Predicting daily water tank level fluctuations by using ARIMA model. A case study

S Mancini¹, A B Francavilla², A Longobardi², G Viccione²ᵃ and C Guarnaccia²ᵇ

¹ Department of Information and Electric Engineering and Applied Mathematics, University of Salerno, via Giovanni Paolo II 132, 84084, Fisciano, SA, Italy
² Department of Civil Engineering, University of Salerno, via Giovanni Paolo II 132, 84084, Fisciano, SA, Italy

E-mail: ⁡ᵃgviccion@unisa.it, ⁡ᵇcguarnaccia@unisa.it

Abstract. The intrinsic dynamical features of water demand highlight the need of proper operational management of tanks in water distribution networks. In addition, due to the water resource scarcity, sustainable management of urban systems is essential. For this purpose, the aid of a predictive model is crucial since it allows to give short term forecasts that can be used to predict the oscillations of relevant parameters, i.e. tanks level and/or water demand. Urban water managers can use these predictions to implement actions aimed at the optimisation of the network function. Among several modelling techniques, the univariate time series analysis is instrumental since it allows forecasting the studied parameter by using the measurements of the parameter itself. In this paper, an autoregressive integrated moving average (ARIMA) model is calibrated on water levels data, measured in an urban tank in Benevento, Campania region (Italy) and then tested on a large dataset not used to tune the parameters. The validation and forecast phases show good performances of the model on a short-term forecast horizon demonstrating the excellent potentiality of this techniques. Finally, the residuals and errors analysis complete the work suggesting possible future implementations and improvements of this technique.

1. Introduction

Water supply and distribution networks belong to the ensemble of critical infrastructures, essential systems for maintaining health, safety, economic and social well-being of citizens. As a primary objective, the water distribution network (WDN) service is oriented to meet the user water demand, providing the user itself a reliable water distribution, that is providing the user with an adequate quantity of water, of good quality and with adequate pressure [1]. The management of water distribution networks (WDNs) relies on water utility operations consisting of both quick responses to either water demand or water tanks and reservoirs recharge variations or long-term management of network ageing effects [2,3]. In this context, storage tanks play a key role, actually acting as lungs [4] - that is, balancing instantaneous flow variations in the water demand pattern as well as compensating abrupt interruptions of the water recharge to the storage tank, such as in cases of drought periods or electricity shortages in pumping stations delivering water, when the tank level fluctuates within a fixed range of levels. Water tank level prediction and forecasting are crucial steps for supporting decision-making regarding operating actions as they represent the pressure heads that affect the reliable WDN operation. Water tank levels can be modelled through hydraulic models when the water demand and the management rules and operations are known. Still, in practical applications, the latter facts are not
always fully known [5,6]. From a “modellistic” point of view, the AutoRegressive Integrated Moving Average (ARIMA) typology of models is well established, having been applied in the field of water demand forecasting for a long time [4,5,7-9]. This is justified by the fact that the model follows the trend at different time scales. On the other side, ARIMA models are also particularly suitable to predict and forecast groundwater levels, typically representing one of the primary sources of reservoir and tanks recharge [10]. Despite the number of applications in urban water demand, there is a gap in the literature concerning the use of ARIMA models for tank water levels. In [11], the link between the water supply, consumer demand and water level at the tank is, however, discussed, with the aim of providing a practical tool for water utilities to take prompt action based on water level variations.

In this paper, a univariate time series of water levels recorded in a urban tank is analysed. The data will be aggregated on one hour basis and an ARIMA modelling technique will be implemented, to produce short term predictions of the hourly water tank levels. The paper is organised in a “Materials and methods” section, in which the theoretical background of the ARIMA models is summarised, together with the case study and dataset presentation. Then, results of the application of the model are presented and discussed extensively, including the exploratory data analysis, the selection and calibration of the best ARIMA model, and the validation and the forecast testing phase. Since the selected model provides a one-step-ahead prediction, the forecast analysis will be performed in a 24-48 hours range. Quantitatively conclusions will be drawn analysing the residuals and the errors of the model, respectively in calibration and testing dataset. The definition of an ARIMA model applied to water levels would allow the prediction and forecasting of water tank levels for hydraulic WDN modelling as the baseline for supporting decision making regarding operating actions. Additionally, it would allow the identification of proper water tank management, improving drinking water waste reduction and drinking water saving, on one hand, and treatment costs related to chlorination or purifying techniques reduction on the other hand. This is an aspect of paramount importance as the circumstance in which the tank is not able to serve due to water scarcity is not rare, whereas there is a waste of the resource when the water inflow is not controlled [3].

