

Urban Water Flow and Water Level Prediction Based on Deep Learning

Haytham Assem¹(✉), Salem Ghariba², Gabor Makrai³, Paul Johnston⁴,
Laurence Gill⁴, and Francesco Pilla²

¹ Cognitive Computing Group, Innovation Exchange, IBM, Dublin, Ireland
haythama@ie.ibm.com

² Department of Planning and Environmental Policy,
University College Dublin, Dublin, Ireland
{salem.ghariba, francesco.pilla}@ucd.ie

³ York Centre for Complex Systems Analysis (YCCSA),
University of York, Heslington, York, UK
gabor.makrai@york.ac.uk

⁴ Department of Civil, Structural, and Environmental Engineering,
Trinity College Dublin, Dublin, Ireland
{pjhnston, laurence.gill}@tcd.ie

Abstract. The future planning, management and prediction of water demand and usage should be preceded by long-term variation analysis for related parameters in order to enhance the process of developing new scenarios whether for surface-water or ground-water resources. This paper aims to provide an appropriate methodology for long-term prediction for the water flow and water level parameters of the Shannon river in Ireland over a 30-year period from 1983–2013 through a framework that is composed of three phases: city wide scale analytics, data fusion, and domain knowledge data analytics phase which is the main focus of the paper that employs a machine learning model based on deep convolutional neural networks (DeepCNNs). We test our proposed deep learning model on three different water stations across the Shannon river and show it out-performs four well-known time-series forecasting models. We finally show how the proposed model simulate the predicted water flow and water level from 2013–2080. Our proposed solution can be very useful for the water authorities for better planning the future allocation of water resources among competing users such as agriculture, demotic and power stations. In addition, it can be used for capturing abnormalities by setting and comparing thresholds to the predicted water flow and water level.

Keywords: Deep learning · Water management
Convolutional neural networks · Urban computing

1 Introduction

Simulating and forecasting the daily time step for the hydrological parameters especially daily water flow (streamflow) and water level with sort of high

accuracy on the catchment scale is a key role in the management process of water resource systems. Reliable models and projections can be hugely used as a tool by water authorities in the future allocation of the water resource among competing users such as agriculture, demotic and power stations. Catchment characteristics are important aspects in any hydrological forecasting and modeling process. The performance of modeling and projection methods for single hydro-metric station varies according to its catchment climatic zone and characteristics. Karran et al. [11] state that methods that are proven as effective for modeling streamflow in the water abundant regions might be unusable for the dryer catchments, where water scarcity is a reality due to the intermittent nature of streams. Climate characteristics may severely affect the performance of different forecasting methods in different catchments and this area of research still requires much more exploration. The understanding of streamflow and water level dynamics is very important, which is described by various physical mechanisms occurring on a wide range of temporal and spatial scales [20]. Simulating these mechanisms and relations can be executed by physical, conceptual or data-driven models. However physical and conceptual models are the only current ways for providing physical interpretations and illustrations into catchment-scale processes, they have been criticized for being difficult to implement for high-resolution time-scale prediction, in addition to the need too many different types of data sets, which are usually very difficult to obtain. In general, physical and conceptual models are very difficult to run and the more resolution they have, the more data they need, which leads to over parametrize complex models [1].

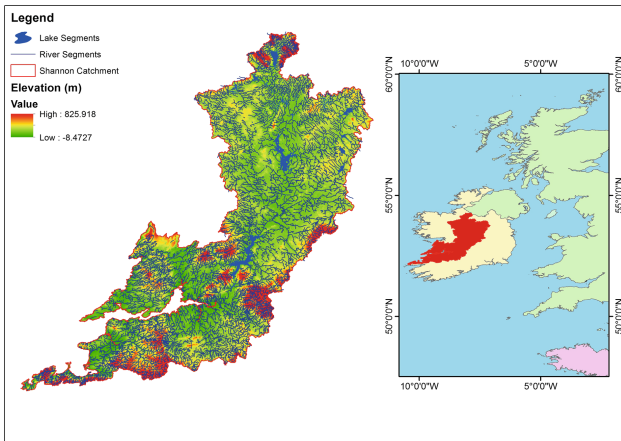


Fig. 1. Shannon river catchments and segments.

