A Water Demand Prediction Model for Central Indiana

Setu Shah, Mahmood Hosseini, Zina Ben Miled
Department of Electrical and Computer Engineering,
Indiana University Purdue University - Indianapolis.
Email: setshah@iupui.edu, hosseima@iupui.edu,
zmiled@iupui.edu

Rebecca Shafer, Steve Berube
Citizens Energy Group,
Indianapolis.
Email: rschaefer@citizensenergygroup.com,
sberube@citizensenergygroup.com

Abstract

Due to the limited natural water resources and the increase in population, managing water consumption is becoming an increasingly important subject worldwide. In this paper, we present and compare different machine learning models that are able to predict water demand for Central Indiana. The models are developed for two different time scales: daily and monthly. The input features for the proposed model include weather conditions (temperature, rainfall, snow), social features (holiday, median income), date (day of the year, month), and operational features (number of customers, previous water demand levels). The importance of these input features as accurate predictors is investigated. The results show that daily and monthly models based on recurrent neural networks produced the best results with an average error in prediction of 1.69% and 2.29%, respectively for 2016. These models achieve a high accuracy with a limited set of input features.

1 Introduction

An accurate water consumption prediction model can help planners meet user demands. Previous research indicated that water consumption is strongly correlated with weather conditions such as temperature and precipitation (Morgan and Smolen 1976), (Hansen and Narayanan 1981), (Bakker et al. 2014). The focus of this paper is on developing an accurate prediction model that can support operational decision-making. Two types of prediction models are developed: a feed-forward back-propagation neural network and a recurrent neural network. The prediction accuracy of these models is compared to previously used water demand prediction models based on linear regression. In addition, the impact of each input feature on the accuracy of the models is discussed.

Two time scales are considered for each model: daily and monthly. The monthly prediction models are more suitable for planning. The daily models are needed to support daily operational decisions. Comparing the models at different time scales will highlight the robustness of the model in tracking changes in demand trends.

The data used to design and validate the proposed models was obtained from the Central Indiana region for the period 1997-2016. The number of customers in the service region grows to over 324,000 in 2016.

2 Related Work

Water demand forecasting is an active area of research as water consumption keeps increasing, especially in urban areas. In (House-Peters and Chang 2011), the benefits of both short-term and long-term predictions are reviewed. Short-term prediction is defined as the prediction of daily or monthly water consumption that is necessary to support operational decisions. Long-term prediction spans several years and depends on the forecasted growth of geographical regions. The authors of (House-Peters and Chang 2011) suggest that economic factors such as water price and household income are key features in water demand prediction. Similarly, (Arumugam et al. 2017) analyzed the spatio-temporal patterns of water consumption across the US and concluded that the efficiency of water-usage is higher in urban areas.

Regression models have traditionally been used in the prediction of water consumption levels. Some of the early research in the field (Morgan and Smolen 1976), (Hansen and Narayanan 1981) used temperature and rainfall as weather features for these regression models. In (Morgan and Smolen 1976), regression models based on climatic indicators such as temperature and precipitation and potential evapotranspiration (sum of evaporation and plant transpiration) minus precipitation were developed. In (Hansen and Narayanan 1981), a multivariate regression model that includes the daily water demand of Salt Lake City, average temperature, total precipitation, and percentage of daylight hours as input features was proposed.

The importance of weather features for water demand forecasting was discussed in (Bakker et al. 2014). The authors tested three models with and without the inclusion of weather features. The results show that the models which take into consideration weather features outperformed the models without weather features.

While several models have been proposed as predictors for water consumption, the importance of each input feature in the model is not well documented. In this paper, we use an extended feature set that includes temperature, customer count, median income, holidays, precipitation in both forms (rainfall and snow), and the previous day's water consumption. Moreover, we analyze the importance of each of

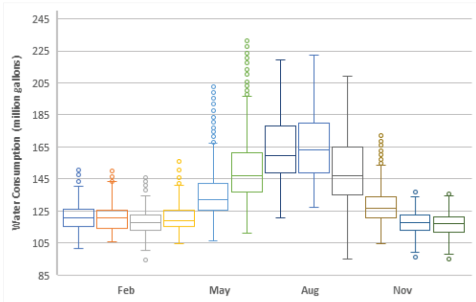

Figure 1: Monthly water consumption from 1997-2016.

these features in the prediction of both daily and monthly consumption. The models investigated in this paper include feed-forward back propagation neural networks (FFNN) and recurrent neural networks (RNN). To the best of our knowledge, recurrent neural networks have not been used for water demand forecasting in previous work.