2. Materials and methods

In this paper, the Box-Jenkins/ARIMA forecasting model is implemented. Autoregressive integrated moving average (ARIMA) was proposed by Box and Jenkins [12] in 1976. The key components of the ARIMA model are the autoregressive or AR component, the differencing component and the moving average or MA component applying the Box-Jenkins method for finding the best fit coefficients. [13]

The order of an ARIMA model is usually denoted by the following notation ARIMA(p,d,q), where p, d and q are the order of the autoregressive part, the differencing and the moving-average process, respectively. The following equation expresses the general source formula:

\[ \phi_p(B)(1 - B)^d Y_t = \theta_q(B) e_t \]  

(1)

\( Y_t \) : value of the series observed at the time t,
\( B \): delay operator,
\( \phi_p \) : autoregressive polynomials,
\( \theta_q \) : moving average polynomials
\( e_t \) is the difference between the observed value \( Y_t \) and the forecasted one \( \hat{Y} \) at the time t.

Any seasonal component is not supposed to be considered when using equation (1).

The general methodology implemented in this paper is summarised in Table 1. It is composed by four main steps. The preliminary phase aims to analyse the dataset to numerically and graphically describe it, in terms of data types, outliers, missing values, distributions. Then, the selection of the best ARIMA model is determined according to either Akaike Information Criterion (AIC), Akaike Information Criterion corrected (AICc) or Bayesian Information Criterion (BIC).

At least, the model validation phase to verify the forecasting abilities of the proposed model and the forecasting are performed. The entire process has been implemented in R software [14].
ARIMA modelling in ‘R’ uses a variation of the Hyndman-Khandakar algorithm [15], minimizing the standard and information criteria above mentioned to obtain the best ARIMA model. In particular, the auto.arima function provides hyperparameters and parameters of the best fitting model, using the conditional-sum-of-squares to find starting values, then the maximum likelihood for finding the best values. Then, residual tests need to be performed in order to examine the appropriateness of the model. So, the values of the autocorrelation and partial autocorrelation functions (ACF/PACF) of the residual series must all be approximately near to zero. If they do not look like white noise, another model should be tried. Once the residuals look like white noise, forecasting can be processed. [16]

Table 1. Workflow of the applied methodology.

| n. | Phase                  | Description                                                                 |
|----|------------------------|------------------------------------------------------------------------------|
| 0  | exploratory data analysis | Dataset analysis and summary of their main characteristics                  |
| 1  | model calibration       | Hyperparameters tuning and best ARIMA model selection, according to the information criteria |
| 2  | model validation        | Test of selected model on a dataset not used in the calibration (fixed parameters but observed data used as input of the model) |
| 3  | model forecasting       | Forecast with the selected model and comparison with observed data (parameters fixed and no observed data in input) |

2.1. Case study and dataset presentation
This study is applied on the time series of the levels observed in one of the tanks of the water supply system of the town of Benevento, Italy, located in the area of Gesuiti. Authors already analysed data from this site in a previous paper where a preliminary study, focused on the use of a stochastic model to fill big sequences of missing data, and the calibration of a model have been described [6].

In this paper, data refer to daily average water levels, expressed in meters, registered from 2 p.m. on 6 September 2018 to 11 p.m. on 2 April 2019. The original dataset has been divided into two subsets: the first 4330 values were used to perform the calibration and, therefore, the estimation of the model parameters (calibration dataset) and the last 672 data were used for the testing (including the validation and forecasting phase). The time plot of the dataset is reported in Fig. 1.

As already graphically revealed by the time plot, the dataset shows no relevant gaps. Indeed, only 13 data are randomly missing for short time ranges. Thus, the gaps were filled by using the mean value among the nearly previous and subsequent data value registered. Just in one case, when 6 hours

**Figure 1.** Time plot of the data with evidence of the subset of calibration (blue) and testing (green).
measurements were missing, the data were imputed with the mean of the observed levels at the same hours of previous and subsequent days. The description of the final data subset is given in Table 2.

Table 2. Description of the calibration and testing dataset.

| Dataset  | From          | To            | Sample size | Missing values | Imputed data |
|----------|---------------|---------------|-------------|----------------|--------------|
| Calibration | September 6, 2019 | March 5, 2020 | 4330        | 13             | 0.3          |
| Testing   | March 6, 2020  | April 2, 2020 | 672         | 0              | 0            |

3. Results and Discussion

Following the workflow presented in Table 1, the first step of the study is an exploratory data analysis that is performed on the original dataset, as well as on the calibration and testing ones, to quantitatively and qualitatively describe the main features.