In this paper, we introduce a water management framework for the aim of providing insights of how to better allocate water resources by providing a highly accurate forecasting model based on deep convolutional neural networks (termed as DeepCNNs in the rest of the paper) for predicting the water flow and water

level for the Shannon river in Ireland, the longest river in Ireland at 360.5 km. It drains the Shannon River Basin which has an area of 16,865 km², one fifth of the area of Ireland. Figure 1 shows Shannon river segments and catchments across Ireland. To the best of our knowledge, this paper is the first to explore and show the effectiveness of the deep learning models in the hydrology domain for long-term projections by employing deep convolutional network model and comparing its performance and showing that it out-performs other well-known time series forecasting models. We organize the paper as follows: Sect. 2 reviews the related work and identify our exact contribution with respect to the state-of-the-art. Section 3 introduces our proposed framework for water management. Section 4 presents the proposed architecture of the deep convolutional neural networks. Section 5 describes the experiments illustrates our results. Finally, we conclude the paper in Sect. 6.

2 Related Work

Artificial Neural Networks (ANNs) have been used in hydrology in many applications such as water flow (stream flow) modeling, water quality assessment and suspended sediment load predictions. The first uses for ANN in hydrology is introduced initially in the early 1990s [3], which find the method useful for forecasting process in the hydrological application. The ANN then has been used in many hydrological applications to confirm the usefulness and to model different hydrological parameters, as stream flow. The multi-layer perceptron (MLP) ANN models seem to be the most used ANN algorithms, which are optimized with a back-propagation algorithm, these models are improved the short-term hydrological forecasts. Examples of recent remarkable published applications for the use of ANN in hydrology are as follows [2, 12]. Support Vector Machines (SVMs) have been recently adapted to the hydrology applications that is firstly used in 2006 by Khan and Coulibaly in [13], who state that SVR model out performs MLP ANNs in 3–12 month water levels predictions of a lake, then the use of SVM in hydrology has been promoted and recommended in many studies' as described in [4] from the use of flood stages, storm surge prediction, stream flow modeling to even daily evapotranspiration estimations.

The limited ability to process the non-stationary data is the biggest concern of the machine learning techniques applied to the hydrology domain, which leads to the recent application of hybrid models, where the input data are preprocessed for non-stationary characteristics first and then run through the post processing machine learning models to deal with the non-linearity issues. Wavelet transformation combined with machine learning models has been proven to give highly accurate and reliable short-term projections. The most popular hybrid model is the wavelet transform coupled with an artificial neural network (WANN). Kim and Valdés [14] is one of the first hydrological applications of the WANN model, which address the area of forecasting drought in the Conchos River Basin, Mexico, then many following published studies provide the application of WANN in streamflow forecasting and many research areas in hydrological modeling and

prediction. In general, all the studies that compare between the ANN and WANN conclude that the WANN models have outperformed the stand alone ANNs [3]. Furthermore, wavelet transform coupled with SVM/SVR (WSVM/WSVR) has been proposed to be used in hydrology applications. To the best of our knowledge, there is a very little research into the application of this hybrid model for streamflow forecasting and there is no application on water level forecasting.

