TAILORING SEASONAL TIME SERIES MODELS TO FORECAST SHORT-TERM WATER DEMAND

Ernesto Arandia,[1] Amadou Ba,[2] Bradley Eck, M.ASCE[3] and Sean McKenna[4]

ABSTRACT

This paper presents a methodology to forecast short-term water demands either offline or online by combining SARIMA (seasonal autoregressive integrated moving average) models with data assimilation. In offline mode, the method frequently re-estimates the models using the latest historical data. In online mode, the method applies a Kalman filter to optimally and efficiently update the models using a real-time feed of data. The tailoring process consists of identifying, estimating and validating the models, along with exploring how the length of demand history used in fitting can improve forecast performance. We obtain a suite of models adequate for 15-min, hourly and daily demands having daily and weekly periodicities. We analyze the model output across temporal resolutions, periodicities and forecasting modes. We find that the normalized forecast deviations range from, approximately, 4.2 to 1.3% in correspondence to a decrease in temporal granularity. Models of the weekly-seasonal type are found to more efficiently remove the autocorrelations with respect to models of the daily-seasonal type. In terms of the forecasting mode, the online implementation is shown to produce a higher performance specially for models with higher temporal resolution. Finally, a case study is conducted where forecasts are compared to the actual water production volumes of the local water utility. The results indicate that a significant improvement may

[1]IBM Research, IBM Technology Campus Bldg. 3, Damastown, Dublin 15, Dublin, Ireland. E-mail: ernestoa@ie.ibm.com

[2]IBM Research, IBM Technology Campus Bldg. 3, Damastown, Dublin 15, Dublin, Ireland. E-mail: amadouba@ie.ibm.com

[3]IBM Research, IBM Technology Campus Bldg. 3, Damastown, Dublin 15, Dublin, Ireland. E-mail: bradley.eck@ie.ibm.com

[4]IBM Research, IBM Technology Campus Bldg. 3, Damastown, Dublin 15, Dublin, Ireland. E-mail: seanmcke@ie.ibm.com

be obtained in estimating the production of water based on the model output.

Keywords: Water demand, forecasting, water production, time series models, real-time estimation, temporal resolution, data assimilation.

INTRODUCTION

Water demand forecasts are essential to water utilities in the operation and planning activities since, virtually, all decisions regarding the management of the utility's physical assets, including treatment plants, wells, pumping stations, reservoirs, tanks, and distribution network rely on predictions about future consumption (Billings and Jones 2008). In the short term, water utilities need an estimate of daily water consumption one or more days in advance to size the amount of water to treat or purchase to satisfy the demand (Donkor et al. 2014). In addition, more refined predictions, usually at hourly level, are required to manage their pumping and storage to benefit from the electricity tariff structure (Alvisi et al. 2007).

The financial pressure due to increasing urbanization, aging infrastructure and stringent water-quality regulations (European Council 1998; EPA 2015b) is forcing water utilities in Europe and the United States to focus their efforts on more efficient, economic, and sustainable resource management. Water-loss and electricity costs account for a significant portion of a utility's operating budget. A conservative estimate of the total annual cost of non revenue water (NRW) to utilities worldwide is US$ 14 billion (Kingdom et al. 2006). Energy is usually the second highest budget item for municipal drinking water and wastewater facilities; it can comprise 30-40% of a utility's total operating costs (EPA 2015a). Therefore, reducing NRW and energy consumption are key objectives. And an increasingly sensitive demand response plan is a powerful strategy to achieve them. The motivation towards acquiring and analyzing the supporting water consumption data in real time will thus continue to grow and prompt the exploitation of currently existing data acquisition systems and the development of dynamic forecasting methods (Arandia et al. 2014a; Arandia et al. 2014b).

During the last four decades, a considerable amount of research on demand forecasting has been published, originally multiple regression and time series analysis techniques were proposed (Maidment et al. 1985; Maidment and Miaou 1986; Fildes et al. 1997; Zhou et al. 2000; Aly and Wanakule 2004). The models attempted to represent the growth and periodic behaviour of monthly, daily or hourly water use. The dependence of water demand on climatic variables such as rainfall and air temperature was also considered (Jain and Ormsbee 2001; Jain et al. 2001; Bougadis et al. 2005; Aly and Wanakule 2004) and temporal disaggregation techniques (Zhou et al. 2000; Zhou et al. 2002) and deterministic components (Aly and Wanakule 2004) were in some cases implemented. Other similar approaches (Gato et al. 2007a; Gato et al. 2007b) decomposed the water demand time series into base and seasonal parts, where base use is characterised by water consumption during the winter months and seasonal use includes the effects of seasonal, climatic and persistence components.

Jain and Ormsbee (2001) developed a decision support system for drought characterization to assist operators and water managers of the water supply system of the city of Lexington, Kentucky. The system included a water demand forecasting module that enables forecasts five days ahead. Various modeling techniques including regression, time series analysis and artificial neural networks were explored for forecasting water demand.

Soon after, Jain et al. (2001) proposed the application of artificial neural networks (ANNs) in forecasting water usage. They compared regression models, univariate time series models, and ANN models in terms of their ability to forecast weekly demands. The authors report that the ANN methods consistently outperformed the conventional techniques; the most complex ANN model was able to achieve an absolute percentage error (APE) in forecasting of 2.4%.

Jain and Ormsbee (2002) examined the suitability of ANN models for use in forecasting daily demands; the ANN performance was compared with results produced using conventional time series and regression models. The effects of rainfall and air temperature were also considered. A number of models of each category were estimated from daily data and ANN were found to outperform the conventional models.

Altunkaynak et al. (2005) presented a fuzzy logic method to forecast monthly water consumption. They examined different model configurations and selected the one with best performance error statistics. The fuzzy model does not require assumptions such as stationarity and ergodicity which are key requirements of conventional stochastic modeling. The model was applied to forecast one-month-ahead demands of the city of Istanbul, Turkey producing relative errors of less than 10%.

Cutore et al. (2008) proposed an approach where the Shuffled Complex Evolution Metropolis algorithm (SCEM-UA) is applied in the calibration of an ANN water consumption forecasting model. The model was used not only to predict water demands but also to quantify the uncertainties of model parameters and demand predictions. In assessing the model performance, the authors reported similar results to those obtained from conventional models.

In another comparative study by Herrera et al. (2010), a suite of models including ANN, projection pursuit regression, multivariate adaptive regression splines, random forests, support vector regression (SVR) and a simple model based on the weighted demand profile were assessed to forecast hourly water usage in Spain. The authors concluded that the SVR model performs best.

The progression of research summarized above signals a clear trend towards embracing artificial intelligence techniques due to their reported performance qualities (Zhang 2001; Jain and Ormsbee 2002; Bougadis et al. 2005; Ghiassi et al. 2008; Adamowski et al. 2012). However, conventional methods such as multiple regression and time series analysis still remain the most common forecasting methods (Adamowski and Chan 2011). The performance advantages of ANN techniques are attributed to the ability to identify non-linear relationships among different variables in water demand time series of different characteristics (Tiwari and Adamowski 2013). One of the main drawbacks of ANN methods is, however, their performance limitations in dealing with *noisy* and non-stationary data (Tiwari and Adamowski 2013).

More recently, there appears to be growing interest in hybrid approaches which exploit the strengths of individual methods and aim to reduce model uncertainty (Srinivasulu and

Jain 2009; Kant et al. 2013; Tiwari and Adamowski 2013). Caiado (2010), for instance, examined the performance of double seasonal univariate time series models in isolation and as an ensemble. Optimally combining forecasts from different model types was found to improve the forecast accuracy. Considering a different and wider arrange of models, Tiwari and Adamowski (2013) presented an application of a hybrid neural network forecasting model as an ensemble of several ANNs built using bootstrap sampling and wavelet analysis. The performance of these models was evaluated for daily, weekly and monthly lead times. In addition, the performance of the method was compared with the autoregressive integrated moving average (ARIMA) and autoregressive integrated moving average model with exogenous input variables (ARIMAX) and conventional ANNs. The authors found that their hybrid method produced more accurate forecasts than the conventional time series and ANN models.

