corresponds to the features that are the furthest way from D_t.

3.5 Developing the simulation model

The choice of a time-step influences the model's computational burden. Nothing in the literature indicates that an 1 h time-step is not enough to faithfully simulate the behaviour of the network.Thus, a conservative 1 h time-step was adopted. The hydraulic simulation model can be described as an input-output model like the one depicted in Figure 1, where:

- $[P_t^1,\ P_t^2,\ ...,\ P_t^n]$ input vectors of values in $[0, 1]$ range representing a fraction of the given timestep from t to Δt, where a certain pump was on at time t. Given by $P_t^i = \frac{t_{on}^i}{\Delta t}$, where P_t^i and t_{on}^i are the status of pump i and the time that pump i was active in a certain timestep Δt, respectively. $P = 1$ means the pump was always active between t and Δt. n represents the total number of pumps.
- $[S_t^1,\ S_t^2,\ ...,\ S_t^y]$ input vectors of values representing the water level, in meters, of the certain tanks at time t, where y represents the total number of storage tanks.
- $[D_t^1,\ D_t^2,\ ...,\ D_t^j]$ input vectors of values representing the water demands, in cubic meters, of each consumption point of the network between time t and $t + \Delta t$, where j represents the total number of consumption point on the network.
- A_t - an input vector of values containing the aggregated demand of every operational window (*e. g.* 24 h) at time t in cubic meters. Given by $A_t = \begin{Bmatrix} \sum_{i=1}^{j} D_t^i,\ t=0 \\ A_{t-1}+\sum_{i=1}^{j} D_t^i,\ t>0 \end{Bmatrix}$
- $[S_{t+\Delta t}^1, S_{t+\Delta t}^2, ..., S_{t+\Delta t}^y]$ output vectors of values representing the water level, in meters, of the certain tanks at time $t + \Delta t$, where y represents the total number of storage tanks.
- $E_{t+\Delta t}$ output vector of values representing the energy consumption, in kilowatts, of the pumps between time t and $t + \Delta t$.

A set of inputs/outputs that map the behaviour of the network for several diverse input scenarios has to be created. In this work, the hydraulic models are used to replace real observations. This is due to the difficulty of obtaining a large amount of reliable data from the water supply systems. Therefore, virtual observations are used instead of real ones. The networks can be accurately emulated using the EPANET software [1] or mathematical models that use differential calculus.

Before starting the training procedure, the data is normalized and split into 3 sets: train, test and validation. The training set is composed of 80% of the total data, 19200 inputs, the test set has 20% - 24 samples. Those remaining 24 samples that compose the validation set. The starting indices of each split are randomly permutated.

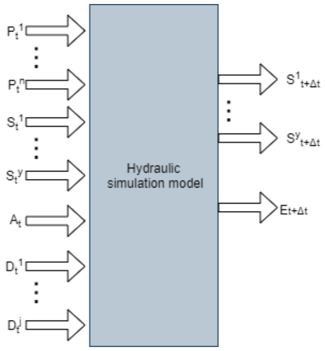


Fig. 1: Representation of the input - output variables of the model.

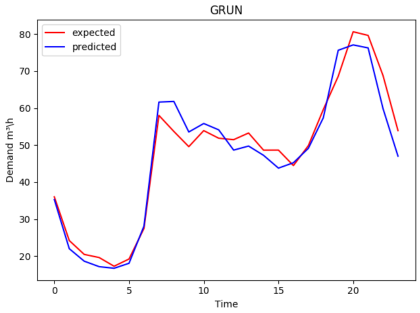
4 Results and Discussion

4.1 Water demand forecasting

A 24 h scenario was selected to test the performance of the models. Figure 2 show the expected, depicted by the red line, and the predicted water demand, depicted by the blue line. Additionally, Table 1 shows the scores from various model evaluation metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) [5]. As expected, the ANN results demonstrate that it able to predict water demands with a small error margin and it is able to keep up with the several variations of the water demand along the day. The GRUN model has a similar performance (see Figure 2 and Table 1), however, this comes at a cost of a much higher computational cost. Furthermore, the GRUN's performance has demonstrated to exceptionally vary depending on outcome of its training, making the model not as stable as the ANN.

	ANN	**GRUN**
RMSE	3,5967	3,8146
MAE	2,7938	2,8395
R^2	0,9601	0.9552

Table 1: ANN's and GRUN's RMSE, MAE and R^2 metric water forecasting scores on a scenario of 24 h with 1 h time-step.



