engineering proceedings

MDPI

Proceeding Paper

Probabilistic Forecasting of Hourly Water Demand †

Panagiotis Kossieris [1,*], Ioannis Tsoukalas [1,2], Dionysios Nikolopoulos [1], Georgios Moraitis [1] and Christos Makropoulos [1]

[1] Department of Water Resources, School of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-15780 Zographou, Greece; itsoukal@civil.duth.gr (I.T.); nikolopoulosdio@central.ntua.gr (D.N.); georgemoraitis@central.ntua.gr (G.M.); cmakro@mail.ntua.gr (C.M.)
[2] Department of Civil Engineering, Democritus University of Thrace, Kimmeria Campus, GR-67100 Xanthi, Greece
* Correspondence: pkossier@mail.ntua.gr
† Presented at the 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024), Ferrara, Italy, 1–4 July 2024.

Abstract: Timeseries forecasting holds a prominent position in the domain of urban water systems. Most forecasting approaches are designed to provide single-point deterministic forecasts, neglecting the uncertainty in model predictions. In this work, we propose a methodological framework, able to provide probabilistic predictions over lead times of operational interest, by combining machine learning (ML) methods with multivariate statistics (i.e., copulas). The idea is that ML methods can be used to provide deterministic forecasts, and copulas can be used to quantify the predictive uncertainty of the forecasts. We showcase the effectiveness of proposed framework using hourly water demand data from a real-world case study.

Keywords: water demand; probabilistic forecasting; machine learning; copulas; uncertainty quantification; random forest; long short-term memory (LSTM) neural network

Citation: Kossieris, P.; Tsoukalas, I.; Nikolopoulos, D.; Moraitis, G.; Makropoulos, C. Probabilistic Forecasting of Hourly Water Demand. *Eng. Proc.* **2024**, *69*, 100. https://doi.org/10.3390/engproc2024069100

Academic Editors: Stefano Alvisi, Marco Franchini, Valentina Marsili and Filippo Mazzoni

Published: 10 September 2024

1. Introduction

The primary objective of urban water systems (UWSs) is to provide water of adequate quantity and quality to consumers. Effectively operating and managing UWSs towards this objective, reliable, and accurate water demand forecasting, both in short- and long-term horizons, is essential. Specifically, accurate water demand forecasts at fine temporal scales, e.g., hourly, are important for real-time control (e.g., optimal pump and valve actuation) and identification of potential water network failures (e.g., bursts) [1]. Nonetheless, it has been considered a difficult and still open problem to solve. In particular, short-term forecasting approaches have ranged from heuristics and regression models to machine learning methods, such as artificial neural networks [2], including complex architectures such as long short-term memory (LSTM)-based models [3], support vector machines, fuzzy and neuro-fuzzy models [4], random forests (RFs) [1], time series models, and hybrid approaches.

The majority of the aforementioned approaches typically neglect predictive uncertainty and address the problem under the typical deterministic time-series forecasting prism, which is focused on providing single-point forecasts [5]. However, providing probabilistic, and thus multi-point, forecasts of water demand is of paramount importance, since this could, on the one hand, encapsulate the uncertainties associated with the forecasting model per se (recall the classic quote of George Box, "*all models are wrong, but some are useful*") and, on the other hand, enhance operational aspects of UWSs by propagating such forecasting uncertainties into the decision-making process [5].

In the light of the above, this work provides a general modeling framework for probabilistic time series forecasting that builds upon, and uses forecasts from, existing deterministic models, thus capitalizing decades of research in the domain and also enabling

(1) the quantification of predictive uncertainty and (2) the generation of multi-point, probabilistic forecasts. The methodology adopts copulas to model the joint distribution among the forecasted and observed data and next transform the deterministic predictions into probabilistic/stochastic ones, via the derived conditional distribution.

2. Methodology

Copulas (see [6,7]) lie at the core of the suggested approach since they allow the construction of joint distributions (by independently modeling the dependence structure and the marginal distributions of the involved random variables), thus enabling the derivation of the conditional distribution of the predictand (e.g., observed water demand) given the forecast's predictions.