Karran et al. [11] compares the use of four different models, artificial neural networks (ANNs), support vector regression (SVR), wavelet-ANN, and wavelet-SVR for one single station in each watershed of Mediterranean, Oceanic, and Hemiboreal watershed, the results show that SVR based models performed best overall. Kisi et al. [16] have applied the WSVR models with different methods to model monthly streamflow and find that the WSVR models outperformed the stand alone SVR. From the previous state-of-the-art work, we have concluded that the previous mentioned machine learning models (ANNs, SVMs, WANNs, and WSVMs) are the most well-studied and well-known in the field of hydrology. Hence, we build in this paper four baselines employing the previous mentioned models for having a fair comparison for our proposed deep convolutional neural networks across three various water stations. To the best of our knowledge, this paper is the first to adapt Deep Learning technique in the hydrology domain and showing better accuracy across three water stations compared to state-of-the-art models used in the hydrology applications.

3 Water Management Framework

In this section, we summarize the three phases for the proposed framework for predicting water flow and water level through multistage analytics process.

(a) City wide scale data analytics: This phase is composed mainly of two steps, the first step utilize the dynamically spatial distributed water balance model integrating the climate and land use changes. This stage use a wide range of input parameters and grids including seasonally climate variables and changes, land use and its seasonal parameters and future changes, seasonal groundwater depth, soil properties, topography, and slope. The output of this step is several parameters including runoff, recharge, interception, evapotranspiration, soil evaporation, transpiration including total uncertainties or error in the water balance. We utilize *runoff* from this step as an extracted feature to be passed to the data storage (please refer to [6] for the description of the used model). In the second step, we gathered the data for the *temp-max* and *temp-min* from Met Eireann¹ from 1983–2013, the national meteorological service in Ireland. We further simulated the future temperatures from 2013–2080 using statistical down scaling model as described in [7]. **(b) Data Fusion:** In this phase, we follow a stage-based fusion method [22] in which we fused the features extracted from the previous stage with the two observed outputs for water flow and water level from 1983–2013. Furthermore, we normalize and scale the data and store it in a

¹ <http://www.met.ie>.

data-storage for further being processed by the next phase. **(c) Domain knowledge data analytics:** This phase is our main focus for the paper in which we consume the features stored in the data-storage and train our proposed model along with the baseline models for the aim of predicting water flow and water level across three different water stations.

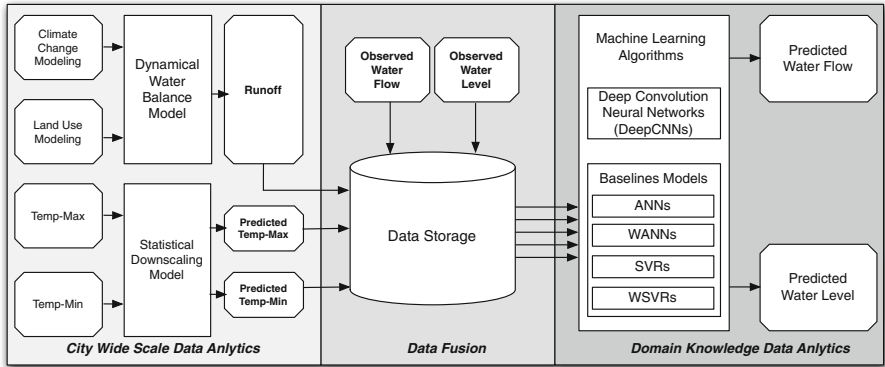


Fig. 2. Water management framework.

4 Deep Convolutional Neural Networks

In order to design an effective forecasting model for predicting water flow and water level across several years, we needed to exploit the time series nature of the data. Intuitively, analyzing the data over a sufficient wide time interval rather than only including the last reading would potentially lead to more information for the future water flow and water level. A first approach is that we concatenate various data samples together and feed them to a machine learning model, this is what we did in the baseline models which boosts the performance achieved. To achieve further improvements, we make use of the adequacy of convolutional neural networks for such type of data [18]. We propose the following architecture, each input sample consists of 10 consecutive readings concatenated together (10 worked best on our datasets). Each of the three input features (Temp-max, Temp-min, and Run-off) is fed to the network to a separate channel. The resulting dataset is a tensor of $N \times T \times D$ dimensions, where N is the number of data points (the total number of records minus the number of concatenated readings). T is the length of the concatenated strings of events and D is the number of collected features. Each of the resulting tensor records, of dimensionality $1 \times T \times D$ is processed by a stack of convolution layers as shown in Fig. 3.