Across the formal forecasting techniques cited above, we identify aspects that remain to be addressed. First, there is a lack of models that deal with sub hourly data, which is increasingly becoming available as a standard in the water industry. Second, models trained offline are of limited use in a real-time context because they lack a data assimilation component. Third, data assimilation is gaining popularity (Hutton et al. 2014; Shang et al. 2006; Nasseri et al. 2011; Preis et al. 2010), but its use in forecasting water demand is, to our knowledge, quite limited (Nasseri et al. 2011). Fourth, only a few studies (Cutore et al. 2008; Hutton and Kapelan 2015) have attempted to quantify short-term water demand forecasting uncertainty. And fifth, SARIMA models have not received much attention in the water domain (Caiado 2010; Arandia 2013) in spite of their qualities, such as parsimony and a straightforward interpretability of their parameters due to explicit mathematical formulations (Box et al. 2008; Shumway and Stoffer 2000). In fact, SARIMA models have been successfully used in applications such as electricity load and traffic flow forecasting (Taylor 2003; Taylor and McSharry 2007; Sevlian and Rajagopal 2014; Sigauke and Chikobvu 2011; Williams and Hoel 2003).

This paper presents a water usage forecasting method that applies SARIMA models and assimilates demand measurements to produce online forecasts and estimates of uncertainty. The data assimilation technique is based on the Kalman filter and requires a state-space model. Unlike ANNs or other "black-box" models, SARIMA can be cast in state-space form. Our tailoring methodology consists of identifying model parametric structures suitable for water demands with temporal resolutions ranging from sub-hourly to daily. We discuss how the method can be used to forecast demands either offline or online. The offline mode is suitable for utility operations, such as sizing daily water production while the online mode may be adequate for other operations, such as scheduling pumps. The forecast horizon is fixed to 24 h for consistency with the daily planning of water utilities.

METHODOLOGY

This section presents a description of the time series models, the model estimation process, the forecasting approach, the optimal filtering methodology and the model performance assessment approach. The relationship among the components of the method is illustrated in Fig. 1. The model estimation module comprises Algorithm 1 (Fig. 2(a)) and yields at least one model that may be used by the offline or online forecasting components. The offline forecasting module requires an estimated model, a starting time t_0 and an ending time t_f as input parameters. The trigger time t_i (time of origin for the forecasts) is initially set equal to t_0 and then Algorithm 2 (Fig. 2(b)) is executed iteratively to retrieve training data, re-estimate the model and compute forecasts; then, the trigger time is updated and the process repeated as long as $t_i \leq t_f$. The online forecasting component comprises Algorithm 3 (Fig. 3); there is a unidirectional flow between Algorithms 1 and 3 because once a model is estimated its parameters are updated by means of the Kalman filter with no need of re-estimation. The sections below explain the models and components in detail.

Time Series Models

SARIMA models were used to forecast water demands. A SARIMA model is denoted as ARIMA$(p, d, q) \times (P, D, Q)_s$ (Shumway and Stoffer 2000) and is compactly formulated as

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)\epsilon_t. \tag{1}$$

The variables x_t and ϵ_t represent, respectively, the measured water demand time series and a random error process with variance σ, where t is the time index. The term B is the backshift operator defined by $B^k x_t = x_{t-k}$. The equation also includes the seasonal autoregressive polynomial

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps},$$

the seasonal moving average polynomial

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs},$$

the ordinary autoregressive polynomial

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p,$$

and the ordinary moving average polynomial

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q,$$

where P, Q, p and q are the respective polynomial orders, and s is the seasonal period. In addition, Eq. (1) contains the seasonal differencing operator $\nabla_s^D = (1 - B^s)^D$ and the ordinary differencing operator $\nabla^d = (1 - B)^d$, where D and d are the differencing orders. Finally, the equation includes the intercept $\delta = \mu(1 - \phi_1 - \ldots - \phi_p)(1 - \Phi_1 - \ldots - \Phi_P)$, where μ is the mean of the demand time series.

Model Selection

The method to select a SARIMA model (Box et al. 2008) is illustrated in Fig. 2(a); it consists of iteratively identifying a suitable model structure, estimating the model parameters and performing a diagnostic checking on the residuals to assess if the model is well fitted. Only when the diagnostic checks are passed, the model is used in forecasting, otherwise a new model needs to be identified. Usually, several prospective models are tested and more than one alternative model is selected for further forecast performance analysis.

Models suitable to the training data are identified by sequentially finding values for the parameters s, d, D, P, Q, p, and q. First, the autocorrelation and partial autocorrelation functions (ACF and PACF) are evaluated in search of non-stationary and periodic behavior. This step aims at identifying the seasonal period s and the differencing operators d and D. Second, the data is filtered using ∇^d and ∇_s^D and the ACF and PACF are again computed. Third, the results are examined to verify the adequacy of s and D and to identify possible combinations of orders P, Q, p, and q. Fourth, all the alternative order combinations are used to fit SARIMA models to the training data. Finally, the models are compared using goodness of fit statistics (see below). The process usually yields more than one combination of suitable orders, thus multiple candidate models.

The estimation task produces maximum-likelihood estimates for the polynomial coefficients of each of the model structures identified. It uses an algorithm that combines the methods by Gardner et al. (1980) and Jones (1980). The algorithm consists of a recursive approach that computes a set of standardized prediction errors and the determinant of the covariance matrix of demand measurements. Together, these two quantities yield the exact likelihood which is maximized by a numerical optimization algorithm that does not require analytic derivatives.

After estimation, the alternative models pass through a diagnostic check which includes the analysis of the standardized residuals as well as model comparisons. The standardized residuals are the normalized differences between the demand measurements used in training and the one-step-ahead predictions based on the fitted model. If the model fits well, the standardized residuals are expected to behave as an independent and identically distributed sequence with mean zero and variance one. This condition is verified graphically (time series plot, histogram, autocorrelation plot, Q-Q plot) and numerically using the Ljung-Box-Pierce Q-statistic (Shumway and Stoffer 2000). The model comparisons are performed using the AIC, AICc and BIC statistics (Shumway and Stoffer 2000) which measure the goodness of fit by balancing the error of the fit (based on the residual sum of squares) against the number of model parameters.

Forecasting Algorithm

The forecasting algorithm is illustrated in Fig. 2(b), where the required input arguments are the data sampling frequency or number of measurements per hour f, the time length of the training window τ, the forecasting horizon h, and the orders of the priorly selected SARIMA model. The *trigger* time t_i in Fig. 2(b) represents, for instance, the time at the end of the day from where forecasts at the desired data resolution will be generated. New model parameters are estimated every time a trigger time is reached. The estimation requires a new training data set of length τ. The fitted model is used to compute and return the forecasts at every time step up to $t + h$. In addition, once historical data is available, the forecast error statistics are computed and reported.

State-Space Model

Casting a SARIMA model in state-space form enables the application of the Kalman filter (Hamilton 1994). The general formulation of the state space model includes an observation equation (Eq. 2) and a state equation (Eq. 3):

$$\boldsymbol{y}_t = \boldsymbol{A}^\top \boldsymbol{u}_t + \boldsymbol{H}^\top \boldsymbol{z}_t + \boldsymbol{w}_t, \tag{2}$$

$$\boldsymbol{z}_t = \boldsymbol{F} \boldsymbol{z}_{t-1} + \boldsymbol{v}_t. \tag{3}$$

where $\boldsymbol{y}_t$ is the vector of observed variables, $\boldsymbol{z}_t$ is the state vector which consists of unobserved variables, $\boldsymbol{u}_t$ is a vector of predetermined variables that possibly are lagged values of $\boldsymbol{y}_t$, $\boldsymbol{A}$ is a predetermined matrix, $\boldsymbol{F}$ and $\boldsymbol{H}$ are parameter matrices, $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are error terms assumed to be distributed as white noise with covariance matrices $\boldsymbol{W}$ and $\boldsymbol{R}$ (Hamilton 1994). Eq. (2) has the form of a linear regression model and Eq. (3) is written as a first order vector autoregressive model (Hamilton 1994).