To elaborate, let X and Y denote two random variables (RVs), which correspond to the observed and forecasted quantities, respectively, while $F_X(x)$ and $F_Y(y)$ denote their cumulative distribution functions (cdf). According to copulas, their joint cdf can be expressed by $F(x,y) = P[X \leq 0, Y \leq 0] = C(F_X(x),\ F_Y(y))$, where $C(\cdot,\ \cdot)$ denotes the copula cdf; $u_X = F_X(x)$ and $u_Y = F_Y(y)$ are uniformly distributed in [0, 1].

The conditional cdf of the RV $X|Y = y$, that is $F_{X|Y=y}(x) = P[X \leq x|Y = y]$ can be obtained via $F_{X|Y=y}(x) = \partial C(F_X(x), F_Y(y))/\partial F_Y(y) = \partial C(u_X, u_Y)/\partial u_Y = C_{X|Y}(u_X|u_Y)$, where $C_{X|Y}$ stands for the so-called conditional copula, while the latter relationship can be inverted as follows, $u_X^{a|u_Y} = C_{X|Y}^{-1}(a|u_Y)$, in order to find the value of u_X that corresponds to a desired probability of non-exceedance $a := C_{X|Y}$ given the (known) value of $u_Y = F_Y(y)$ (compactly written as $u_X^{a|u_Y}$). Finally, to obtain the quantile that corresponds to that conditional probability level, the inverse cdf (icdf) of X, that is $F_X^{-1}(u)$, is employed. This operation reads as follows:

$$x_{a|u_Y} = F_X^{-1}\left(u_X^{a|u_Y}\right) \tag{1}$$

These expressions are general and can be used along with any bivariate copula, regardless of whether the copula has a direct expression for $C_{X|Y}$ and/or $C_{X|Y}^{-1}$. In the case of the Gaussian copula (used in this work), the parameter, say θ, can be identified via the Pearson correlation coefficient among X and Y, since it depends on their marginals; therefore, this is often called the equivalent correlation parameter (see [8,9]). Given the above, one may answer questions including the following:

Given that the estimate of the forecasting model is equal to y, what is the probability that the true (i.e., observed) value is smaller than x? The answer can be given by estimating the probability: $P\ [X\ \leq\ x|\ Y = y] = F_{X|\ Y=y}(x)$.

Given that the estimate of the forecasting model is equal to y, and for a given uncertainty level (say $a = 90\%$), what are plausible values for the upper and lower limits of true (i.e., observed) value (say x_U and x_L)? The answer to this question is given by estimating the quantities x_U and x_L (via Equation (1)) that correspond to the following probabilities. $\left[F_{X|\ Y=y}(x_L),\ F_{X|\ y=y}(x_U)\right] = [p_L,\ p_U\ ] = \left[\frac{1-a}{2}, \frac{1+a}{2}\right]$.

Finally, for the generation of multi-point forecasts for a given time step (i.e., an ensemble based on the available deterministic forecast y), it suffices to generate n (size of ensemble) random variables uniformly distributed in $[0,\ 1]$ and employ Equation (1).

3. Case Study and Results

To demonstrate the suggested approach, we used the hourly water demand data from the Battle of Water Demand Forecasting, organized in the context of the 3rd International WDSA-CCWI Joint Conference. Particularly, we employed data from DMA E, which is a residential/commercial district close to the city center. To this end, we build two different machine learning (ML) forecasting models, each forecasting water demand at different lead times (1 step ahead and 24 steps ahead).

The first forecasting model (lead time: 1 step ahead) implements a long short-term memory (LSTM) neural network architecture, with a proven capacity in time series forecast-

ing. We utilize an LSTM layer of 50 memory cells connected downstream to a single dense output layer to forecast water demand across DMAs. The LSTM employs 17 predictors, including water demands, meteorological parameters, and temporal features (month, day, and hour). The mean absolute error and Nash–Sutcliffe coefficient for the LSTM predictions (1 step ahead) for DMA E are 3.89 m^3/h and 0.87, respectively.