The first convolution layer utilizes a set of three-channel convolution filters of size l . We do not employ any pooling mechanisms since the dimensionality of the data is relatively low. In addition, zero padding was used for preserving the input data dimensionality. Each of these filters provides a vector of length 10, each of its

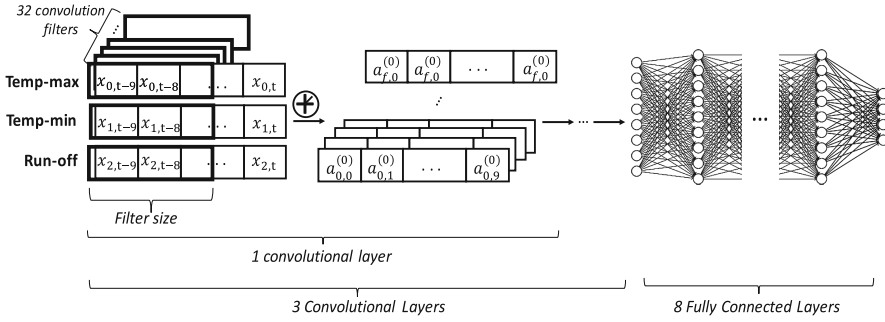


Fig. 3. The proposed convolutional neural network architecture (DeepCNNs).

elements further goes to non linear transformation using ReLu [19] as a transfer function. The resulting outputs are further processed by another similar layers of convolutional layers, with as many channels as convolution filters in the previous layer. Given an input record x , we can therefore define the entries output by filter f of convolution layer l at position i as shown in Eq. 1. Finally, the last convolution layer is flattened and further processed through a feedforward fully connected layers.

$$a_{f,i}^{(l)} = \begin{cases} \phi(\sum_{j=0}^2 \sum_{k=0}^{c-1} w_{fjk}^{(l)} x_{j,i+k-c/2} + b_{fl}), & \text{if } l = 0 \\ \phi(\sum_{j=0}^{n^{(l-1)}-1} \sum_{k=0}^{c-1} w_{fjk}^{(l)} a_{j,i+k-c/2}^{(l-1)} + b_{fl}), & \text{otherwise} \end{cases} \quad (1)$$

where ϕ is the non-linear activation function. $x_{j,i}$ is the value of a channel (which corresponds to a feature) j at position i of the input record (if i is negative or greater than 10, then $x_{j,i} = 0$). $w_{fjk}^{(l)}$ is the value of channel j of convolution filter f of layer l at position k , and b_{fl} is the bias of filter f at layer l . $n^{(l)}$ is the number of convolutions filters at layer l .

5 Experiments

In this section, we first describe the dataset used in our experiments, then we give an overview on the used baseline models, and finally we show our results discussing the key findings and observations.

5.1 Dataset

Following the procedures described in Fig. 2, the resulted datasets stored in the data storage are comprised of five parameters named, max-temp, min-temp, run-off, water flow and water level where the first three represents the featuresof the

trained models while the later two represents the outputs of the models. The used parameters can be defined as follows:

- **max-temp, min-temp:** These are the highest and lowest temperatures recorded in °C during each day in the dataset.
- **run-off:** Runoff is described as the part of the water cycle that flows over land as surface water instead of being absorbed into groundwater or evaporating and is measured in *mm*.
- **water flow:** Water flow (streamflow) is the volume of water that moves through a specific point in a stream during a given period of time (one day in our case) and is measured in m^3/sec .
- **water level:** This parameter indicates the maximum height reached by the water in the river during the day and is measured in *m*.