To obtain the state space form, a SARIMA$(p,d,q) \times (P,D,Q)_s$ model of the univariate water demand series y_t is written as an equivalent ARMA$(p+sP, q+sQ)$ (Š̈avås 2013) for a transformed variable y_t^*. The equivalent model is expressed in state-space form as

$$\boldsymbol{y}_t = \boldsymbol{H}^\top \boldsymbol{z}_t, \tag{4}$$

$$\boldsymbol{z}_t = \boldsymbol{F}\boldsymbol{z}_{t-1} + \boldsymbol{v}_t, \tag{5}$$

where $\boldsymbol{A}$, $\boldsymbol{u}_t$, and $\boldsymbol{R}$ have all been set to zero because there are no exogenous variables and no noise in the measurements is considered. The state-space matrices have dimension $r = max(p+sP, q+sQ+1)$ and are specified as

$$\boldsymbol{y}_t = y_t^*, \quad \boldsymbol{z}_t^\top = [z_{1,t}, z_{2,t}, \ldots, z_{r,t}], \quad \boldsymbol{H}^\top = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix},$$

$$\boldsymbol{F}^\top = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{r-1} & \phi_r \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \quad \boldsymbol{v}_t = \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{bmatrix} \epsilon_t,$$

$$\boldsymbol{W}^\top = \sigma^2 \begin{bmatrix} 1 & \theta_1 & \ldots & \theta_{r-1} \\ \theta_1 & \theta_1^2 & \ldots & \theta_1\theta_{r-1} \\ \vdots & \vdots & \ldots & \vdots \\ \theta_{r-1} & \theta_1\theta_{r-1} & \ldots & \theta_{r-1}^2 \end{bmatrix}.$$

Before applying the Kalman filter, the matrices $\boldsymbol{F}$, $\boldsymbol{H}$, $\boldsymbol{v}$, and $\boldsymbol{W}$ need to be computed using the previously estimated parameters of the SARIMA model.

The Kalman Filter

The Kalman filter allows updating the state vector of Eq. (5) every time there is a new observation of the possibly multivariate series $\boldsymbol{y}_t$ (Harvey 1989). The filter consists of a sequence of steps where $\hat{\boldsymbol{z}}_t$ is linearly estimated from known values of $\hat{\boldsymbol{z}}_{t-1}$ and $\boldsymbol{y}_t$. An initial step is required where the prior state $\boldsymbol{z}_1$ is computed from prior information or assumed to be zero in the absence of such information. Also, the prior state's variance needs to be computed from

$$\boldsymbol{P}_{1|0} = vec(\boldsymbol{P}_{1|0}) = [\boldsymbol{I} - (\boldsymbol{F} \otimes \boldsymbol{F})]^{-1} \times vec(\boldsymbol{W}), \tag{6}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{F}$ and $\boldsymbol{W}$ are previously known, and $vec(\boldsymbol{P}_{1|0})$ is the column vector which is directly transformed into the quadratic matrix $\boldsymbol{P}_{1|0}$.

Following the prior state estimation, the values of the state variables are updated iteratively by recursions on the equation below:

$$\hat{\boldsymbol{z}}_{t+1|t} = \boldsymbol{F}\hat{\boldsymbol{z}}_{t|t-1} + \boldsymbol{F}\boldsymbol{P}_{t|t-1}\boldsymbol{H}(\boldsymbol{H}^\top \boldsymbol{P}_{t|t-1}\boldsymbol{H})^{-1}(\boldsymbol{y}_t - \boldsymbol{H}^\top \hat{\boldsymbol{z}}_{t|t-1}), \tag{7}$$

where $\hat{\boldsymbol{z}}_{t|t-1}$ is the forecast of the true state $\boldsymbol{z}_t$ based on the linear function of the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}$ and $\boldsymbol{P}_{t|t-1}$ is the variance of this forecast. The forecast variance $\boldsymbol{P}_{t+1|t}$ needs to be updated accordingly using the equation below:

$$\boldsymbol{P}_{t+1|t} = \boldsymbol{F}\left[\boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{H}(\boldsymbol{H}^\top \boldsymbol{P}_{t|t-1}\boldsymbol{H})^{-1}\boldsymbol{H}^\top \boldsymbol{P}_{t|t-1}\right]\boldsymbol{F}^\top + \boldsymbol{W}, \tag{8}$$

Eqs. (7) and (8) are implemented in the algorithm of Fig. 3, where the recursions are separated in smaller components.

Throughout the remainder of the paper, we will use the terms *offline* to refer to forecasts that are obtained by means of parameter re-estimation (Algorithm 2) and *online* to mention forecasts that result from updating the model parameters (in state-space form) by applying the Kalman filter (Algorithm 3).

Model Performance Assessment

The forecasting algorithm of Fig. 2(b) is applied to all the alternative models and the results are analysed to assess performance. The analysis includes the computation of forecast error statistics and the comparison of results across the candidate models. In addition, forecasts are produced and assessed after the implementation of the Kalman filter algorithm of Fig. 3.

One of the performance statistics computed is the root mean squared error (RMSE),

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{x}_t - x_t)^2}{n}}, \tag{9}$$

where $\hat{x}_t$ represents the forecasted value and n is the total number of measurements in the number of forecast days. Another statistic is the normalized RMSE (NRMSE),

$$\text{NRMSE} = \frac{\text{RMSE}}{x_{\max} - x_{\min}}, \tag{10}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum value of the demand measurements. Finally, a statistic used as meaningful to perform comparison across models' forecasts and utility operations is the mean absolute percentage error (MAPE),

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{(x_t - \hat{x}_t)}{x_t}\right|. \tag{11}$$

RESULTS AND DISCUSSION

In this section, the model estimation, forecasting, and assessment results are presented. The selection of model structures is achieved first by means of an extensive testing of alternatives; only the best candidates are presented and assessed.

Water Demand Data

Data from three sources was used in this study, namely, A, B, and C. Source A is a set of flow rate measurements generated by over 550 individual telemetry instruments distributed across 190 district metered areas (DMAs) in the city of Dublin, Ireland. The temporal resolution of these data sets is 15 min and their total length is 19 months, spanning from January 2010 throughout July 2011. The measurements from individual DMAs were preprocessed and aggregated to obtain a total demand series. The relevant descriptive statistics of the series, namely, minimum value, maximum value, mean, median and standard deviation are, respectively, 78.46, 199.93, 144.11, 151.93, 24.42 ML/d. The mean value (144.11 ML/d) represents a fraction of roughly 30% of Dublin's total consumption; however, from a modeling perspective the data corresponds to a high level of aggregation and qualitatively has the characteristics of a total system demand series.

Source B consists of hourly demands obtained from approximately 600 smart meters distributed across three DMAs of the water network of a European city. This data was provided by partners of the iWIDGET Project funded by the European Union (Barry et al. 2014). The individual-meter demand measurements were aggregated to produce a 10-month time series (November 2008 throughout August 2009) having minimum value, maximum value, mean, median and standard deviation of, respectively, 0, 1.63, 0.36, 0.33 and 0.22 ML/d. This data constitutes a different set of measurements for model testing which, in contrast to source A, has a low level of spatial aggregation.

Source C comprises daily flow rate measurements of total system demand and total water production in Dublin in 2013; these measurements are made publicly available by the local water utility (DCC 2013). This data source was used to assess the benefits of the proposed forecasting methods with respect to the real water production performance of the water utility. The minimum value, maximum value, mean, median and standard deviation of this dataset are, respectively, 496.21, 576.02, 540.15, 541.61 and 11.54 ML/d.

Estimation Results

As a general rule, each data source A, B and C was divided in training and validation segments comprising, respectively, about 60 and 40% of the available data. Hence, respective to sources A, B and C, the specific time series lengths used in training were 11, 6, and 7 months; and the remainder 8, 4, and 5 months were used in validating the models.

The model selection process described above was performed by first examining the periodicity and non-stationarity of the data. Fig. 4 illustrates the ACF of a training data segment. The slow decay of the ACF indicates non-stationarity and the periodicity is clearly noticeable from the daily and weekly peaks. Hence, both seasonalities ($s = 1$ d and $s = 7$ d) were found suitable. Several combinations of orders were analyzed from an examination of the ACF/PACF computed after differencing the data. In general, the results suggested that a seasonal moving average process is most suitable; thus, the seasonal-autoregressive and seasonal-moving-average orders were assigned values $P = 0$ and $Q = 1$, respectively. Moreover, the autocorrelation analysis suggested trying combinations of orders p and q ranging from 0 to 4. By fitting different models and comparing the AIC, AICc, BIC and residual randomness across the alternatives, the autoregressive and moving-average orders were assigned values $p = 0$ and $q = 4$, respectively. Consequently, the resulting left-hand-side polynomials are $\Phi_P(B^s) = 1$, $\phi(B) = 1$, $\nabla_s^D = (1 - B)^s$ and $\nabla^d = (1 - B)$; and the right-hand-side polynomials are $\Theta_Q(B^s) = 1 + \Theta_1 B^s$ and $\theta(B) = 1 + \theta_1 B + \ldots + \theta_4 B^4$. Therefore, the suitable general structure is

$$(1 - B^s)(1 - B)x_t = \delta + (1 + \Theta_1 B^s)\left(1 + \sum_{i=1}^{4} \theta_i B^i\right)\epsilon_t, \tag{12}$$

where s depends on the resolution and the seasonal correlation period; the vector of parameters to estimate is $\boldsymbol{\Theta} = (\Theta_1, \theta_1, \theta_2, \theta_3, \theta_4, \sigma, \delta)^T$. Since the number of parameters is the same in every model, there is no reason to account for number of parameters in the performance assessment using statistics such as the AIC, BIC, or K-L distance, etc. The suite of alternative models identified under the general structure is summarised in Table 1, where the models are distinguished by their seasonal period and their associated data resolution. For instance, an S-96 model has $s = 96$ which is the product of the measurement frequency (inverse of the resolution) $f = 4$ h^{-1} and the seasonal period, equal to 24 h. The table also indicates the data sources that correspond to each model structure; the key data attribute in this relationship is the resolution. For example, models of type S-96 and S-672 apply only to data source A because they require a measurement resolution of 1/4 h (15 min).