Table 3 reports a preliminary descriptive analysis of the datasets, including the main statistics of the data. It can be noticed that the calibration and the full datasets have similar distributions. The same conclusion can be drawn looking at the histograms and boxplots in Figure 2, and the correlograms and partial autocorrelation functions in Figure 3. The testing dataset differs from the calibration and the full ones because it missed the left tail (water tank levels below 3 m).

![Figure 2](image1.png)

**Figure 2.** Histograms (left) and boxplots (right) of the data for a) full dataset, b) calibration dataset and c) testing dataset.
Table 3. Summary statistics of the datasets of the water levels observed in the tank of Gesuiti.

| Dataset    | Mean [m] | Std. Dev. [m] | Median [m] | Min [m] | Max [m] | Skewness | Kurtosis |
|------------|----------|---------------|------------|---------|---------|----------|----------|
| Calibration | 4.32     | 0.60          | 4.37       | 1.54    | 5.36    | -0.53    | 0.41     |
| Testing    | 4.12     | 0.55          | 4.10       | 3.07    | 5.36    | 0.18     | -0.74    |
| Total      | 4.30     | 0.60          | 4.34       | 1.54    | 5.36    | -0.43    | 0.18     |

Figure 3. Autocorrelation plots (left) and Partial autocorrelation plots (right) of the data for a) full dataset, b), calibration dataset and c) testing dataset.

3.1. Model selection and calibration of hyperparameters
The best ARIMA model has been chosen according to the auto.arima function implemented in the “forecast” package of R software [14]. This function compares different models according to the common information criteria, such as Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC). Running on the calibration dataset, this function suggested an ARIMA (4,1,1) model. The predicting formula of this model is:
\[ \hat{Y}_t = \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + \phi_3(Y_{t-3} - Y_{t-4}) + \phi_4(Y_{t-4} - Y_{t-5}) + \theta_1 e_{t-1} \]  

(2)

This model requires in input the previous observations thus it is a one-step-ahead model. The model parameters were estimated using again the statistical program R and the results are reported in Table 4.

|       | AR1, $\phi_1$ | AR2, $\phi_2$ | AR3, $\phi_3$ | AR4, $\phi_4$ | MA1, $\theta_1$ |
|-------|---------------|---------------|---------------|---------------|-----------------|
| Estimated Value | 1.88 | -1.10 | 0.18 | -0.03 | -0.91 |
| Standard error | 0.017 | 0.032 | 0.032 | 0.016 | 0.008 |

The plot of the simulated levels (red dots) overlapped with the observed data (black solid line) is reported in Figure 4. The single point simulated in period 536, i.e. hour 9 p.m. of September 28, 2018, completely misses the prediction. The prediction is 0.13 m and comes after a range of 6 data that were imputed because of missing observations in the raw dataset. The measurements were probably missing because of a maintenance of the tank, which caused a lowering of the levels. This oscillation was reproduced by the model, with a one-step delay, as shown in Figure 5, in periods 535 and 536.

Figure 4. Comparison between observed (solid black line) and simulated (red dots) water tank levels plotted versus time in the calibration dataset.

Figure 5. Comparison between observed (solid black line) and simulated (red dots) water tank levels plotted versus time in the 500-550 periods range.

3.2. Residuals analysis
In order to quantitatively evaluate the model’s performances, the analysis of the difference between observed and simulated water tank levels in the calibration phase (residuals) has been executed. The results are reported in Table 5 (summary statistics) and in Fig. 6 (slope, autocorrelation and histogram).

Table 5. Summary statistics of the residuals of ARIMA (4,1,1) model in the calibration phase.

| Mean [m] | Std. Dev. [m] | Median [m] | Min [m]  | Max [m]  | Skewness | Kurtosis |
|---------|---------------|------------|----------|----------|----------|----------|
| 0.00    | 0.05          | 0.00       | −1.60    | 1.41     | −2.15    | 298.72   |

It can be noticed that apart from the oscillation observed in the periods 535 and 536 for the reasons mentioned above, the residuals are very close to zero. This is confirmed by the frequency distribution that is very narrow and with basically a zero mean. The autocorrelation presents a significant peak in correspondence of a 24 hours lag. This suggests that a seasonal model could be implemented in a further analysis [17].

Figure 6. Time slope, autocorrelation plot and frequency distribution of the residuals in the calibration dataset.