3 Data Set

The water consumption data set was collected for the central Indiana region. The data included the daily water consumption (in million gallons) along with the customer count in the service area from 1997-2016. Figure 1 shows the minimum, lower quartile, median, upper quartile and maximum water consumption for each month. The outliers in the data are shown as dots in the graph. As can be seen from Figure 1, the water consumption is the highest during the summer months of May through September, with a large number of outliers. The deviation from the median values is also higher for these months. For the remainder of the year (i.e., January through April and October through December), the consumption is largely constant with a small deviation from the median values.

The other component of the water consumption data set is the customer count which is available only on a yearly basis with an average increase of 1.4% from one year to the next.

The second dataset is related to weather. Weather data from seven weather stations within the service area were obtained from the National Climatic Data Center (NCEI 2017) database. The daily values of minimum temperature, maximum temperature, rainfall, snow and snow depth from each of the seven stations for each day were extracted, when available. The mean values across all of the seven stations were then calculated and used as the second data set. For some stations, data were missing for some days (i.e., were not recorded for the day). These missing values were not included in the calculation of the daily mean values.

The third dataset consists of the median household income of the region, and was obtained from the U. S. Census Bureau (Bureau 2017). The highest median income was in 2007 and the lowest was in 1997. The median income for 2016 was unavailable at the time this paper was written, and thus the data from 2016 was not used. However, we do present testing results for 2016 using a variant model that does not include median income as an input feature.

The fourth data set includes all the major public holidays, bank holidays and observances (e.g., Halloween).

4 Daily Prediction Models

Two daily neural network models were developed. The first is a feed-forward model with back propagation (Larose 2014). The second is a recurrent network that is based on the approach presented in (Elman 1990). A regression model was also developed and the accuracy of all models was compared.

All the network models investigated in this paper include three-layers (input layer, hidden layer and output layer). Moreover, the first 14 years of the data-set (from 1997 to 2010) were used for training and the remaining 5 years (from 2011 to 2015) were used for the testing. The output of the nodes in the hidden layer is calculated as follows:

$$H_j = f(h_j) \quad where \quad h_j = \sum_{i=1}^{n} w_{ij}^h \cdot x_i \tag{1}$$

where x_i represents the input node i and w_{ij}^h represents the value of the weight function from the input layer node x_i to the hidden layer node H_j, n is the total number of input nodes and f is the activation function. Similarly, the output of the network is calculated as follows:

$$\tilde{O} = f(\tilde{o}) \quad where \quad \tilde{o} = \sum_{j=1}^{m} w_j^o \cdot H_j \tag{2}$$

where $\tilde{O}$ is the output node and w_j^o represents the weight function from the hidden layer node H_j to the output node, and m is the total number of hidden nodes.

The accuracy of the models is measured by comparing the predicted values to the actual observations. The error (e) for each predicted value and the average error ($\bar{e}$) for each model are calculated as follows:

$$e = \frac{|\hat{O} - \tilde{O}|}{\hat{O}} \times 100 \quad and \quad \bar{e} = \frac{1}{p}\sum_{k=1}^{p} e_k \tag{3}$$

where $\hat{O}$ is the observed water demand, p is the number of data points and e_k is the error for the corresponding data point k (i.e., day or month).

In addition, the coefficient of determination score (C_D) (Pedregosa et al. 2011) is used to evaluate each model. A C_D value of 1.0, the maximum possible score, shows high correlation between the predicted values and the input features. In case of poor correlation, the C_D can be 0 or negative.

$$C_D = 1 - \frac{\sum(\hat{O} - \tilde{O})^2}{\sum(\hat{O} - \hat{O}_{mean})^2} \tag{4}$$

The importance of each input feature in the prediction of the output is evaluated by using the following metric:

$$IW_i = \frac{\sum_{j=1}^{m} |w_{ij}^h|}{\sum_{k=1}^{n}\sum_{j=1}^{m} |w_{kj}^h|} \times 100 \tag{5}$$

where the numerator corresponds to the weights from the input layer to the hidden layer for a specific input feature i.