The previous parameters in the dataset are for 30-years (1983–2013) resulting in 11,392 samples where each sample represents a day. The datasets formulated related to three different water hydrometric stations named, Inny, lower-shannon, and suck.

5.2 Baselines

In this section we describe the baseline models that have been developed for assessing the performance of the proposed deep convolutional neural network. We choose two very popular ordinary machine learning algorithms that has already shown success in hyrdology, Artificial Neural Networks (ANNs) [8] and Support Vector Machines (SVMs) [4]. In addition, we choose two wavelet transformation models that have shown stable outcomes and in particular for the time-series forecasting problems, Wavelet-ANNs (WANNs) and Wavelet-SVMs (WSVMs).

- **ANNs:** We developed three layer feed-forward neural network employing backpropogation algorithm. An automated RapidMinder algorithm proposed in [17] is utilized for optimizing the number of neurons in the hidden layer with setting the number of epochs to 500, learning rate to 0.1 and momentum to 0.1 as well.
- **SVMs:** We developed SVM with non-linear dot kernel which requires two parameters to be configured by the user, namely cost (C) and epsilon (ϵ). We set C to 0.0001 and ϵ to 0.001. The selected combination was adjusted to the most precision that could be acquired through a trial and error process for a more localized optimization for the model parameters.

We used Discrete Wavelet Transforms (DWTs) to decompose the original time series into a time-frequency representation at different scales (wavelet sub-times series). In this type of baselines, we set the level of decomposition to 3, two levels of details and one level of approximations. The signals were decomposed using the redundant trous algorithm [5] in conjunction with the non-symmetric db1 wavelet as the mother function². Three sets of wavelet sub-time series were

² The using of the átrous algorithm with the db1 wavelet mother function is a result of the optimizing Python Wavelet tool [9].

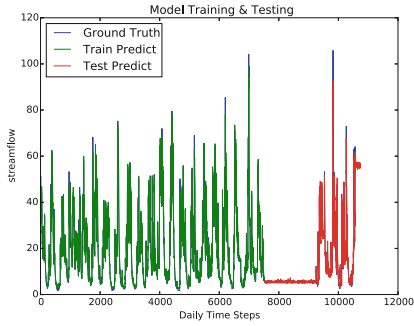
created, including a low-frequency component (Approximation) that uncovers the signal’s trend, and two sets of high-frequency components (Details). The original signal is always recreated by summing the details with the smoothest approximation of the signal. All the input time series are gone through the designed wavelet transform and the resulted sub datasets have been used by the following models:

- **WANNs:** The decomposed time series are fed to the ANN method for the prediction of water flow and water level for one day ahead. The WANNs model employs discrete wavelet transform to overcome the difficulties associated with the conventional ANN model, as the wavelet transform is known to overcome the non-stationary properties of time series.
- **WSVMs:** The WSVR are built in the same way as the WANN model.