Forecasting Results

The forecasting algorithm (Fig. 2(b)) was applied on the data sources A, B and C using the model inputs of Table 1. In addition to the model parameters and the demand measurements, the algorithm requires values for the length of the training window τ and the forecasting horizon h. The length τ was selected through experimentation with values ranging from 7 to 28 d. A 7-d window represents the minimum length of data required to fit a model with weekly seasonal period. Even though the models with daily period require smaller windows, the same lower limit was used for all models to facilitate a comparison between periods.

The forecasting algorithm was executed on a validation data segment taking random trigger days t_i and fixing h to 24 h for each time window and each model structure; the number of runs was 450. Table 2 presents the RMSE of the forecasts for dataset A in correspondence to different values of τ. For hourly and sub-hourly models, the RMSE increases in response to an increase of the training window. On the contrary, for the daily models, the RMSE decreases as τ increases. The reason for such trends lies in the tradeoff between model uncertainty and temporal resolution. For hourly and sub-hourly demands, as more estimation data is added the higher levels of data noise produce an increase in model uncertainty that outweighs the benefits from adding information through measurements. On the other hand, since daily data is considerably smoother, the opposite effect is observed. However, we do not generalize this experimental findings but point out that the length of the training window may be dynamically learned for optimal results. In our analysis, the "best" observed windows of Table 2 were selected, i.e., $\tau = 7$ d was used in fitting models S-96 through S-168 and $\tau = 28$ d was used in estimating models S-1 and S-7.

A sample of the offline forecasts for the sub-hourly, hourly and daily models is presented in Fig. 5. Each plot shows a data segment with the one-day-ahead forecasts and the uncertainty bands (95% confidence level) in grey for the corresponding data source and resolution. The figure illustrates how the different models respond to the data characteristics. For instance, Fig. 5(a) displays a good agreement between data and forecast at the level of total system demand and when the temporal resolution is high. Fig. 5(b) also shows a good forecast behavior for a coarser temporal resolution of 1 d; although there appear to be larger deviations (due to the display scale), daily models perform better than higher resolution models because there is an effect of noise reduction or smoothing of the demand signal when temporal aggregation is performed.

A similar pair of plots for data source B (Figs. 5(c) and Fig. 5(d)) shows that when the spatial aggregation level is low the model performance is poorer. This behavior is expected and is due to an increase in the randomness of the demand signal as the aggregation level approaches the individual service connection.

Using a set of SARIMA parameters estimated during the offline forecasting stage, the Kalman filter (Fig. 3) was applied to generate a new online set of forecasts for each data source and model structure. The parameters were not recalculated in online mode, hence the online computational performance was substantially higher, e.g., an S-672 model took approximately 4020 and 108 seconds to forecast one day ahead in offline and online mode, respectively. In most cases, the quality of the forecasts was considerably improved as well. A graphical comparison is presented in Fig. 6 where it is clear that for the S-672 model the filter noticeably increases the prediction accuracy; for the S-7 model the improvements are smaller and not evident without a statistical metric. A complete picture of model performance is presented in the assessment section below.

Model Performance Assessment

This section refers to the offline and online forecasts obtained with the model structures summarized in Table 1 and by means of Algorithms 2 and 3 (Figs. 2(b) and 3). Forecasts were generated with the models for the validation segments of datasets A, B and C. In order to assess the performance, the RMSE (Eq. 9), the NRMSE (Eq. 10) and the MAPE (Eq. 11) were computed from the deviations of the forecasts with respect to the validation data. Table 3 summarises the results.

All three statistics follow similar trends for the six model types since the same validation datasets were used. However, the MAPE is considered most meaningful in cross comparison because it normalises the errors at every measurement. Thus, Fig. 7 graphically summarizes the distribution of the MAPE for each model of Table 3.

The results of Fig. 7(a) for data source A indicate that the median percentage value for all models is below 2.5%. Among the offline models with sub-hourly resolution S-672 performs better than S-96 because the weekly seasonal period is more suitable for the data. A similar relationship is observed for the hourly models S-168 and S-24 as well as for the daily models S-1 and S-7. For offline prediction purposes and when high resolution forecasts are required, the S-672 and S-168 are recommended. For hourly forecasts the S-168 offers good performance and higher computational efficiency. When daily water production forecasts are required, the S-7 model is recommended.

The statistics in Table 3 and Fig. 7(a) indicate that the application of the Kalman filter considerably reduces the magnitude and variability of the forecast errors for the sub-hourly and hourly models. Also, it practically homogenizes the performance across all resolutions. In the case of the daily models, the filtered forecasts are slightly better (SF-7) or even worse (SF-1) in quality. The reason is the lack of updating of the SARIMA parameters. For daily forecasting therefore it is recommended to perform a frequent (daily) re-estimation of parameters.

The results of Table 3 and Fig. 7(b) for data source B show a much higher level of errors where the median APE reaches up to 30%. As mentioned above, the reason for such magnitude of errors is the randomness in demands when the aggregation scale is small. For this dataset, the daily-periodic model S-24 outperforms the weekly-periodic model S-168 and is thus recommended for hourly prediction. On the other hand, model S-7 is slightly more accurate than model S-1 and is preferred for daily forecasting. Once more, the online representation of the models leads to better performance except for SF-1, for the reason explained above.

Once the capabilities of the models were identified, a practical test was conducted using data source C. The purpose was to illustrate the capabilities of the forecasting models in improving water utility operations. Specifically, the best daily model S-7 was used to forecast water demands in Dublin and the results compared to the public records of water production by the local utility. To illustrate, Fig. 8 presents a segment of the results where total demand, production, and forecasts are displayed. The MAPE of production by the utility in 2013 is 1.90% (Table 3) which may be considered an acceptable performance. However, as shown in Table 3 and Fig. 8, the output of the model has considerably lower deviations with a MAPE of 1.08%. Thus the resulting error reduction is 43% which amounts to the daily consumption of approximately 30,000 people.

SUMMARY AND CONCLUSIONS

The methodology presented couples SARIMA models with data assimilation to forecast sub-hourly, hourly and daily water demand. We tailored the models to three qualitatively different datasets by uncovering the demand seasonality, identifying suitable model structures, and estimating the size of the training data sample. SARIMA models have a state-space representation which allows for elegant and straightforward implementation of data assimilation using the Kalman filter. The models were applied to forecast demands 24 h in advance in offline mode (re-estimation of parameters) and online mode (assimilation of measurements).

We analyze and compare the performance of the models, and demonstrate the application of the method in estimating daily water production for the local water utility. In offline mode and for demands at total system level, models with weekly seasonality perform better than models with daily periodicity. At sub-DMA aggregation scales, models with both types of seasonality behaved in a similar manner. An overall trend observed is a decay of the mean and dispersion of the APE as the temporal resolution increases. This result is expected, because smoothing the data has the effect of reducing model uncertainty. In online mode, the data assimilation method significantly improves forecast accuracy both for total-system and sub-DMA demands while also improving computational performance. Future research may consider an adaptive sizing of the training window to further tailor the models.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Dublin City Council for the telemetry data provided. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 318272

REFERENCES

Adamowski, J. and Chan, H. F. (2011). "A wavelet neural network conjunction model for groundwater level forecasting." *Journal of Hydrology*, 407(1-4), 28–40.