3.3. Model testing: validation and forecast

Once the model is calibrated and the residuals have been analysed, the model can undergo a testing phase, i.e. a comparison between results of the models and measured data not used in the calibration phase. The key point of this phase is to check if the model has a good accuracy once the parameters have been fixed. The difference between validation and forecast subphases is in the usage of observed data in input of the prediction formula: during the validation the theoretical assumption is that the observed data are available and can be used to obtain the simulations (but not to tune the parameters). On the contrary, in the forecast, the observed data cannot be used to produce simulations because the model is predicting values in the future, but they can be compared with forecasts for performances analysis.
It is worth underlining that ARIMA models, adopting the last observed data to compute the predictions, have a short forecast horizon. The ARIMA (4,1,1) selected in this work has a one step ahead prediction and, thus, the forecast tends to a constant value, according to what suggested in TSA literature. In facts, when predicting water tank levels in the future, the model is fed by simulations and it lacks the variability of the observed data.

As already stated above, the ARIMA model was validated on the 672 periods of the testing dataset. The overlap between observed, simulated (in validation) and forecasted water tank levels is reported in Figure 7 for the entire validation dataset. It can be immediately noticed that the validation confirms the goodness of the model when working with observed data in input. On the contrary, the forecast tends to a constant value after about 48 hours. For this reason, a zoom on the first 24 hours is reported in Figure 8. Also, the error analysis presented in the following has been restricted to a two-day window.

![Figure 7](image1.png)

**Figure 7.** Comparison between observed (black), simulated in validation (red) and forecasted (blue) water tank levels plotted versus time in the testing dataset.

![Figure 8](image2.png)

**Figure 8.** Comparison between observed (black), simulated in validation (red) and forecasted (blue) water tank levels plotted versus time in the first 24 hours of the testing dataset.

### 3.4. Error analysis in validation and forecast
The performances of the models are tested in the validation and forecast phases by means of error analysis. The summary statistics of the validation and forecast errors are reported in Table 6. It can be noticed that the statistics of the errors are extremely good in the validation phase. In the 48 hours forecast, the mean error is reasonable, as well as the standard deviation. Figures 9 and 10 report the errors versus time, the autocorrelation and the frequency distribution, respectively for the validation and the forecast on 48 periods. The validation errors plots confirm the good performances already shown in the errors statistics, in fact the errors are almost normally distributed, with a narrow distribution and with an almost constant time slope. The autocorrelation plot again suggests that part of the variance of the data has not been explained and that a seasonal model could be an option for this dataset. The forecast errors, instead, have an oscillation over the 48 hours window, due to the convergence to a constant value of the forecast, which leads to the variation of the error.

Table 6. Summary statistics of the errors of ARIMA (4,1,1) model in the testing phase.

|                      | Mean [m] | Std. Dev. [m] | Median [m] | Min [m] | Max [m] | Skewness | Kurtosis |
|----------------------|----------|---------------|------------|---------|---------|----------|----------|
| Validation (672 periods) | 0.00     | 0.04          | 0.00       | -0.15   | 0.15    | -0.24    | 0.59     |
| Forecast (48 periods)   | -0.27    | 0.36          | -0.36      | -0.77   | 0.37    | 0.27     | -1.23    |

Figure 9. Time slope, autocorrelation plot and frequency distribution of the errors in validation.
Figure 10. Time slope, autocorrelation plot and frequency distribution of the errors in the forecast.

4. Conclusions
In this paper, the modelling and forecasting of the time series of water levels in a tank connected to the supply system of a city has been discussed.

The univariate time series is related to the daily measurements of the water levels, observed from September 2018 to April 2019 (5002 hours), in the tank of Gesuiti site, connected to the urban water system of the city of Benevento (Italy). Only few measurements (13 periods) were missing, probably due to maintenance operations in the tank. These missing data have been reconstructed imputing the average value of the two temporally closest measurements, one previous and one following the missing period. In this way, a continuous time series of 5002 periods was obtained.

Once statistically characterised the time series, the best ARIMA model has been chosen, according to the main information criteria. In particular, an ARIMA (4,1,1) has been implemented.

The statistical analysis showed that the series of levels exhibits evident periodicity or seasonality, which could be used to construct a seasonal model (lag = 24 hours) as future research development.

The modelling results were very encouraging both by graphically comparing the observed level with the simulated one, and by analysing the distribution of the forecast error from a quantitative point of view. The selected model provided good predictions, even if limited to a two-day forecast horizon. The use of these techniques on water tank levels data shows, as already demonstrated in previous papers, excellent potential, presenting a possible way for the integration between the current water service monitoring systems and the forecasts obtained with this easy and simple-to-implement methodology.