4.1 Multiple Linear Regression

A multiple linear regression model (dMLR) is developed for the data set using the approach proposed in (Pedregosa et al. 2011). The input features used for this model are day of the year ($DoYr$), maximum temperature ($MxTP$), minimum temperature ($MnTp$), rainfall (Rnf), snow ($Snow$), snow depth (SnD), number of customers (Cst), median income (Inc) and holiday (Hol).

4.2 Feed-Forward Neural Network

The second model is a feed-forward neural network with back propagation (dFFNN). As discussed in (Larose 2014), choosing a large number of hidden nodes would increase the complexity of the model whereas the network's ability to learn may be affected by a reduced number of hidden nodes. After examining different configurations, a model with 12 hidden nodes was selected. This configuration was compared to models with 8 and 16 hidden nodes. The average errors ($\bar{e}$) for the dFFNN models with 8, 12 and 16 hidden nodes were 9.17%, 8.84% and 9.54%, respectively.

A single dFFNN was used for the entire data-set. A model consisting of multiple networks, one for each month, was also investigated. The result of this investigation showed that a single dFFNN produced better results and suggests that FFNNs perform better when trained on a large data set as opposed to a data set segmented on a monthly basis.

The same input features used for dMLR were also used for dFFNN with the addition of a bias with a value of one. Normalized data was fed to the input nodes of the network and the hyperbolic tangent ($\tanh$) was used as the activation function for these nodes. The hidden nodes and the output node of the network are calculated by using equations 1 and 2, respectively with a Sigmoid as the activation function.

4.3 Recurrent Neural Network

The third model is a recurrent neural network (dRNN). A different recurrent neural network is developed for each month. The hidden nodes used for the prediction of the previous day are used as input in the calculation of the next day's predicted water consumption level.

The first nine of the input nodes correspond to $DoYr$, $MxTp$, $MnTp$, Rnf, $Snow$, SnD, Cst, Inc, Hol. Adding a bias and the previous day's hidden layer output (PHL) brings the total number of input nodes to sixteen for dRNN. The hidden layer is composed of six nodes and a bias node. The hyperbolic tangent ($\tanh$) function was used as the activation function for the input, hidden and output layers.

As in the case of the dFFNN model, Offline training is performed using the data spanning 1997–2010. This step is followed by an online training and testing step using the data from 2011 to 2015. The online training process updates weights once for each data point.

4.4 Analysis and Discussion

The average prediction errors ($\bar{e}$) for each month of the year for all models are included in Table 1. The last row of the table shows the cumulative average error for the entire year. This table shows that the cumulative average error of dMLR during 2011–2015 is 12.73%. In most cases, the predicted water demand is much lower than the observed water demand. The average errors ($\bar{e}$) for January and February are low (7.5%). However, the error increases from March and remains high through December. The model is not able to adjust to dynamic changes in water consumption and the errors increase from one year to the next. Furthermore, dMLR has a coefficient of determination (C_D) of 0.26. This is an indication that the model is disregarding the input features.

Table 1: Monthly average error (Avg ($\bar{e}$) for all daily models during the period 2011–2015.

	dMLR	*dFFNN*	*dRNN*	*dMRNN*
Jan	7.58	8.76	4.25	3.34
Feb	7.55	6.89	3.74	2.32
Mar	11.78	7.58	3.74	2.30
Apr	18.37	7.59	2.85	2.15
May	15.97	10.58	3.52	3.27
Jun	12.75	8.85	4.50	4.45
Jul	12.08	11.34	3.98	3.47
Aug	10.60	10.98	4.08	3.91
Sep	11.74	10.93	5.74	4.91
Oct	16.47	7.68	3.34	2.50
Nov	15.16	7.53	3.62	2.61
Dec	12.38	7.25	2.77	2.69
Cum	*12.73*	*8.83*	*3.84*	*3.17*

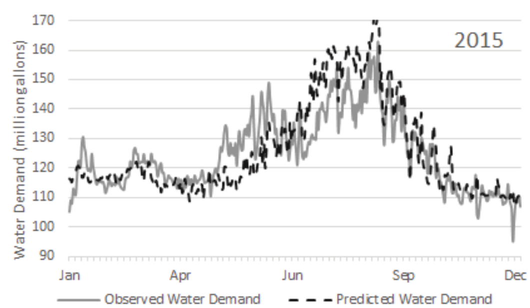

Figure 2: dFFNN water consumption predictions for 2015.

