NUS BED 18-19

Interim Report

Rainfall Modelling Using Artificial Neural Network

ZHU MORAN

A0142451R

1. Background Introduction

In order to derive at a more accurate description of rainfall runoff processes in Singapore, a monitoring programme at pilot catchment (Kent Ridge Catchment at campus of National University of Singapore) was established to collect dense hydrological data over a representative area. The monitoring program collects hydrological data during entire year. The data are used to provide background information for development of an accurate hydrological distributed model for the pilot catchment.

2. Project objectives:

1. Use suitable artificial neural network architecture to produce 10, 20- and 60-minute forecast of flow rates at location Main Drain_04;
2. Present and discuss forecast accuracy as function of lead time;
3. Explain which neural network architecture did you use and why;
4. Discuss choice of training, cross-validation and testing data sets;
5. Present and discuss results using testing data set

3. Scope of project:

A set of data consisting of 41 events were collected during year 2011 and this project will use this set of data as reference. The data consists time series at resolution of 1 minutes of following quantities:

1. Catchment averaged rainfall intensities;
2. Time series of discharges at location Central Library, OPPRLink, Main_Drain_01, Main_Drain_02;
3. Time series of discharges at location Main_Drain_04

Detailed description of number of data sets for each event is included in Annex A.

4. Artificial Neural Network

Learning from the past to understand the present and predict the future has been the motivation for researchers in various academic fields. Many mathematical models were created and refined over the years to find the pattern or correlation of data sets. However, most of the modelling methods suffer from imprecision and uncertainty which lead to unsatisfactory prediction of results.

Artificial Neural Networks (ANN) become popular since the 1980s. ANNs have the ability to solve complicated problems like pattern recognition and nonlinear modelling. It could be a more accurate prediction tool if carefully trained.

Structure of ANNs

ANNs process information through nodes and links. The link connecting nodes in each layer has weights which represent the connection strength. Mathematically speaking, the weight acts as the coefficient of functions in calculation of the output values. The activation functions are often sigmoid functions. Sigmoid functions are easy to use due to its simple derivative equation. It could also model non-linear process.

Training

Training, also referred to as learning, is the process of obtaining reasonable weights and bias factors for the specific ANN using experiment data. Most of the hydrologic applications use supervised learning, a process where the external operator adjusts the weights and bias for nodes of ANN based on the difference between output data and actual data, Overfitting, or over-training refers to the situation where the ANN is only tuned to the training data sets. The operator should also understand when to terminate training.

Issues in Modelling

Although ANNs are powerful tools in modelling complex problems, it would still be beneficial for the users to consider several issues before training the ANNs. First of all, selecting appropriate variables is crucial in modelling ANNs. There could be many different variables available such as temperature, water level, evaporation rate, tidal flows and so on. Instead of blindly including all related variables and create a large network, the user could study the problem and find the most relevant variables. Not only to produce a succinct network, but also saves time for training by reducing the number of nodes and links. At the same time, the output variables should also be carefully checked to find out probable problems with the existing ANN.

Also, the pre-processing of data plays a significant role in creating a good ANN. The reliability of relevant data should be checked before inputting into ANN, especially if the data is obtained remotely or by external parties. Sieving out anomaly could greatly improve the speed of training. In common practices, the first-handed data needs to be normalised or de-trended before feeding to ANN. The suitable and reasonable method to pre-process data is important. Inappropriate ways to transform data could result in unnecessary difficulties during the training step.

Choice of ANN structure

Time-Delay Neural Network (TDNN) would be the most appropriate network structure. Since the data obtained is in time sequence at resolution of 1 minute, TDNN is able to recognize this time difference and select a certain window of events for training. Unlike traditional Multi-layer network (MLP) which considers all inputs at the same time, TDNN will not destroy the time series signal.

5. Project tools:

NeuroSolutions 7 and Microsoft Excel will be used for this project. NeuroSolutions 7 is an easy-to-use and powerful software to build, train and test Artificial Neural Networks.