References
[1] Longobardi A., G. Fattoruso, G. Guarnieri, A. Di Pietro, L. La Porta, M. Pollino. Water distribution network perspective in RAFAEL Project, a system for critical infrastructure risk analysis and forecast. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Proceedings of the ICCSA 2021, International Conference on Computational Science and its Applications, RRS2021 Workshop, Cagliari, 13-16-Settembre 2021
[2] Bello, O. Abu-Mahfouz, A.M., Hamam, Y., Page, P.R., Adedeji, K.B., Piller, O. Solving management problems in water distribution networks: A survey of approaches and mathematical models. Water 11, 2019, 562, doi:10.3390/w11030562

[3] Viccione, G., Ingenito, L., Evangelista, S., Cuozzo, C. Restructuring a water distribution network through the reactivation of decommissioned water tanks. Water 11, 2019, 1740, doi:10.3390/w11091740

[4] Tripathi, A., Kaur, S., Sankaranarayanan, S., Narayanan, L.K., Tom, R.J. Water demand prediction for housing apartments using time series analysis. Int. J. Intell. Inf. Technol. 2019, 15, 57–75, doi:10.4018/IJIIT.2019100104

[5] Guarnaccia, C.; Tepedino, C.; Viccione, G.; Quartieri, J. Short-Term Forecasting of Tank Water Levels Serving Urban Water Distribution Networks with ARIMA Models. In Frontiers in Water-Energy-Nexus—Nature-Based Solutions, Advanced Technologies and Best Practices for Environmental Sustainability; Advances in Science, Technology & Innovation Naddeo, V., Balakrishnan, M., Choo, K.H., Eds.; Springer: Cham, Switzerland, 2018; doi:10.1007/978-3-030-13068-86

[6] Guarnaccia C., Longobardi A., Mancini S., Viccione G. Drinking Water Tank Level Analysis with ARIMA Models: A Case Study. In: V. Kanakoudis, The 4th EWA International Conference: Valuing the Water, Carbon, Ecological Footprints of Human Activities, 2020, 1-8

[7] Muhammad, A.U., Li, X., Feng, J. Artificial Intelligence Approaches for Urban Water Demand Forecasting: A Review. In: International Conference on Machine Learning and Intelligent Communications Zhai X., Chen B., Zhu K., Eds.; Springer: Cham, Switzerland, 294, 2019, pp. 595–622, doi:10.1007/978-3-030-32388-2_51

[8] Zhao, L., Zhang, J., Chen, T. Application of product seasonal ARIMA model to the forecast of urban water supply. J. Water Resour. Water Eng. 2011, 22, 58–62

[9] Billings, R.B., Jones, C.V. Forecasting Urban Water Demand; American Water Works Association: Denver, 2008, CO, USA. ISBN 978-1-58231-537-1

[10] Boulariah O., Mikhailov P.A., Longobardi A., Elizariev A.N., Aksenov S.G. Assessment of prediction performances of stochastic models: Monthly groundwater level prediction in Southern Italy. J. Groundw. Sci. Eng., 2021, 9(2), 161-170. doi: 10.19637/j.cnki.2305-7068.2021.02.008

[11] Viccione, G., Guarnaccia, C., Mancini, S., Quartieri, J. On the use of ARIMA models for short-term water tank levels forecasting. Water Supply 20, 2020, 787–799, doi:10.2166/ws.2019.190

[12] Box, G., Jenkins, G. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day, 1976, CA,USA

[13] Khanarsa, O., Luangsodsa, A., Sinapiromsaran K. Self-Identification ResNet-ARIMA Forecasting Model. WSEAS transactions on systems and control, 2020,15, 196-211. ISSN: 2224-2856

[14] R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2014. Available online: http://www.R-project.org/ (last accessed on May 3, 2021)

[15] Hyndman, R. J., & Khandakar, Y. Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 2008, 27(3), 1–22.

[16] Merabet, F., Zeghdoud H. On Modelling seasonal ARIMA series: Comparison, Application and Forecast (Number of Injured in Road Accidents in Northeast Algeria), WSEAS transactions on systems and control, 2020, 15, 235-246. ISSN: 2224-2856

[17] Guarnaccia, C., Quartieri, J. Tepedino, C. Deterministic decomposition and seasonal ARIMA time series models applied to airport noise forecasting, in AIP Conference Proceedings 1836, 2017, 020079, doi: 10.1063/1.4982019