The second forecasting model (lead time: 24 steps ahead) is based on the widely known random forest (RF) bagging algorithm [10], which builds an ensemble of individually trained (de-correlated) decision trees. The RF model utilizes 300 decision trees, while for the other hyperparameters the default values are used. As predictors, we use the consumption lag-1, -2, -24 and -168 hourly consumption, meteorological predictors and temporal features (month, day, hour). The mean absolute error and Nash–Sutcliffe coefficient for the RF predictions (24 steps ahead) for DMA E are 11 m^3/h and 0.30, respectively.

In this work, we fitted and examined two copulas, in particular the Gaussian and Clayton copula. Furthermore, in all cases (both observed series and forecasted time series), we employed the Generalized Gamma distribution, fitted using the L-moments method. The results of the analysis are presented in the next two figures.

Particularly, Figure 1 depicts the quantification of predictive uncertainty at uncertainty level 90% for the LSTM (panel (a)) and RF (panel (b)) model, using the two types of copulas discussed above. For both models and copulas, the empirical coverage (i.e., the percentage of points that lie within the theoretical uncertainty level, in this case 90%) resembles with high accuracy the theoretical counterpart. Figure 2 provides a time series view of the methodology, illustrating that the method also acts as a bias adjustment technique, enabling much better matching between the observed and modeled values.

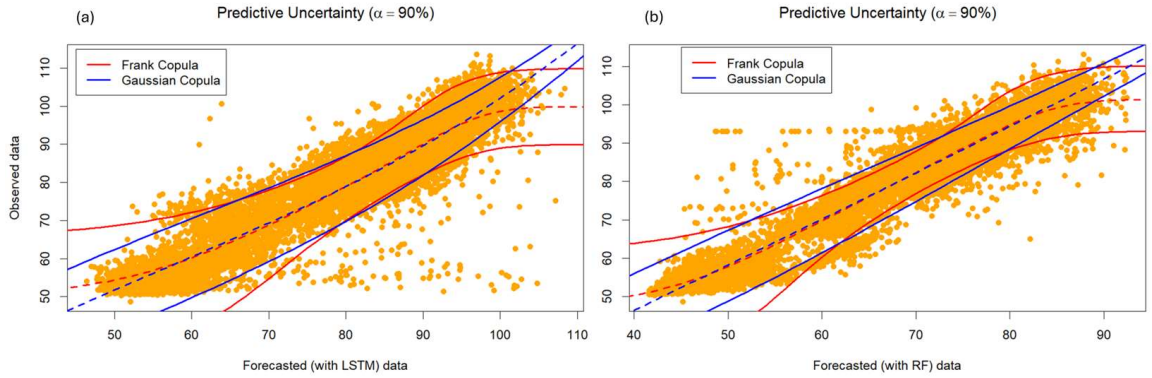

Figure 1. Quantification of predictive uncertainty ($a = 90\%$, thus $p_L = 0.05$ and $p_U = 0.95$) for (**a**) 1 step ahead LSTM forecasting model and (**b**) 24 step ahead RF forecasting model. In both cases, both the Gaussian and Clayton copula have been fitted to the data, while dashed lines represent the median values.

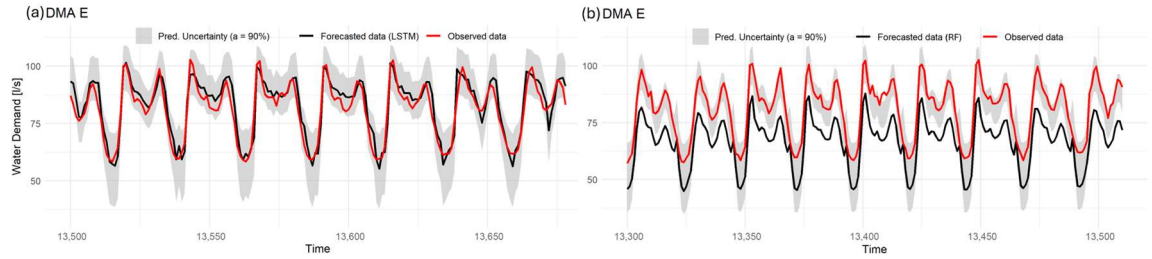

Figure 2. Comparison of observed and forecasted water demand time series for (**a**) LSTM (1 step ahead) and (**b**) RF (24 steps ahead) forecasting models. In both cases, the uncertainty intervals (gray bands) of $p_L = 0.05$ and $p_U = 0.95$ (i.e., $a = 90\%$) are depicted, which were estimated by the suggested copula-based method, via the Clayton copula.