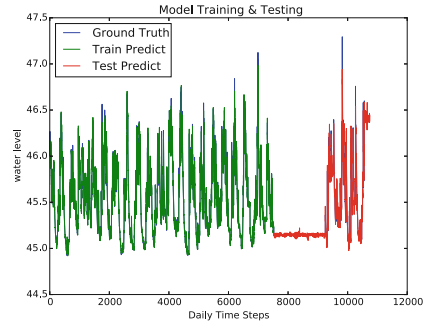
5.3 Results and Discussion

We followed the previous design for the convolutional neural network in which we performed a random grid search of the hyperparameter space and choose the best performing set. We found that the best performing model is composed of 3 convolutional layers, each of which learns 32 convolution patches of width 5 employing zero padding. After the convolutional layers, we employed 8 fully stacked connected layers. The convolutional layers are regularized using dropout technique [21] with a probability of 0.2 for dropping units. All dense layers employ L2 regularization with $\lambda = 0.000025$. All layers are batch normalized [10] and use ReLu units [19] for activation with an exception for the output layer because it is a regression problem and we are interested in predicting numerical values directly without transform. The efficient ADAM optimization algorithm [15] is used and a mean squared error loss function is optimized with a minibatches of size 10. We set aside 30% from the whole data for testing the performance of the trained model while the 70% rest of the data act as the training dataset. From the training dataset, we select 90% for training each model, and the remaining 10% as the validation set, we used it to export the best model if any improvements on the validation score, we continue the whole process for 200 epochs. In addition, we reduce the learning rate by a factor of 2 once learning stagnates for 20 consecutive epochs. Figures 4, 5 and 6 show the output of the previous training process for the Inny, lower-shannon, and suck water stations respectively where the x axis represents the daily time steps while y axis indicates the output whether it is the water flow (streamflow) or water level. The blue line in the figures indicates the original dataset (ground truth), the green indicates the output of the model on the training dataset, while the red line indicates the output of the model on the test data in which it has not been exposed at all to the model during the training procedures.

We compare our proposed model with the other baseline models described in the previous section. We use the following three evaluation metrics for our comparisons: (a) *Root-mean-square error (RMSE)*: is the most frequently used

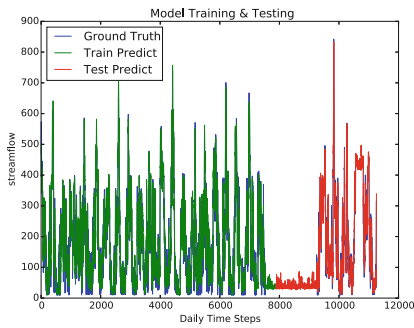


(a) Water flow

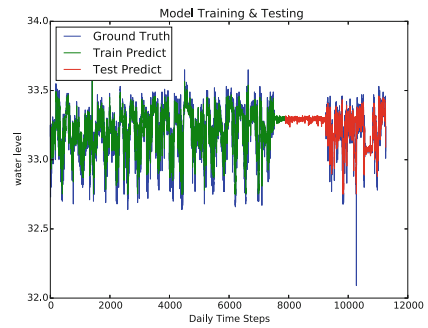


(b) Water level

Fig. 4. Inny water station.

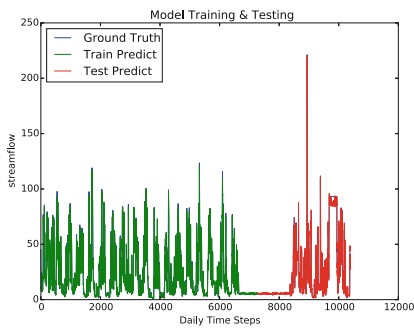


(a) Water flow

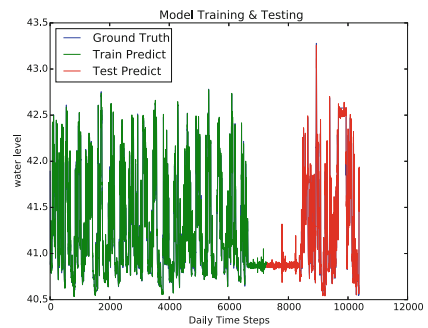


(b) Water level

Fig. 5. Lower-shannon water station.



(a) Water flow



(b) Water level

Fig. 6. Suck water station.

Table 1. Comparison between baseline models and DeepCNNs.