Adamowski, J., Chan, H. F., Prasher, S. O., Ozga-Zielinski, B., and Sliusarieva, A. (2012). "Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in montreal, canada." *Water Resources Research*, 48(1).

Altunkaynak, A., Özger, M., and Çakmakc, M. (2005). "Water consumption prediction of Istanbul city by using fuzzy logic approach." *Water Resour. Manage.*, 19(5), 641–654.

Alvisi, S., Franchini, M., and Marinelli, A. (2007). "Short-term forecasting for urban water consumption." *J. Water Resour. Plann. Manage.*, 9(1), 39–50.

Aly, A. and Wanakule, N. (2004). "Short-term forecasting for urban water consumption." *J. Water Resour. Plann. Manage.*, 130(5), 405–410.

Arandia, E. (2013). "Spatial-temporal statistical modeling of treated drinking water usage." Ph.D. thesis, University of Cincinnati, Cincinnati, OH.

Arandia, E., Eck, B., and McKenna, S. (2014a). "The effect of temporal resolution on the accuracy of forecasting models for total system demand." *Procedia Engineering*, 89, 916–925.

Arandia, E., Uber, J., Boccelli, D., Janke, R., Hartman, D., and Lee, Y. (2014b). "Modeling automatic meter reading water demands as nonhomogeneous point processes." *J. Water Resour. Plann. Manage.*, 140(1), 55–64.

Barry, M. G., Purcell, M. J., Eck, B. J., Hayes, J., and Arandia, E. (2014). "Web services for water systems: The iWIDGET REST API." *Procedia Engineering*, 89, 1120–1127.

Billings, R. B. and Jones, C. V. (2008). *Forecasting urban water demand.* American Water Works Association, Denver, CO, 2nd edition.

Bougadis, J., Adamowski, K., and Diduch, R. (2005). "Short-term municipal water demand forecasting." *Hydrolog. Process.*, 19(1), 137–148.

Box, G., Jenkins, G., and Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control.* Wiley Series in Probability and Statistics. Wiley.

Caiado, J. (2010). "Performance of combined double seasonal univariate time series models for forecasting water demand." *J. Hydrol. Eng.*, 15(3), 215–222.

Cutore, P., Campisano, A., Kapelan, Z., Modica, Z., and Savic, D. (2008). "Probabilistic prediction of urban water consumption using the SCEM-UA algorithm." *Urban Water J.*, 5(2), 125–132.

DCC (2013). "Dublin City Council, your drinking water: Regional demand during 2010 and 2013, <http://www.dublincity.ie/main-menu-services-water-waste-and-environment/your-drinking-water>.

Donkor, E. A., Mazzuchi, T. A., Soyer, R., and Roberson, J. A. (2014). "Urban water demand forecasting: Review of methods and models." *J. Water Resour. Plann. Manage.*, 140(2), 146–159.

EPA (2015a). "Energy efficiency for water and wastewater utilities", <http://water.epa.gov/infrastructure/sustain/energyefficiency.cfm>. Accessed: 2015-03-11.

EPA (2015b). “Safe Drinking Water Act (SDWA)”, <http://water.epa.gov/lawsregs/rulesregs/-sdwa/index.cfm>. Accessed: 2015-03-10.

European Council (1998). “Council directive 98/83/EC of 3 november 1998 on the quality of water intended for human consumption (OJ L 330, 5.12.1998 p. 32).

Fildes, R., Randall, A., and Stubbs, P. (1997). “One-day ahead demand forecasting in the utility industries: Two case studies.” *J. Oper. Res. Soc.*, 48, 15–24.

Gardner, G., Harvey, A., and Phillips, G. (1980). “Algorithm AS154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering.” *Applied Statistics*, 29(3), 311–322.

Gato, S., Jayasuriya, N., and Roberts, P. (2007a). “Forecasting residential water demand: Case study.” *J. Water Resour. Plann. Manage.*, 133(4), 309–319.

Gato, S., Jayasuriya, N., and Roberts, P. (2007b). “Temperature and rainfall thresholds for base use urban water demand modelling.” *J. Hydrol.*, 337(3-4), 364–376.

Ghiassi, M., Zimbra, D., and Saidane, H. (2008). “Urban water demand forecasting with a dynamic artificial neural network model.” *J. Water Resour. Plann. Manage.*, 134(2), 138–146.

Hamilton, J. (1994). *Time series analysis*. Princeton University Press, Princeton, New Jersey.

Harvey, A. (1989). *Forecasting structural time series models and the Kalman filter*. Cambridge University Press, Cambridge.

Herrera, M., Torgo, L., Izquierdo, J., and Perez-Garcia, R. (2010). “Predictive models for forecasting hourly urban water demand.” *J. Hydrol.*, 387(1-2), 141–150.

Hutton, C., Kapelan, Z., Vamvakeridou-Lyroudia, L., and Savić, D. (2014). “Dealing with uncertainty in water distribution system models: A framework for real-time modeling and data assimilation.” *J. Water Resour. Plann. Manage.*, 140(2), 169–183.

Hutton, C. J. and Kapelan, Z. (2015). “A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting.” *Environmental Modelling and Software*, 66(2), 87–97.

Jain, A. and Ormsbee, L. (2001). "A decision support system for drought characterization and management." *Civ. Eng. Environ.*, 18(2), 105–140.

Jain, A. and Ormsbee, L. E. (2002). "Short-term water demand forecast modeling techniques: Conventional methods versus AI." *J. Amer. Water Works Assoc.*, 94(7), 64–72.

Jain, A., Varshney, A., and Joshi, U. (2001). "Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks." *Water Resour. Manage.*, 15(5), 299–321.

Jones, R. (1980). "Maximum likelihood fitting of arma models to time series with missing observations." *Technometrics*, 22(3), 389–395.

Kant, A., Suman, P. K., Giri, B. K., Tiwari, M. K., Chatterjee, C., Nayak, P. C., and Kumar, S. (2013). "Comparison of multi-objective evolutionary neural network, adaptive neuro-fuzzy inference system and bootstrap-based neural network for flood forecasting." *Neural Computing and Applications*, 23(1), 231–246.

Kingdom, B., Liemberger, R., and Marin, P. (2006). "The challenge of reducing non-revenue water (NRW) in developing countries." *Water Supply and Sanitation Sector Board Discussion Paper Series, Paper no 8*, The World Bank.

Maidment, D. R. and Miaou, S. P. (1986). "Daily water use in nine cities." *Water Resources Research*, 22(6), 845–851.

Maidment, D. R., Miaou, S. P., and Crawford, M. M. (1985). "Transfer function models of daily urban water use." *Water Resources Research*, 21(4), 425–432.

Nasseri, M., Moeini, A., and Tabesh, M. (2011). "Forecasting monthly urban water demand using extended kalman filter and genetic programming." *Expert Systems with Applications*, 38(6), 7387 – 7395.

Preis, A., Whittle, A. J., Ostfeld, A., and Perelman, L. (2010). "Efficient hydraulic state estimation technique using reduced models of urban water networks." *J. Water Resour. Plann. Manage.*, 137(4), 343–351.

S̈avås, F. N. (2013). "Forecast comparison of models based on SARIMA and the Kalman filter for inflation." M.S. thesis, Uppsala University, Sweden.

Sevlian, R. and Rajagopal, R. (2014). “Short term electricity load forecasting on varying levels of aggregation.” *arXiv preprint arXiv:1404.0058*.

Shang, F., Uber, J. G., van Bloemen Waanders, B. G., Boccelli, D., and Janke, R. (2006). “Real time water demand estimation in water distribution system.” *8th Annual water Distribution Systems Analysis Symposium*.

Shumway, R. and Stoffer, D. (2000). *Time Series Analysis and Its Applications*. Springer-Verlag GmbH.

Sigauke, C. and Chikobvu, D. (2011). “Prediction of daily peak electricity demand in south africa using volatility forecasting models.” *Energy Economics*, 33(5), 882 – 888.

Srinivasulu, S. and Jain, A. (2009). “River flow prediction using an integrated approach.” *J. Hydrol. Eng.*, 14(75), 75–83.

Taylor, J. W. (2003). “Short-term electricity demand forecasting using double seasonal exponential smoothing.” *J. Oper. Res. Soc.*, 54(8), 799–805.