The dFFNN model resulted in a cumulative average error of 8.83% (Table 1). The months of February through April and October through December have low average errors. For the remaining months, the average error is higher than 8% but less than 12%. It was also found that dFFNN had difficulties adapting to changes in water demand for months with higher average temperature (Figure 2).

Table 2: IW for each input feature for the dFFNN model.

DoYr	*MxTp*	*MnTP*	*Rnf*	*Snow*
16.31	7.55	7.18	18.68	2.37
SnD	*Cst*	*Inc*	*Hol*	*Bias*
0.99	14.63	9.45	0.26	22.59

Table 2 shows the importance of the weight (IW) of each input feature in dFFNN as per Equation 5. This table indicates that Snow, Snow depth and Holiday are not as significant as the remaining features.

The online training and testing of the third model, dRNN, for the period 2011–2015 produced a cumulative average error ($\bar{e}$) of less than 4% (Table 1). An example of the predicted water demand compared to the observed water demand for 2015 is shown in Figure 3. The error kept decreasing as the model was trained with data from additional years.

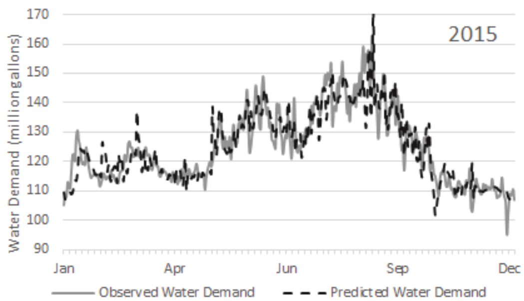

Figure 3: dRNN water consumption predictions for 2015.

Table 3: IW for each input feature of the dRNN model.

	DoYr	*MxTp*	*MnTp*	*Rnf*	*Snow*
Jan	10.85	17.25	6.38	4.48	6.60
Feb	9.44	4.26	7.96	8.63	12.58
Mar	13.86	4.45	9.59	9.67	6.85
Apr	10.94	7.27	8.24	8.60	3.35
May	10.48	20.28	3.46	22.03	9.94
Jun	7.12	12.70	0.86	27.96	12.94
Jul	0.86	4.90	7.99	36.38	15.07
Aug	5.59	14.53	1.41	5.58	8.02
Sep	3.79	2.52	0.16	45.97	14.49
Oct	41.29	4.95	2.23	18.08	5.19
Nov	20.33	10.39	2.42	12.15	4.53
Dec	13.85	5.62	2.53	9.79	3.07
Cum	*10.10*	*6.01*	*2.73*	*24.10*	*8.78*

	SnD	*Cst*	*Inc*	*Hol*	*Bias*	*PHL*
Jan	8.93	6.51	3.89	6.12	5.77	23.23
Feb	3.06	2.86	1.34	14.13	15.61	20.14
Mar	14.24	4.87	7.58	6.57	5.04	17.30
Apr	4.50	2.59	4.74	24.28	3.87	21.64
May	9.81	1.01	0.96	0.91	10.54	10.57
Jun	12.25	1.45	0.29	0.83	13.07	10.53
Jul	15.24	0.51	0.43	0.35	15.49	2.79
Aug	7.20	3.70	1.18	4.34	3.89	44.56
Sep	14.59	0.31	0.26	1.38	15.32	1.22
Oct	5.46	4.94	3.29	0.58	5.31	8.69
Nov	3.38	7.70	3.45	7.57	3.88	24.20
Dec	5.88	5.98	15.07	13.07	4.03	21.12
Cum	*9.79*	*3.21*	*6.29*	*6.79*	*9.37*	*12.83*

Table 3 shows the importance of the feature weights for the dRNN model. The results indicate that input features such as day of the year as well as rainfall and snow when considered together are the most significant features in the prediction of the water demand, while features like minimum temperature, total number of customers, median income and holiday are not as significant. Also, the significance of the previous hidden layer (PHL) is almost 13%.