Water station	Model	Water flow			Water level		
		RMSE	MAE	R^2	RMSE	MAE	R^2
Inny	ANNs	2.721	1.249	0.977	0.061	0.025	0.982
	SVMs	2.712	0.956	0.977	0.06	0.022	0.983
	WANNs	2.785	1.389	0.977	0.061	0.026	0.982
	WSVMs	2.673	0.933	0.978	0.06	0.023	0.983
	DeepCNNs	2.14	0.92	0.98	0.05	0.02	0.99
Lower-Shannon	ANNs	27.1	16.665	0.974	0.063	0.039	0.853
	SVMs	29.782	18.191	0.969	0.066	0.037	0.842
	WANNs	27.335	16.622	0.973	0.063	0.038	0.854
	WSVMs	30.715	19.89	0.968	0.065	0.036	0.842
	DeepCNNs	22.30	13.43	0.98	0.05	0.03	0.87
Suck	ANNs	4.25	2.09	0.985	0.08	0.042	0.986
	SVMs	3.831	1.783	0.987	0.079	0.031	0.986
	WANNs	4.252	1.954	0.985	0.079	0.039	0.986
	WSVMs	4.075	1.469	0.985	0.08	0.031	0.986
	DeepCNNs	3.46	1.43	0.99	0.06	0.03	0.99

metric for assessing time-series forecasting models which measures the differences between values predicted by a model and the values actually observed. (b) *Mean absolute error (MAE)*: is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. (c) *Coefficient of determination (R^2)*: is a metric that gives an indication about the goodness of fit of a model in which a closer value to 1 indicates a better fitted model. Table 1 illustrates the results of the comparisons between our proposed model and all baselines across the previous described three performance metrics for the three different water stations. Interestingly, we noticed that the proposed deep convolutional neural network model outperforms all baselines across the three different performance metrics. This suggests that predicting water flow and water level in rivers manifests itself in a complex fashion, and motivates further research in the application of deep learning methods to the water management domain. From such comparison, it is observed as well that SVMs is the second best performing model for the Inny and Suck water stations. ANNs is the second best performing model for the lower-shannon water station for both outputs.

Finally, and based on the forecasted/simulated values for the features (Temp-max, Temp-min and run-off) from 2013–2080, we show in Fig. 7a and b the prediction for water flow and water level respectively for the lower-shannon station employing our trained model based on the DeepCNNs. Based on the predictions by our proposed model, it is worth noting from Fig. 7a that there will be a significant increase in the water flow crossing 250 m³ in several days across 2028, 2040 and 2059 while a less but still significant increase across several days in 2047, 2048, 2076 and 2078. It could be observed as well from Fig. 7b that there will be

a significant rise of water level crossing 33.4m in several days in 2021 and bit less in 2032, 2044, 2045 and others as well. These results should be very useful for further being assessed by water authorities for building mitigation plans for the impact of such increase as well as better planning for water allocation across various competing users.

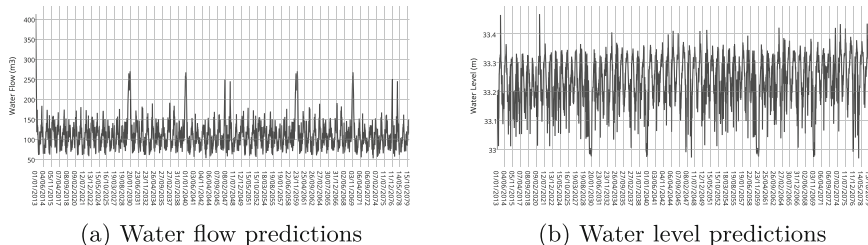


Fig. 7. Predictions of water flow and water level for lower-shannon water station from 2013–2080 using the DeepCNNs proposed model.

6 Conclusion and Outlook

This paper presents the application of a new data-driven methods for modeling and predicting daily water flow and water level on the catchment scale for the Shannon river in Ireland. We have designed a deep convolutional network architecture to exploit the time-series nature of the data. Using several features captured at real across three various water stations, we have shown that the proposed convolutional network outperforms other four well-known time series forecasting models (ANNs, SVMs, WANNs and WSVMs). The inputs to the models consist of a combination of 30-years daily time series data sets (1983–2013), which can be divided between observed data sets (maximum temperature, minimum temperature, water level and water flow) and simulated data set, runoff. Based on the proposed deep convolutional network model, we further show the predictions of the water flow and water level for the lower-shannon water station from the duration of 2013–2080. Our proposed solution should be very useful for water authorities in the future allocation of water resources among competing users such as agriculture, demotic and power stations. In addition, it could formulate the basis of a decision support system by setting thresholds on water flow and water level predictions for the sake of creating accurate emergency alarms for capturing any expected abnormalities for the Shannon river.