6. Training Attempts

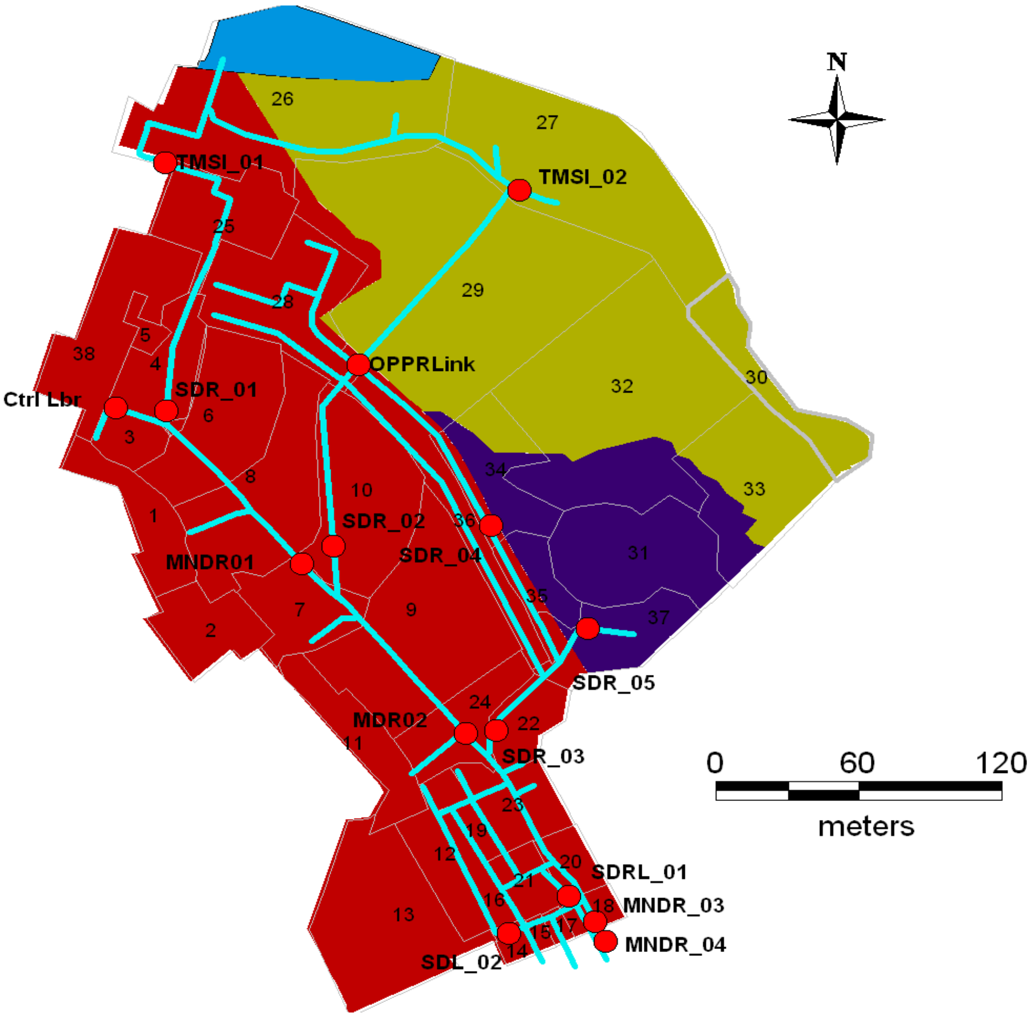


Fig 6.1 Map of drains

As it can be seen from Fig 6.1, Main_Drain_04 has the largest stream because all rainwater collected by other drains will flow downstream into Main_Drain_04. Therefore, all other data need to be considered as inputs in building the network.

The five inputs are:

1. Rainfall (unit:mm)
2. Q_Main_Drain_01 (unit: l/s)
3. Q_Main_Drain_02 (unit: l/s)
4. Q_CentralLibrary (unit: l/s)
5. Q_OppResearchLink (unit: l/s)

The desired output is

1. Q_MainDrain_04

6.1 Straight-forward approach

The first attempt is training a TDNN using the original data given for lead time of 5 and 10 minutes. The aim is to getting familiar with the software and looking for areas of improvement at the same time.

6.1.1 Lead time 5 minutes

Training and testing procedure:

1. Shift Q_MD04 up by 5 rows.
2. Delete last 5 rows of data from each event.
3. Identify five input columns and one desired column
4. Choosing training, cross-validation and testing data sets.