The largest average errors are observed for the summer months. For these months, the dependence of the prediction on previous day's water demand (PHL) varies from 1.22% to 44.56%. From Figure 1, it can be seen that the May–September months are those with the highest water consumption. These months also have a large number of outliers which led to high prediction errors. Typically, high errors are observed for days with maximum temperature greater than 29°C. While some days with such high maximum temperatures have unusually high water consumption, other days have normal or low consumption. The maximum temperature has low IW for the months of February, March, July, September and October. Despite this varying importance of the maximum temperature throughout the year, the high value of IW during January, May, June, August and November indicates a strong dependence of the model on maximum temperature. Moreover, for the May–September period, which has no snowfall or snow depth, the high IW of these two input features suggests that they are behaving like negative biases in the model.

Table 3 also shows that the low demand months (i.e., January–April and November–December), have similar IW values for previous day's hidden layer (PHL) nodes (18-23%). This confirms that the previous day's consumption is an important feature in the prediction of water demand.

Based on the results of Table 3, input features with low IW were omitted. A modified configuration of the model (dMRNN) consisting of four input nodes: 1) day of the year as a bitonic function ranging from 1-183, 2) maximum temperature, 3) precipitation ($Prcp$) as the sum of rainfall and snow, and 4) an input bias was created.

Table 4: IW for each input feature of the dMRNN model.

	DoYr	*MxTp*	*Prcp*	*Bias*	*PHL*
Jan	29.53	2.42	0.50	26.41	41.15
Feb	16.39	2.57	6.64	14.08	60.32
Mar	9.82	1.68	2.66	9.92	75.92
Apr	10.28	6.65	6.62	20.60	55.85
May	10.97	30.79	3.67	37.33	17.24
Jun	15.82	27.21	1.97	40.79	14.21
Jul	15.86	39.01	2.46	26.77	15.90
Aug	4.83	24.73	3.02	28.33	39.09
Sep	6.93	23.86	5.72	31.31	32.18
Oct	16.06	14.26	4.79	20.60	44.29
Nov	13.68	2.08	5.80	14.13	64.31
Dec	12.65	3.32	1.70	15.72	66.61
Cum	*13.23*	*22.17*	*3.39*	*29.20*	*32.01*

The values of IW obtained for the dMRNN model are shown in Table 4. These IWs show a strong correlation between water demand predictions and the day of the year ($DoYr$) as well as the previous day's hidden layer (PHL)

values throughout the year. The dependence of the water demand prediction on the maximum temperature is low for the low demand months of January–April and November–December. Whereas, the dependence on maximum temperature is much higher for the high demand months of May–September. This demonstrates the direct correlation between maximum temperature and high water demand during the summer months. Moreover, the model's dependence on precipitation ($Prcp$) is low and indicative of the lesser importance of precipitation in the prediction.

Table 1 shows that the average errors of the dMRNN model are lower than those for the dRNN model throughout the year. The cumulative average error for the dMRNN model (3.17%) is also lower. Figure 4 shows the predicted water demand for 2015 and 2016 using dMRNN. The average error reduces throughout the online training with the lowest errors observed in 2015 (Table 5). This indicates that the network learns over time, and has the ability to adjust to changes in water demand trends. Because the dMRNN model doesn't use the median income as an input feature, the predictions for 2016 were also generated resulting in an average error ($\bar{e}$) of 1.69%. Compared to that of 2015, this result reinforces the claim that dMRNN continues to learn and generate more accurate predictions over time.