(a) (b)

Fig. 2: Forecasting results for both the ANN and GRUN model on a 24 h scenario with a 1 h time-step.

4.2 Hydraulic simulation

Two networks were taken into consideration in order to test the capabilities of both the water demand forecasting and water network simulation methodologies. Fontinha and Richmond, presented in Figures 3a and 3b respectively.

In order to generate data to train the ANN models, an analytical hydraulic model was used in the case of Fontinha, and a numerical hydraulic model simulated using EPANET for Richmond's. The simulation models were specifically optimized for both networks using a grid search method that exhaustively tested several combinations of parameters. The optimization results are demonstrated in Table 2.

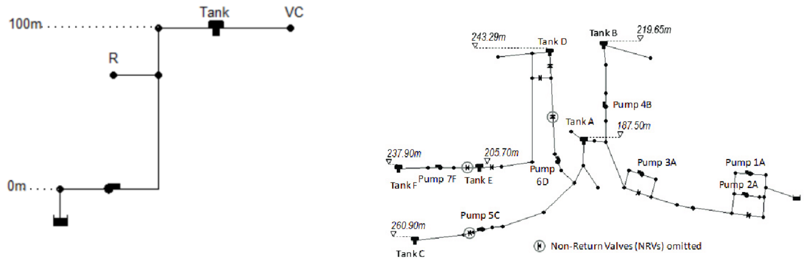
After the construction, optimization and training the ANNs can be validated. A 24 h scenario is chosen from the sample dataset to examine the performance of the models.

Figure 4 demonstrates both the expected and the predicted energy and tank level values of Fontinha's model. The red and blue line represent the expected and simulated values, respectively. There results demonstrate that the model was able to reproduce the behaviour of the network.

Richmond is a much more complex network, as such, the results are expected to be somewhat worse than Fontinha. Figure 5 displays the several outputs of Richmond's simulation model. A slight discrepancy between the expected and the predicted values is noticeable. Even so, the difference between the two values are very minor and does not exceed 10 cm in the worst case.

After the construction, optimization and training the ANNs can be validated. A 24 h scenario is chosen from the sample dataset to examine the performance of the models.

Figure 4 demonstrates both the expected and the predicted energy and tank level values of Fontinha's model. The red and blue line represent the expected and simulated values, respectively, and this result demonstrates that the model was able to faithfully recreate the behaviour of the network.

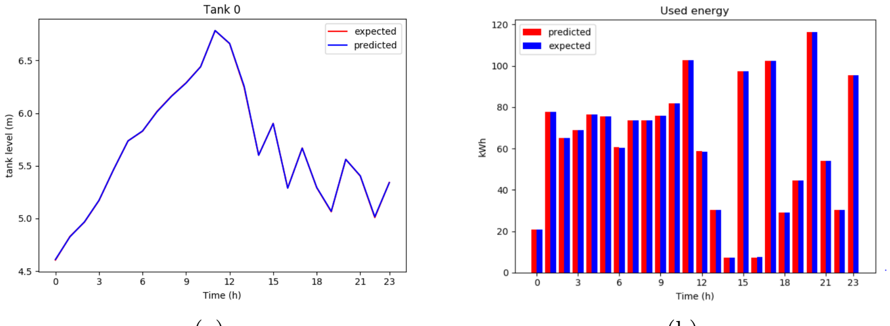


(a) (b)

Fig. 3: Schematic representations of Fontinha's and Richmond's [12] network.

	Fontinha's model	Richmond's model
Number of hidden layers	1	
Number of nodes	15	22
Batch size	50	100
Learning rate (constant)	0.002	0.0025
Activation	ReLU	
Optimizer	Adam	

Table 2: Optimization results for the simulation models of Fontinha and Richmond.



(a) (b)

Fig. 4: Fontinha's simulation model outputs on a 24 h scenario.

	Fontinha's Model		**Richmond's model**	
	Tank level (m)	**Energy (kW)**	**Tank level (m)**	**Energy (kW)**
RMSE	0,0036	0,0708	0,0312	0,2314
MAE	0,0030	0,0570	0,0247	0,0927
R^2	0,9999	0,9999	0,9909	0,9869

Table 3: RMSE, MAE and R^2 metric scores for the tank level and energy simulation of the both Fontinha's and Richmonds models on a 24 h scenario.