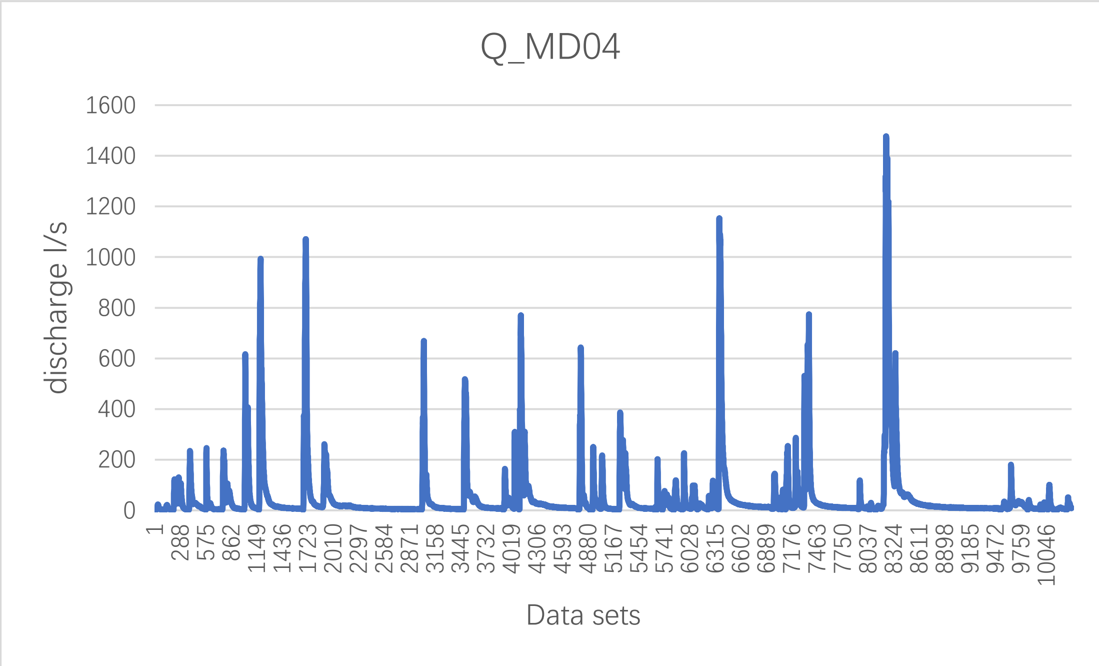


Fig 6.1.1 flow distribution of MD04

As it can be seen in Fig 6.1.1, the most extreme case occurs between No.8037 and No.8611 data sets. According to the principles of ANN training, the training data should include the most extreme case. Therefore, in this project, last 60% of data sets are used for training, the first 20% of data sets are used for testing and the rest 20% is used for cross-validation.

5. Build and train a regression TDNN using NeuroSolutions 7.

6. Results and discussion

Since the lead time is relatively short, the TDNN shows satisfactory results.

Best Networks	Training	Cross Validation
Epoch #	3601	1601
Minimum MSE	0.000407042	0.000183124
Final MSE	0.000407042	0.000228288

Table 6.1.1.1 Straight-Forward Approach Training Results (Lead time 5 minutes)

The training and cross-validation Mean Squared Error (MSE) are all small.

Performance	5 mins Q_MD04
RMSE	31.21571416
NRMSE	0.021161087
MAE	14.9497133
NMAE	0.010134389
Min Abs Error	0.001121289
Max Abs Error	312.17984
r	0.975496217
Score	96.73571975

Table 6.1.1.2 Straight-Forward Approach Testing Results (Lead time 5 minutes)

Using the first 2000 data sets for testing, the R value is 0.9755 and the test score is 96.74.

Detailed training and testing report could be found in the appendix.

6.1.2 Lead time 10 minutes

Training and testing procedure:

1. Shift Q_MD04 up by 10 rows.
2. Delete last 10 rows of data from each event.
3. Build and train TDNN.
4. Results and discussion

   Since the lead time is still relatively short, the TDNN also shows quite satisfactory results.

Best Networks	Training	Cross Validation
Epoch #	6907	4908
Minimum MSE	0.002025743	0.001362867
Final MSE	0.002025744	0.001409998

Table 6.1.2.1 Straight-Forward Approach Training Results (Lead time 10 minutes)

The training and cross-validation Mean Squared Error (MSE) are all small.

Performance	10 min Q_MD04
RMSE	66.90674329
NRMSE	0.045358813
MAE	27.43848831
NMAE	0.018601671
Min Abs Error	0.000416696
Max Abs Error	641.6971404
r	0.87332119
Score	90.7375785

Table 6.1.2.2 Straight-Forward Approach Testing Results (Lead time 10 minutes)

Using the first 2000 data sets for testing, the R value is 0.8733 and the test score is 90.74

Detailed training and testing report could be found in the appendix.

6.2 Taking the difference of flow

Although the straight-forward approach could give good results for short lead time, it can be predicted that the accuracy of forecast would decrease by a large extent for long lead time. Measures should be taken to increase the accuracy for long time prediction.