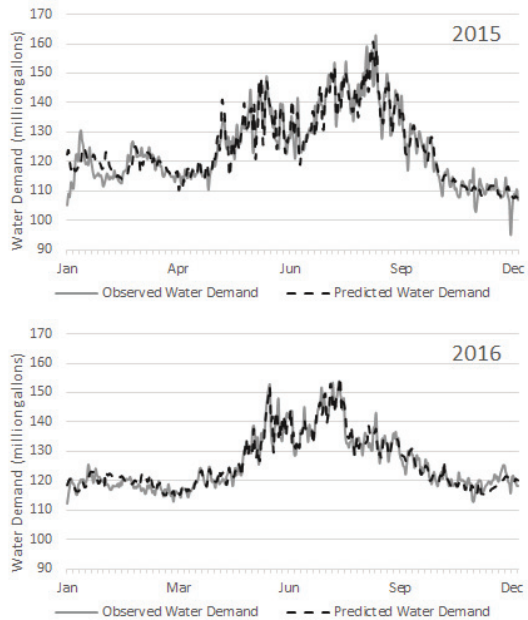

Figure 4: dMRNN water consumption predictions for 2015 and 2016.

Table 5: Yearly average error ($\bar{e}$) for all models during the online testing period of 2011–2015.

Year	*dMLR*	*dFFNN*	*dRNN*	*dMRNN*
2011	10.01	5.04	3.59	3.14
2012	16.57	13.56	4.61	3.82
2013	12.34	11.03	3.87	3.04
2014	12.07	8.14	3.81	3.20
2015	12.64	6.44	3.30	2.63
Cum	*12.73*	*8.84*	*3.84*	*3.17*
C_D	0.26	0.55	0.89	0.93

For the data set used in this study, mandatory restrictions were enforced only once during the period from 07/13/2012 to 09/05/2012. Voluntary restrictions were also advised twice between 1997–2015: from 06/14/2007 to 06/19/2007 and from 07/06/2012 to 07/12/2012. These restrictions caused higher average errors in prediction for all models including dMRNN in 2012 (Table 5). Because of the lack of training data, using restrictions as an input feature produced the exact same results as the model without restrictions.

In summary, Tables 5 shows that the average error for the dMLR model is the highest. The coefficient of determination (C_D) is also the lowest for dMLR. The coefficient of determination (C_D) of the dRNN model is 0.89, which indicates a high correlation between the input features and the predictions. The C_D and average error ($\bar{e}$) of the modified recurrent network dMRNN model improve on the dRNN model.

5 Monthly Prediction Models

The data sets for the monthly models were generated from the daily data sets. Figure 1 shows the monthly average water demand which are obtained from daily data. These monthly averages maintain the daily trend with high consumption during the summer months of May–September and relatively low consumption during the rest of the year. Average monthly values for the other input features were also derived from the daily values (e.g., average monthly maximum temperature).

Monthly models based on multiple linear regression (mMLR), feed forward neural network (mFFNN) and recurrent neural network (mMRNN) were designed. As opposed to the daily dMRNN model, the monthly mMRNN model used a single network for all months.

Table 6: Average error ($\bar{e}$) for the monthly models during the period 2011–2016.

	mMLR	*mFFNN*	*mMRNN*
2011	10.84	5.85	3.99
2012	13.93	11.26	7.79
2013	10.35	11.26	4.08
2014	12.40	8.27	4.44
2015	13.30	12.61	2.41
2016	13.45	11.58	2.29
Cum	*12.38*	*10.14*	*4.17*
C_D	0.15	0.34	0.86

Figure 5 shows the monthly predictions for all three models. The inability of the mMLR model to adapt to changing patterns can be observed in the errors of the model (Table 6) which resulted in predictions that are higher than the observed values for the period 2014–2016 (Figure 5) .

The cumulative average error of mFFNN is 10.14% (Table 6). As in the case of the daily model, when the training data set is small, the mFFNN has difficulties predicting water demand with high accuracy.

The predictions for mMRNN follow the actual demand (Figure 5). During the August–September 2012 period, the mandatory water usage restrictions resulted in lower water consumption than previous years. On the contrary, higher water demand is observed during the January-February 2014 period and unusually low water demand is seen during August 2014 compared to the same period in 2011-2013. While the model fails to predict these outliers correctly, it predicts the demands for other periods with a high level of accuracy. Developing a model that can correctly predict extreme events including the ones triggered by restrictions as well as droughts is the subject of ongoing research.

As can be seen from Figure 5 and Table 6, the mMRNN model learns the reduced water demand in 2014 and predicts the water demand of 2015 with less than 2.5% average errors. This confirms the model's ability to adapt to changing water demand patterns. For 2016, the mMRNN model continued to improve and produced an average error of 2.29%.