Richmond is a much more complex network, as such, the results are expected to be somewhat worse than Fontinha. Figure 5 displays the several outputs of Richmond's simulation model. A slight discrepancy between the expected and the predicted values is noticeable. Even so, the difference between the two values are very minor and does not exceed 10 cm in the worst case.

In addition to the several outputs of the models for the 24 h scenarios, the results from various model evaluation metrics, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) [5], can be consulted in Table 3. The metric scores shows that Fontinha's simulation model is able to predict the tank level with a millimetric error while the energy's error is in the hundredths of a kilowatt. Richmond's simulation model is able to predict the tank levels with an error measured in centimetres, and the energy with an error between the tenths and the hundredths of a kilowatt.

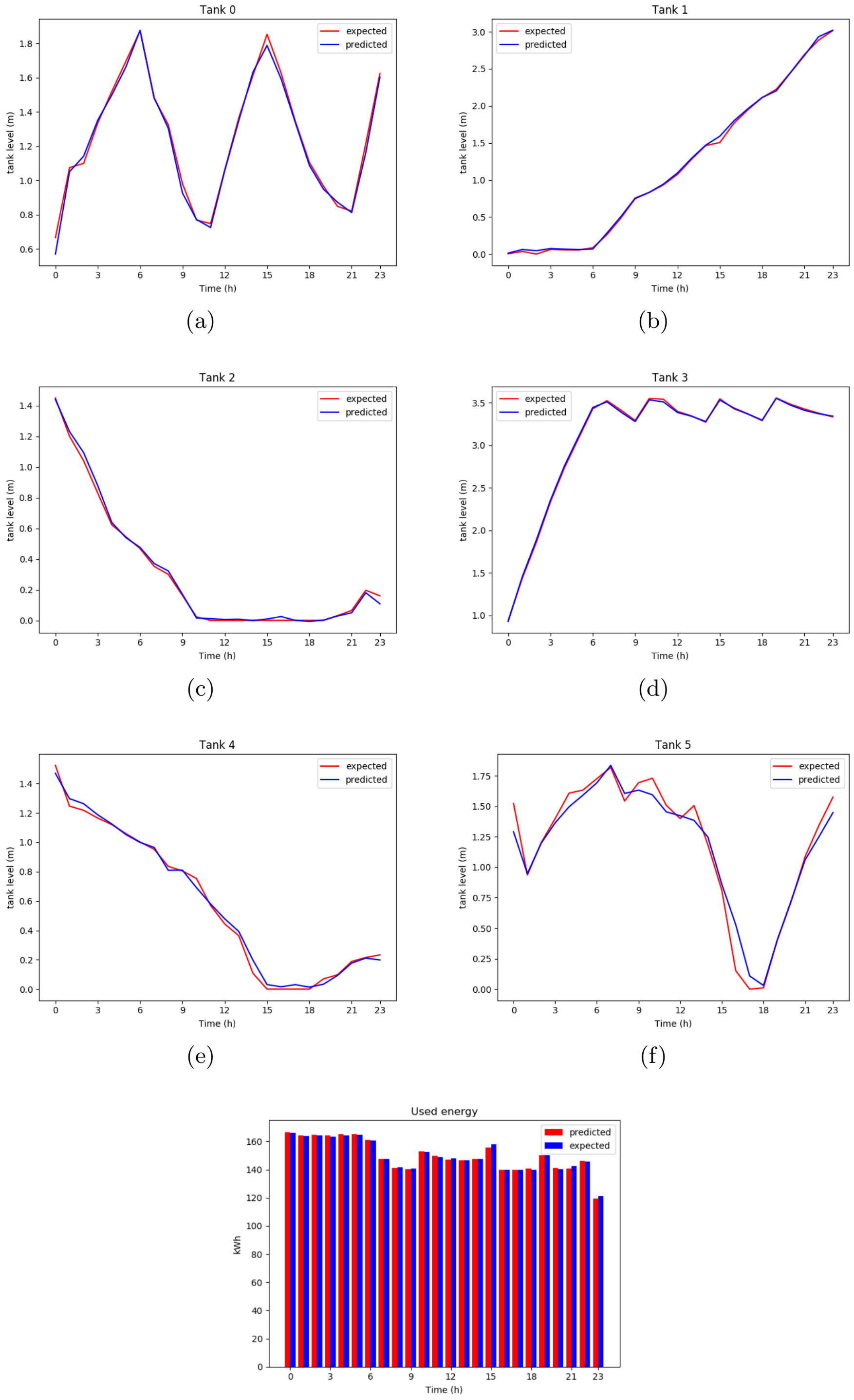
5 Conclusion

Given the results presented in this work, the ANN model already offers a reliable solution to forecast water demands, while the GRUN model, despite having similarly good results, requires much more computational effort and its complexity make it somewhat inconsistent. In this context, the standard ANN seems to be the better solution. The good performance displayed by both models is due in part to the way input data is arranged. Each demand that took place $24n$ hours prior to a certain timestep is considered as a feature. This, not only solves the problem of forecasting a whole operational horizon (*e.g* 24 h), but it also captures the seasonality of each hour of the day.