Since the rainfall amount for each minute is given, the difference in water flow in the drains will be more correlated to the rainfall. Therefore, instead of predicting the flow in MD04 directly, now the aim is to predict the difference in flow.

$$Q(t+1) = Qt + \widehat{dQ}$$

Predict dQ for lead time 5 minutes and 10 minutes.

6.2.1 Lead time 5 minutes:

Training and testing procedure:

1. Compute the differences of flow for 5 minutes for 4 drains, namely MD01, MD02, CentralLibrary and OPPRLink.
2. Compute the differences of flow for 5 minutes for MD04.
3. Delete last 5 rows of data for each event
4. Build and train TDNN
5. Results and discussion

   The results are also satisfactory.

Best Networks	Training	Cross Validation
Epoch #	3471	1471
Minimum MSE	6.99225E-05	4.8834E-05
Final MSE	6.99225E-05	5.42509E-05

Table 6.2.1.1 dQ Approach Training Results (Lead time 5 minutes)

The training and cross-validation Mean Squared Error (MSE) are all small.

Performance	desired
RMSE	20.94367043
NRMSE	0.01420505
MAE	6.188470897
NMAE	0.004197332
Min Abs Error	3.92309E-05
Max Abs Error	480.992967
r	0.942325741
Score	95.4559074

Table 6.2.1.2 dQ Approach Testing Results (Lead time 5 minutes)

Using the test data sets for testing, the R value is 0.9423 and the test score is 95.46

Detailed training and testing report could be found in the appendix.

6.2.2 Lead time 10 minutes:

Training and testing procedure:

1. Compute the differences of flow for 10 minutes for 4 drains, namely MD01, MD02, CentralLibrary and OPPRLink.
2. Compute the differences of flow for 10 minutes for MD04.
3. Delete last 10 rows of data for each event
4. Build and train TDNN
5. Results and discussion

   The results show improvement comparing with the straight-forward approach.

Best Networks	Training	Cross Validation
Epoch #	5479	4322
Minimum MSE	6.55272E-05	8.71113E-05
Final MSE	6.55276E-05	8.74086E-05

Table 6.2.2.1 dQ Approach Training Results (Lead time 10 minutes)

The training and cross-validation Mean Squared Error (MSE) are all smaller comparing with previous method.

Performance	Q_MD04
RMSE	23.57424181
NRMSE	0.011975605
MAE	9.152484964
NMAE	0.00464942
Min Abs Error	0.000100591
Max Abs Error	224.5343607
r	0.975070177
Score	97.21656249

Table 6.2.2.2 dQ Approach Testing Results (Lead time 10 minutes)

Using the test data sets for testing, the R value is 0.9751 and the test score is 97.22

Detailed training and testing report could be found in the appendix.

6.3 Moving Average

The concept of moving average refers to the calculation step of taking average of several data sets. Moving average could help to smooth the data sets and filter out "noise" in the data sets. In this project, there was no information about abnormal data therefore using moving average could help to define a better trend of flow. The method used here is Simple Moving Average (SMA), which is the calculation of averaging a certain number of data sets.

6.3.1 Lead time 5 minutes

Training and testing procedure:

1. Compute moving average of flow for 5 minutes for 4 drains, namely MD01, MD02, CentralLibrary and OPPRLink.
2. Compute difference of flow for 5 minutes for 4 drains, namely MD01, MD02, CentralLibrary and OPPRLink.

   Note: the lagging of moving average is eliminated in this step
3. Compute the differences of flow for 5 minutes for MD04.
4. Delete last 5 rows of data for each event
5. Build and train TDNN
6. Results and discussion

Best Networks	Training	Cross Validation
Epoch #	2547	548
Minimum MSE	8.94239E-05	5.2564E-05
Final MSE	8.9426E-05	5.36957E-05

Table 6.3.1.1 SMA dQ Approach Training Results (Lead time 5 minutes)

The training and cross-validation Mean Squared Error (MSE) are all small.

Performance	Q_MD04
RMSE	14.78095853
NRMSE	0.012179199
MAE	6.28020593
NMAE	0.005174758
Min Abs Error	0.000321534
Max Abs Error	193.4613219
r	0.967082275
Score	96.80130163

Table 6.3.1.2 SMA dQ Approach Testing Results (Lead time 5 minutes)

Using the test data sets for testing, the R value is 0.9671 and the test score is 96.80

Detailed training and testing report could be found in the appendix.