The C_D (Table 6) for the mMRNN model is 0.86. While this score is lower than the C_D of the daily model (dMRNN in Table 5), it still shows a high correlation between the input features and the predicted values.

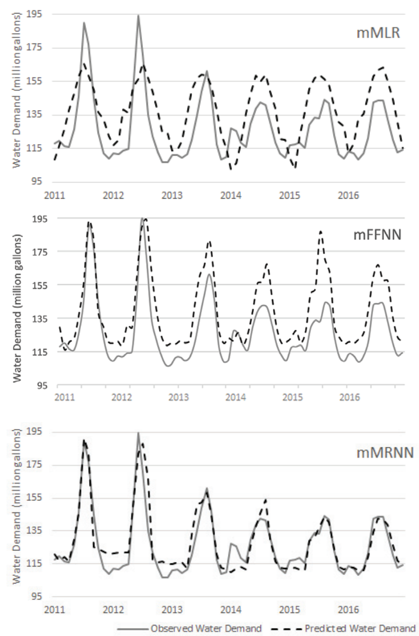

Figure 5: Water demand prediction for monthly models.

6 Conclusion

This paper compares different models for the prediction of daily water demand in an urban area by taking into consideration input features like temperature, rainfall, snow, snow depth, number of customers, median income, holidays and day of the year. The initial configuration of the daily dMLR, dFFNN and dRNN models, used all of the above input features. The average errors produced by these models were 12.73%, 6.45% and 3.84%, respectively. The highest correlation between the input features and the predicted values (C_D) is associated with the dRNN model (i.e., 0.89).

After analyzing the weights of the different input features and reducing the input feature set to maximum temperature, day of the year and precipitation, a modified recurrent neural network (dMRNN) model was proposed. This modified model showed significant improvements with an average error of 3.17% and a C_D of 0.93. Further improvement for dMRNN was observed as online training continued in 2016 and resulted in an average error of 1.69%. It was also observed that the dMRNN model shows better resilience. Indeed, this model has the ability to self-correct within 2 days of mis-predictions in nearly all cases.

For the monthly time scale, mMRNN had a higher coefficient of determination (C_D) and a lower average error ($\bar{e}$) than both mMLR and mFFNN. The average error for mMRNN was 4.17% during 2011-2015 and further reduced to 2.29% in 2016. These accuracy levels are sufficient to support daily and monthly operational decisions. Moreover, the proposed models can be integrated with a weather forecast service to form an application for water demand prediction and what-if-scenario testing.

References

Arumugam, S.; Sabo, J.; Larson, K.; Seo, S.; Sinha, T.; Bhowmik, R.; Vidal, A. R.; Kunkel, K.; Mahinthakumar, G.; Berglund, E.; et al. 2017. Synthesis of public water supply use in the us: Spatio-temporal patterns and socio-economic controls. *Earth's Future*.

Bakker, M.; Van Duist, H.; Van Schagen, K.; Vreeburg, J.; and Rietveld, L. 2014. Improving the performance of water demand forecasting models by using weather input. *Procedia Engineering* 70:93–102.

Bureau, U. C. 2017. Small area income and poverty estimates - interactive data and mapping - u.s. census bureau. https://www.census.gov/did/www/saipe/data/interactive/saipe.html. Accessed 07-01-2017.

Elman, J. L. 1990. Finding structure in time. *Cognitive science* 14(2):179–211.

Hansen, R. D., and Narayanan, R. 1981. A monthly time series model of municipal water demand. *JAWRA Journal of the American Water Resources Association* 17(4):578–585.

House-Peters, L. A., and Chang, H. 2011. Urban water demand modeling: Review of concepts, methods, and organizing principles. *Water Resources Research* 47(5).

Larose, D. T. 2014. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Morgan, W. D., and Smolen, J. C. 1976. Climatic indicators in the estlmation of municipal water demand. *JAWRA Journal of the American Water Resources Association* 12(3):511–518.

NCEI. 2017. National centers for environmental information (ncei) formerly known as national climatic data center (ncdc). https://www.ncdc.noaa.gov/. Accessed 07-01-2017.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(Oct):2825–2830.