Taylor, J. W. and McSharry, P. E. (2007). “Short-term load forecasting methods: An evaluation based on european data.” *Power Systems, IEEE Transactions on*, 22(4), 2213–2219.

Tiwari, M. K. and Adamowski, J. (2013). “Water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models.” *Water Res. Research*, 49(10), 6486–6507.

Williams, B. M. and Hoel, L. A. (2003). “Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results.” *J. Transport. Eng.*, 129(6), 664–672.

Zhang, G. P. (2001). “An investigation of neural networks for linear time-series forecasting.” *Computers & Operations Research*, 28(12), 1183 – 1202.

Zhou, S., McMahon, T., Walton, A., and Lewis, J. (2000). “Forecasting daily urban water demand: A case study of Melbourne.” *J. Hydrol.*, 236(3-4), 153–164.

Zhou, S., McMahon, T., Walton, A., and Lewis, J. (2002). “Forecasting operational demand for an urban water supply zone.” *J. Hydrol.*, 259(1-4), 189–202.

NOTATION

The following symbols are used in this paper:

x_t	=	water demand time series
$\hat{x}_t$	=	water demand forecast
ϵ_t	=	white noise
B	=	backshift operator
ϕ	=	autoregressive polynomial of order p
θ	=	moving average polynomial of order q
Φ_P	=	seasonal autoregressive polynomial of order P
Θ_Q	=	seasonal moving average polynomial of order Q
∇_t^d	=	differencing operator of order d
∇_t^D	=	seasonal differencing operator of order D
δ	=	drift of the demand time series
τ	=	length of the estimation window
h	=	forecast horizon
$\boldsymbol{y}_t$	=	vector of observed demands
$\boldsymbol{z}_t$	=	state vector
$\boldsymbol{F}$	=	matrix of autoregressive parameters
$\boldsymbol{H}$	=	vector of state parameters
$\boldsymbol{W}$	=	covariance matrix of the error in the process
$\boldsymbol{R}$	=	covariance matrix of the error in the measurements
$\boldsymbol{P}$	=	matrix of state variance
$\boldsymbol{w}_t$	=	white noise vector
$\boldsymbol{u}_t$	=	white noise vector

List of Tables

1 Model structures identified 26
2 Effect of the training window length on forecasting error for data source A . 27
3 Statistics for assessment of model performance 28

TABLE 1. Model structures identified

Model ID	Data Source	Resol. (h)	f (h^{-1})	Seasonal Period (h)	Structure
S-96	A	1/4	4	24	$\mathrm{ARIMA}(0,1,4) \times (0,1,1)_{96}$
S-672	A	1/4	4	168	$\mathrm{ARIMA}(0,1,4) \times (0,1,1)_{672}$
S-24	A, B	1	1	24	$\mathrm{ARIMA}(0,1,4) \times (0,1,1)_{24}$
S-168	A, B	1	1	168	$\mathrm{ARIMA}(0,1,4) \times (0,1,1)_{168}$
S-1	A, B, C	24	1/24	24	$\mathrm{ARIMA}(0,1,4) \times (0,1,1)_{1}$
S-7	A, B, C	24	1/24	168	$\mathrm{ARIMA}(0,1,4) \times (0,1,1)_{7}$

TABLE 2. Effect of the training window length on forecasting error for data source A

Temp. Resol.	Model ID	RMSE (ML/d) by length τ 7 d	14 d	21 d	28 d
15 min	S-96	9.57	10.35	12.19	12.13
	S-672	9.31	9.94	11.88	12.09
Hourly	S-24	8.65	8.92	11.11	11.98
	S-168	9.30	9.45	13.40	13.27
Daily	S-1	4.32	3.10	2.64	2.48
	S-7	3.27	4.01	2.82	2.36

TABLE 3. Statistics for assessment of model performance

Model	Data	RMSE (ML/d)		NRMSE (%)		MAPE (%)	
	Source	Offline	Online	Offline	Online	Offline	Online
S-96	A	9.57	3.04	6.72	2.50	4.21	1.49
S-672	A	9.31	2.52	6.47	2.11	3.55	1.43
S-24	A	8.80	3.21	6.18	2.64	3.51	1.49
S-168	A	9.30	4.95	6.53	4.07	3.53	1.83
S-1	A	2.48	3.08	1.72	2.53	1.29	1.96
S-7	A	2.36	2.33	1.64	1.92	1.27	1.10
S-24	B	0.11	0.10	11.09	10.79	38.12	21.33
S-168	B	0.15	0.12	14.81	11.97	50.98	25.92
S-1	B	0.10	0.08	23.01	19.49	42.03	24.71
S-7	B	0.12	0.07	27.02	14.68	52.48	18.92
Util.-2013	C	13.33	-	2.47	-	1.90	-
S-7	C	8.22	8.18	1.52	1.51	1.09	1.08

List of Figures

1 Relationship among components of the method. 30

2 (a) Algorithm 1: model selection; (b) Algorithm 2: offline forecasting. 31

3 Algorithm 3: online forecasting by Kalman filter. 32

4 ACF of a segment of training data (source A). 33

5 Sample of the 24-h-ahead demand forecasts by the models with sub-hourly, hourly and daily resolutions. The grey bounds indicate the uncertainty (95% confidence level). (a) S-672 data source A; (b) S-7 data source A; (c) S-168 data source B; (d) S-7 data source B. 34

6 Comparison of the forecasts on data source A before and after applying the Kalman filter. (a) S-672 (b) S-7. 35

7 Comparison of model performance. The filtered (online) models are denoted by the prefix SF and by a different shade of grey. The model's seasonal period (daily or weekly) and data temporal resolution (15 min, 1 h or 1 d) are indicated on the x-axis. (a) Data source A; (b) Data source B. 36

8 Plots of total system demand, water production, and forecasts at daily level. 37

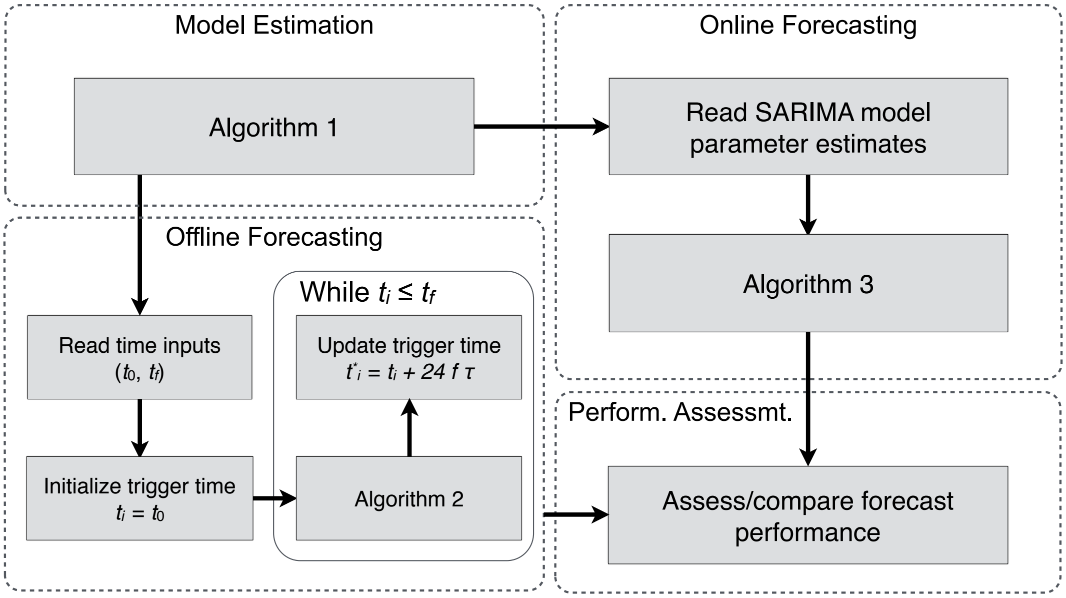


FIG. 1. Relationship among components of the method.

1: Read in training data $\{x_t\}_T$
2: Identify model orders $s, (d, D), (P, Q), (p, q)$
3: Estimate model parameters
4: Run diagnostic checking
5: **if** *diagnostic checking passes* **then**
6: Select model
7: **else**
8: Go to 2
9: **end if**
10: Return model orders and diagnostics

(a)

Read input parameters $f, \tau, h, (p, d, q, P, D, Q, s)$
while *Trigger date/time* t_i *is available* **do**
Get training data $\{x_t : t_i - \tau \leq t \leq t_i\}$,
Run Algorithm 1
Compute forecasts and deviations
Return forecasts and error statistics
end while

(b)

FIG. 2. (a) Algorithm 1: model selection; (b) Algorithm 2: offline forecasting.