Regarding the hydraulic simulation, it has been shown that ANNs are also capable of thoroughly mimicking the behaviour of water distribution networks of various sizes and complexities. ANNs provide a computational efficient solution that is capable of learning by itself the behaviour of the network without requiring a tedious and prolonged calibration process, thus, eliminating the need to use hydraulic simulators.

Acknowledgements

This work was partially supported by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UID/CEC/00127/2019.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 5: Richmond's simulation model outputs on a 24 h scenario.

References

1. Agency, U.S.E.P.: Epanet, https://www.epa.gov/water-research/epanet
2. Antunes, A., Andrade-Campos, A., Sardinha-Lourenço, A., Oliveira, M.: Short-term water demand forecasting using machine learning techniques. Journal of Hydroinformatics **20**(6) (2018). https://doi.org/10.2166/hydro.2018.163
3. Bunn, S.M., Reynolds, L.: The energy-efficiency benefits of pump-scheduling optimization for potable water supplies. IBM Journal of Research and Development **53**(3), 5:1–5:13 (2009). https://doi.org/10.1147/JRD.2009.5429018, http://ieeexplore.ieee.org/document/5429018/
4. Byrd, R., Peihuang, L., Nocedal, J.: A limited-memory algorithm for bound-constrained optimization. Tech. rep., Argonne National Laboratory (ANL), Argonne, IL (mar 1996). https://doi.org/10.2172/204262, http://www.osti.gov/servlets/purl/204262-FVeKR4/webviewable/
5. Dodge, Y.: Coefficient of Determination. In: The Concise Encyclopedia of Statistics, pp. 88–91. Springer New York, New York, NY (2008). https://doi.org/10.1007/978-0-387-32833-1_62, http://www.springerlink.com/index/10.1007/978-0-387-32833-1_62
6. Guancheng, G., Shuming, L., Yipeng, W., Junyu, L., Ren, Z., Xiaoyun, Z.: Short-Term Water Demand Forecast Based on Deep Learning Method. Journal of Water Resources Planning and Management **144**(12), 4018076 (2018). https://doi.org/10.1061/(ASCE)WR.1943-5452.0000992, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000992
7. Jordan, M., Kleinberg, J., Scho, B.: Pattern Recognition And Machine Learning. Springer (2006)
8. Lindeil E., O., Kevin E., L.: Optimal control of water supply pumping systems. J. Water Resour. Plann. Manage **2**(10-11), 237–252 (1994). https://doi.org/10.1016/0042-207X(85)90371-9
9. Rao, Z., Alvarruiz, F.: Use of an artificial neural network to capture the domain knowledge of a conventional hydraulic simulation model. Journal of Hydroinformatics **9**(1), 15 (2007). https://doi.org/10.2166/hydro.2006.014, http://jh.iwaponline.com/cgi/doi/10.2166/hydro.2006.014
10. Salomons, E., Goryashko, A., Shamir, U., Rao, Z., Alvisi, S.: Optimizing the operation of the Haifa-A water-distribution network. Journal of Hydroinformatics **9**(1), 65 (2007). https://doi.org/10.2166/hydro.2006.018, http://jh.iwaponline.com/cgi/doi/10.2166/hydro.2006.018
11. Walski, T.M., Brill, E.D., Gessler, J., Goulter, I.C., Asce, A.M., Jeppson, R.M., Asce, M., Lansey, K., Lee, H.l., Members, S., Liebman, J.C., Mays, L.: Battle of the network models: Epilogue. Journal of Water Resources Planning and Management **113**(2), 191–203 (1987)
12. Zyl, J.E.V., Savic, D.a., Walters, G.a.: Operational optimization of Water Distribution Systems using a hybrid Genectic Algorithm. Journal of Water Resources Planning and Management **130**(2), 160–170 (2004)