References

1. Beven, K.J., et al.: Streamflow Generation Processes. IAHS Press, Wallingford (2006)
2. Chattopadhyay, P.B., Rangarajan, R.: Application of ann in sketching spatial non-linearity of unconfined aquifer in agricultural basin. *Agric. Water Manag.* **133**, 81–91 (2014)

3. Daniell, T.: Neural networks. Applications in hydrology and water resources engineering. In: National Conference Publication - Institute of Engineers. Australia (1991)
4. Deka, P.C., et al.: Support vector machine applications in the field of hydrology: a review. *Appl. Soft Comput.* **19**, 372–386 (2014)
5. Dutilleul, P.: An implementation of the algorithm à trous to compute the wavelet transform. In: Combes, J.M., Grossmann, A., Tchamitchian, P. (eds.) *Wavelets*. IPTI, pp. 298–304. Springer, Heidelberg (1989). https://doi.org/10.1007/978-3-642-75988-8_29
6. Gharbia, S.S., Alfatah, S.A., Gill, L., Johnston, P., Pilla, F.: Land use scenarios and projections simulation using an integrated GIS cellular automata algorithms. *Model. Earth Syst. Environ.* **2**(3), 151 (2016)
7. Gharbia, S.S., Gill, L., Johnston, P., Pilla, F.: Multi-GCM ensembles performance for climate projection on a GIS platform. *Model. Earth Syst. Environ.* **2**(2), 1–21 (2016)
8. Govindaraju, R.S., Rao, A.R.: *Artificial Neural Networks in Hydrology*, vol. 36. Springer Science & Business Media, Heidelberg (2013). <https://doi.org/10.1007/978-94-015-9341-0>
9. Hanke, M., Halchenko, Y.O., Sederberg, P.B., Hanson, S.J., Haxby, J.V., Pollmann, S.: PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics* **7**(1), 37–53 (2009)
10. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint [arXiv:1502.03167](https://arxiv.org/abs/1502.03167) (2015)
11. Karran, D.J., Morin, E., Adamowski, J.: Multi-step streamflow forecasting using data-driven non-linear methods in contrasting climate regimes. *J. Hydroinformatics* **16**(3), 671–689 (2014)
12. Kenabatho, P., Parida, B., Moalafhi, D., Segosebe, T.: Analysis of rainfall and large-scale predictors using a stochastic model and artificial neural network for hydrological applications in Southern Africa. *Hydrol. Sci. J.* **60**(11), 1943–1955 (2015)
13. Khan, M.S., Coulibaly, P.: Application of support vector machine in lake water level prediction. *J. Hydrol. Eng.* **11**(3), 199–205 (2006)
14. Kim, T.W., Valdés, J.B.: Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. *J. Hydrol. Eng.* **8**(6), 319–328 (2003)
15. Kingma, D., Ba, J.: ADAM: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
16. Kisi, O., Cimen, M.: A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *J. Hydrol.* **399**(1), 132–140 (2011)
17. Klinkenberg, R.: *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Chapman and Hall/CRC, Boca Raton (2013)
18. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. In: *The Handbook of Brain Theory and Neural Networks*, vol. 3361, No. 10 (1995)
19. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 807–814 (2010)
20. Sivakumar, B.: Forecasting monthly streamflow dynamics in the Western United States: a nonlinear dynamical approach. *Environ. Model. Softw.* **18**(8), 721–728 (2003)

21. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
22. Zheng, Y.: Methodologies for cross-domain data fusion: an overview. *IEEE Trans. Big Data* **1**(1), 16–34 (2015)