Compute matrices $\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{v}, \boldsymbol{W}$
Initialize state: $\hat{\boldsymbol{z}}_t \leftarrow \boldsymbol{0}$
Initialize forecast variance: $vec(\boldsymbol{P}_{1|0}) \leftarrow [\boldsymbol{I} - (\boldsymbol{F} \otimes \boldsymbol{F})]^{-1} \times vec(\boldsymbol{W})$
while *new obs. at time* t **do**
    Predict state: $\hat{\boldsymbol{z}}^*_{t+1} \leftarrow \boldsymbol{F}^\top \hat{\boldsymbol{z}}_t$
    Predict the water demand: $\hat{y}_{t+1} \leftarrow \boldsymbol{H}^\top \hat{\boldsymbol{z}}^*_{t+1}$
    Store/report $\hat{y}_{t+1}$
    Estimate the error: $\hat{\epsilon}_{t+1} \leftarrow y_{t+1} - \hat{y}_{t+1}$
    Compute the variance: $\hat{\sigma}^2_t \leftarrow \boldsymbol{H}^\top \boldsymbol{P}_t \boldsymbol{H}$
    Update the state: $\hat{\boldsymbol{z}}_t \leftarrow \hat{\boldsymbol{z}}^*_{t+1} + \boldsymbol{F}\boldsymbol{P}_t\boldsymbol{H}(\hat{\sigma}^2_t)^{-1}\hat{\epsilon}_{t+1}$
    Update the forecast variance: $\boldsymbol{P}_{t+1} \leftarrow \boldsymbol{F}\left[\boldsymbol{P}_t - \boldsymbol{P}_t\boldsymbol{H}(\hat{\sigma}^2_t)^{-1}\boldsymbol{H}^\top\boldsymbol{P}_t\right]\boldsymbol{F}^\top + \boldsymbol{W}$
end while

FIG. 3. Algorithm 3: online forecasting by Kalman filter.

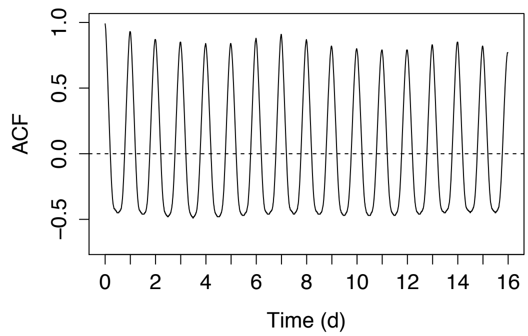
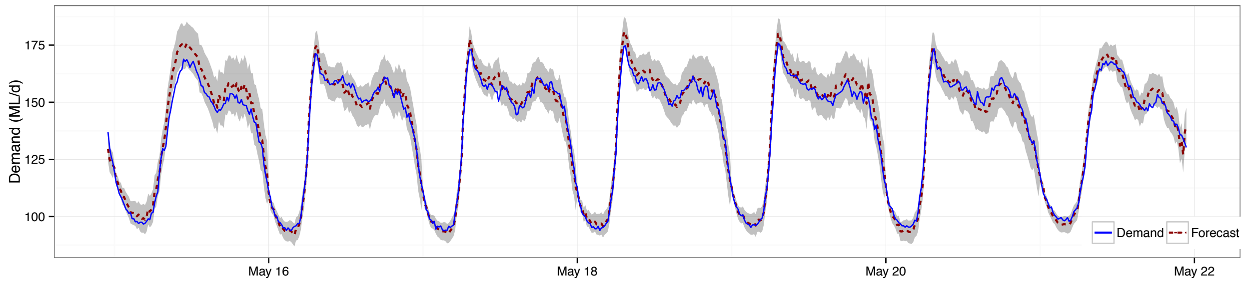
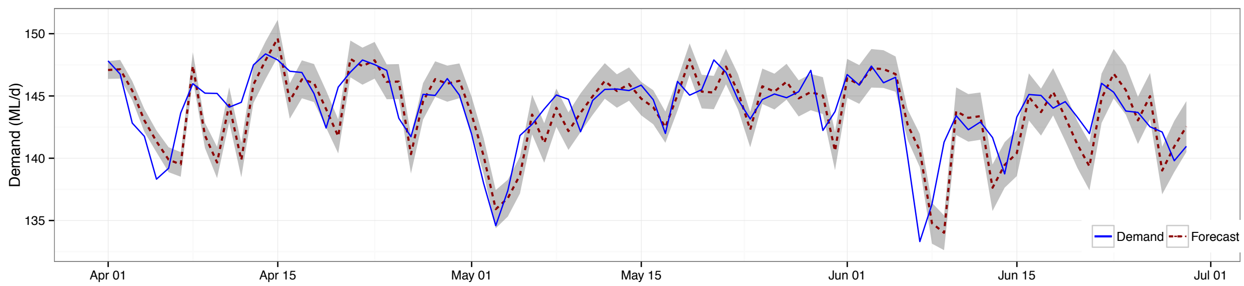


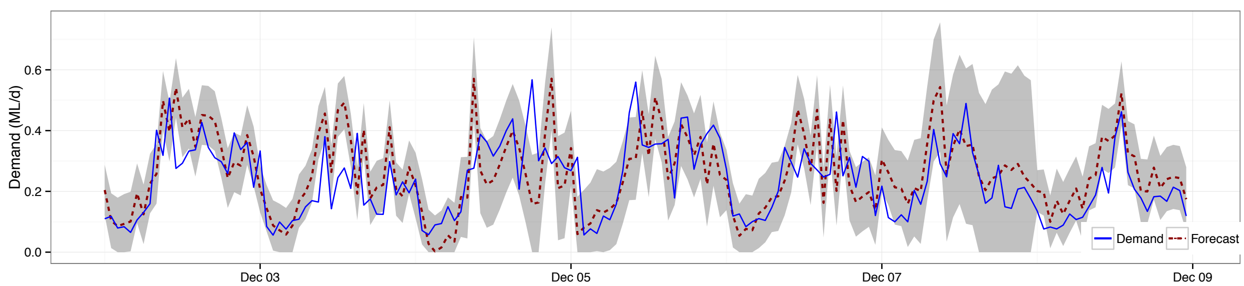
FIG. 4. ACF of a segment of training data (source A).



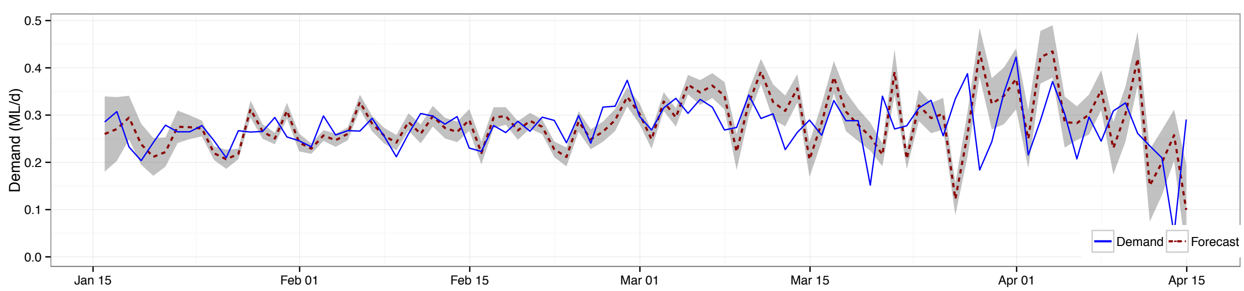
(a)



(b)



(c)



(d)

FIG. 5. Sample of the 24-h-ahead demand forecasts by the models with sub-hourly, hourly and daily resolutions. The grey bounds indicate the uncertainty (95% confidence level). (a) S-672 data source A; (b) S-7 data source A; (c) S-168 data source B; (d) S-7 data source B.