4. Conclusions

The suggested copula-based approach allows the transformation of any deterministic forecasting model into a probabilistic one, and thus allows one to (a) derive the conditional distribution of the observed variable (e.g., water demand) given the value provided by the forecasting model (e.g., ML model) for each lead time (quantification of predictive uncertainty) and (b) generate ensembles of equiprobable forecasts, on the basis of the deterministic one. The results from the employed case study suggest that the copula-based approach provides an effective and efficient way to quantify the predictive uncertainty of deterministic forecasting models, regardless of their type (e.g., random forest, ANN, or any other) and the forecast's lead time (e.g., 1 step ahead or multi-step ahead).

Author Contributions: Conceptualization: P.K. and I.T.; methodology: I.T., P.K., G.M. and D.N.; software: P.K., I.T. and G.M.; formal analysis, P.K., I.T., D.N. and G.M.; investigation, P.K., I.T., D.N. and G.M.; resources, C.M.; writing—original draft preparation, P.K., I.T., D.N. and G.M.; writing—review and editing, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research work was partly supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "Third Call for HFRI PhD Fellowships" (Fellowship Number: 6349).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Herrera, M.; Torgo, L.; Izquierdo, J.; Pérez-García, R. Predictive Models for Forecasting Hourly Urban Water Demand. *J. Hydrol.* **2010**, *387*, 141–150. [CrossRef]
2. Alvisi, S.; Franchini, M.; Marinelli, A. A Short-Term, Pattern-Based Model for Water-Demand Forecasting. *J. Hydroinform.* **2006**, *9*, 39–50. [CrossRef]
3. Bakker, M.; Vreeburg, J.H.G.; van Schagen, K.M.; Rietveld, L.C. A Fully Adaptive Forecasting Model for Short-Term Drinking Water Demand. *Environ. Model. Softw.* **2013**, *48*, 141–151. [CrossRef]
4. Ghalehkhondabi, I.; Ardjmand, E.; Young, W.A.; Weckman, G.R. Water Demand Forecasting: Review of Soft Computing Methods. *Environ. Monit. Assess.* **2017**, *189*, 313. [CrossRef] [PubMed]
5. Donkor, E.A.; Mazzuchi, T.A.; Soyer, R.; Alan Roberson, J. Urban Water Demand Forecasting: Review of Methods and Models. *J. Water Resour. Plan. Manag.* **2012**, *140*, 146–159. [CrossRef]
6. Sklar, A. Random Variables, Distribution Functions, and Copulas: A Personal Look Backward and Forward. *Lect. Notes-Monograph Ser.* **1996**, *28*, 1–14.
7. Embrechts, P.; Lindskog, F.; Mcneil, A. Modelling Dependence with Copulas and Applications to Risk Management. In *Handbook of Heavy Tailed Distributions in Finance*; Elsevier: Amsterdam, The Netherlands, 2003; pp. 329–384.
8. Tsoukalas, I.; Efstratiadis, A.; Makropoulos, C. Stochastic Periodic Autoregressive to Anything (SPARTA): Modeling and Simulation of Cyclostationary Processes With Arbitrary Marginal Distributions. *Water Resour. Res.* **2018**, *54*, 161–185. [CrossRef]
9. Tsoukalas, I.; Kossieris, P.; Makropoulos, C. Simulation of Non-Gaussian Correlated Random Variables, Stochastic Processes and Random Fields: Introducing the AnySim R-Package for Environmental Applications and Beyond. *Water* **2020**, *12*, 1645. [CrossRef]
10. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]