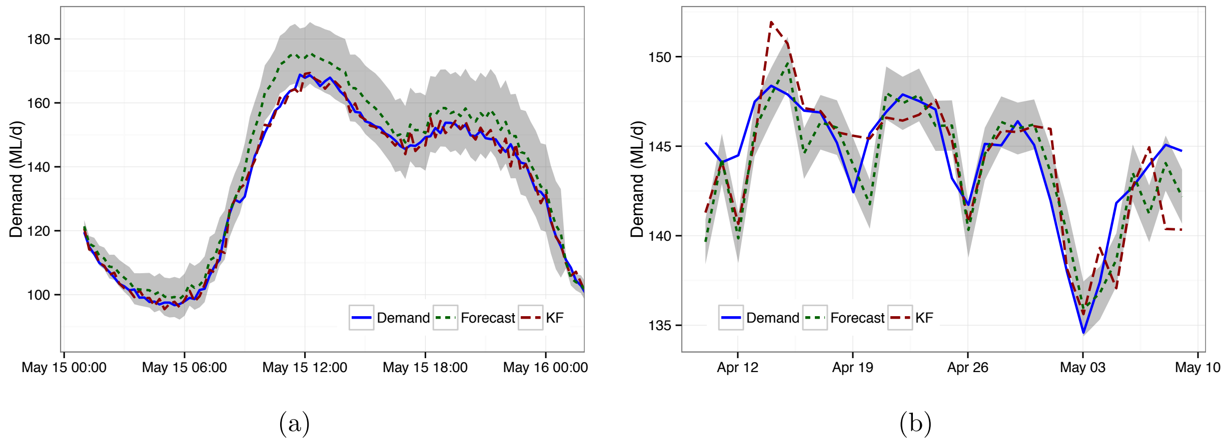
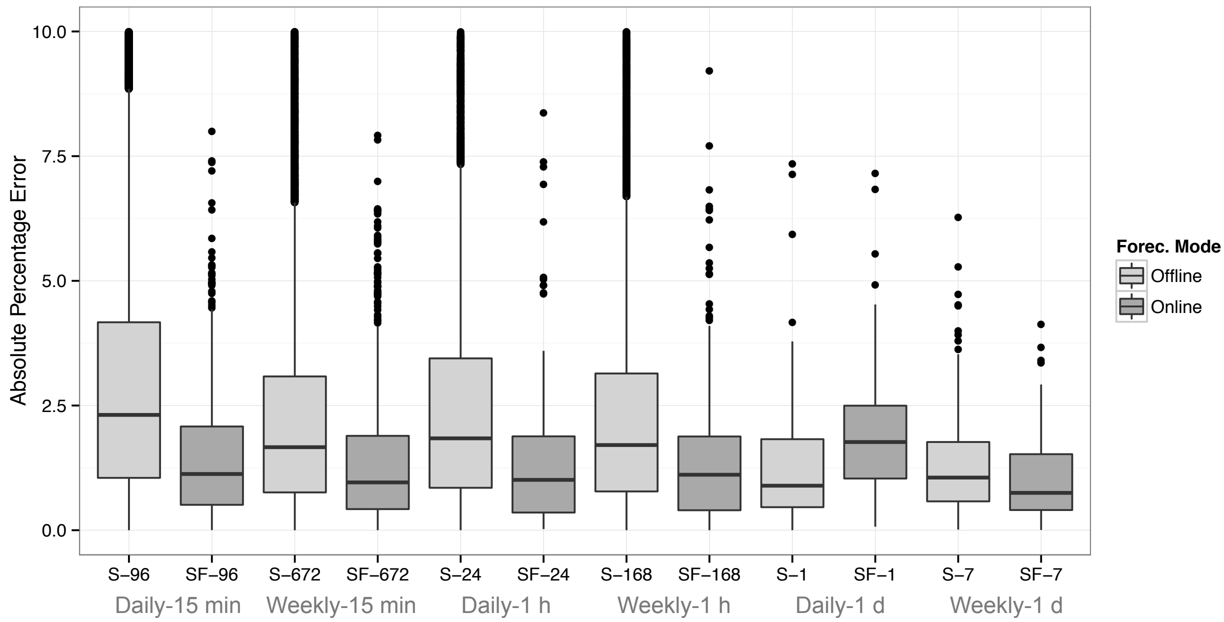
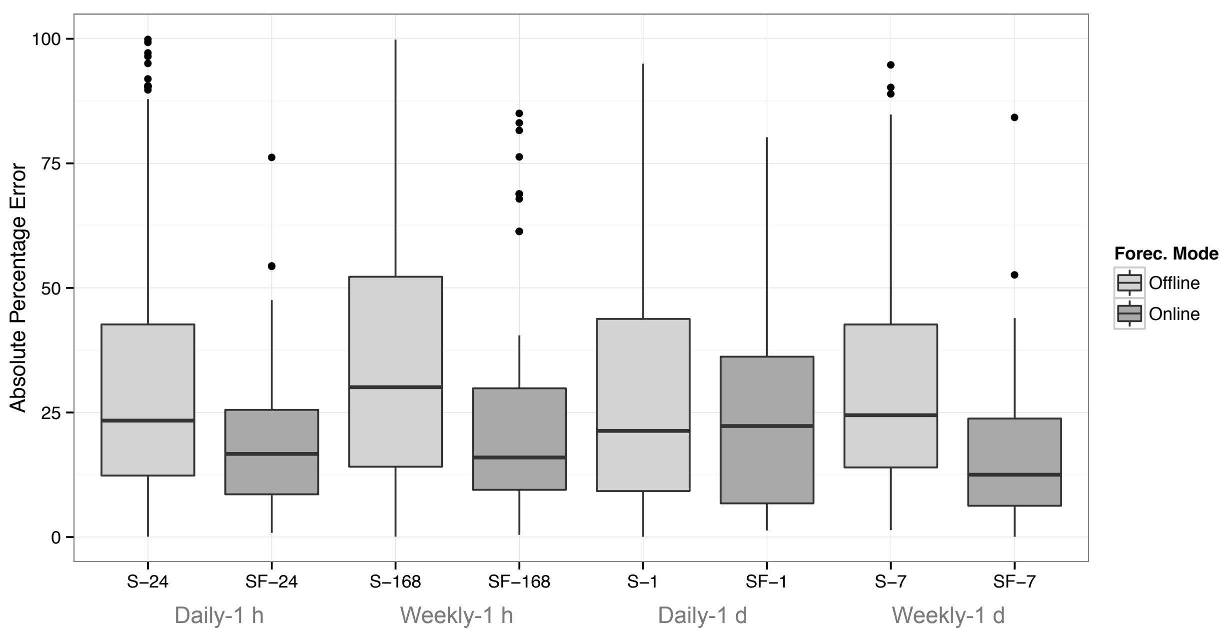


FIG. 6. Comparison of the forecasts on data source A before and after applying the Kalman filter. (a) S-672 (b) S-7.



(a)



(b)

FIG. 7. Comparison of model performance. The filtered (online) models are denoted by the prefix SF and by a different shade of grey. The model's seasonal period (daily or weekly) and data temporal resolution (15 min, 1 h or 1 d) are indicated on the x-axis. (a) Data source A; (b) Data source B.

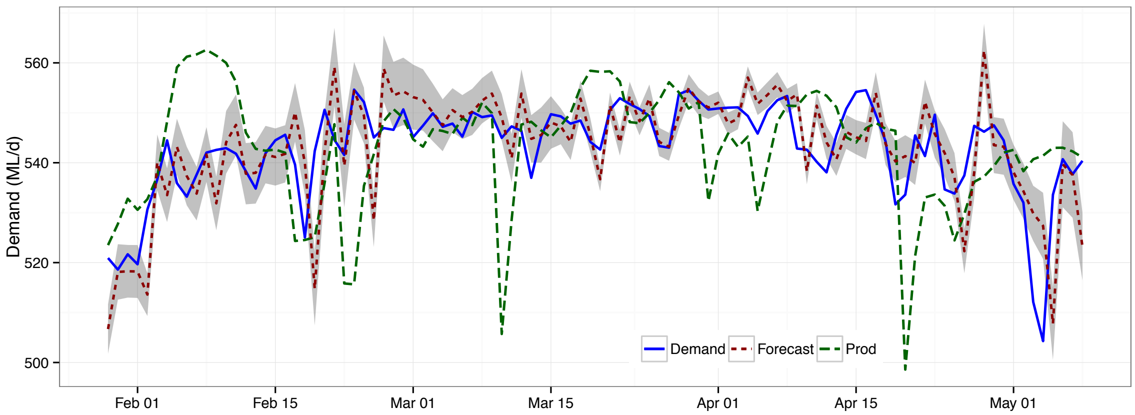


FIG. 8. Plots of total system demand, water production, and forecasts at daily level.