6.3.2 Lead time 10 minutes

Training and testing procedure:

1. Compute moving average of flow for 10 minutes for 4 drains, namely MD01, MD02, CentralLibrary and OPPRLink.

2. Compute difference of flow for 10 minutes for 4 drains, namely MD01, MD02, CentralLibrary and OPPRLink.

Note: the lagging of moving average is eliminated in this step

3. Compute the differences of flow for 10 minutes for MD04.

4. Delete last 10 rows of data for each event

5. Build and train TDNN

6. Results and discussion

Best Networks	Training	Cross Validation
Epoch #	4113	2113
Minimum MSE	6.70739E-05	5.45222E-05
Final MSE	6.70739E-05	6.20011E-05

Table 6.3.2.1 SMA dQ Approach Training Results (Lead time 10 minutes)

The training and cross-validation Mean Squared Error (MSE) are all small.

Performance	Q_MD04
RMSE	16.18299636
NRMSE	0.010726951
MAE	6.500686628
NMAE	0.004309001
Min Abs Error	0.001506964
Max Abs Error	166.6028272
r	0.98400815
Score	97.742609

Table 6.3.2.2 SMA dQ Approach Testing Results (Lead time 10 minutes)

Using the test data sets for testing, the R value is 0.9840 and the test score is 97.74.

The testing results improved comparing with the previous method.

Detailed training and testing report could be found in the appendix.

7. Summary and future works

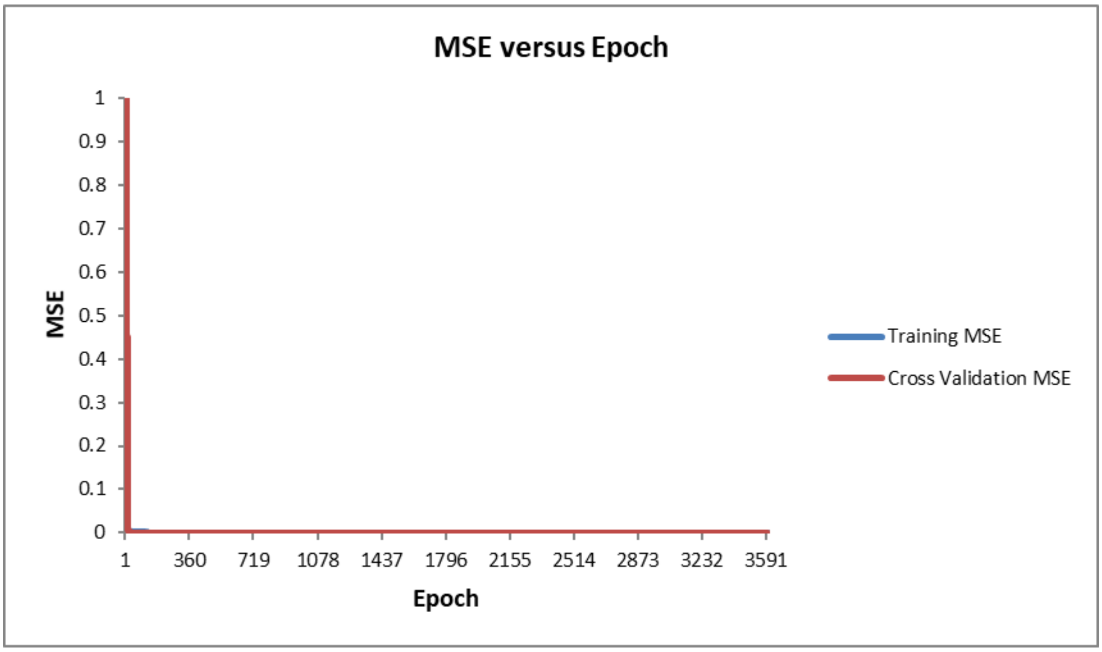
Based on the previous experiment attempts, it could be concluded that the TDNN could have excellent results for short lead time. However, the quality of prediction will decrease as lead time increases. Also, predicting the difference of flow could have better results comparing with directly predicting the actual magnitude of flow. Introducing moving average could filter out potential “noise” in the data sets, create a smoother trend of data hence getting better prediction results.

For the next part of the project, the focus should be getting good prediction results for longer lead time (20 and 60 minutes). Futhermore, other appropriate data processing method should be adopted to improve the prediction accuracy.

Appendix A: Number of data sets of events

Event number	Number of data sets
1	99
2	82
7	174
8	217
12	183
13	251
14	162
15	501
16	1313
17	483
18	451
19	129
20	59
21	662
22	7
23	125
24	315
25	416
26	24
27	62
28	125
29	94
30	175
31	115
32	120
33	604
34	79
35	36
36	43
37	816
38	128
39	61
40	1403
61	87
62	230
63	135
64	38
65	58
66	56
67	141
75	71

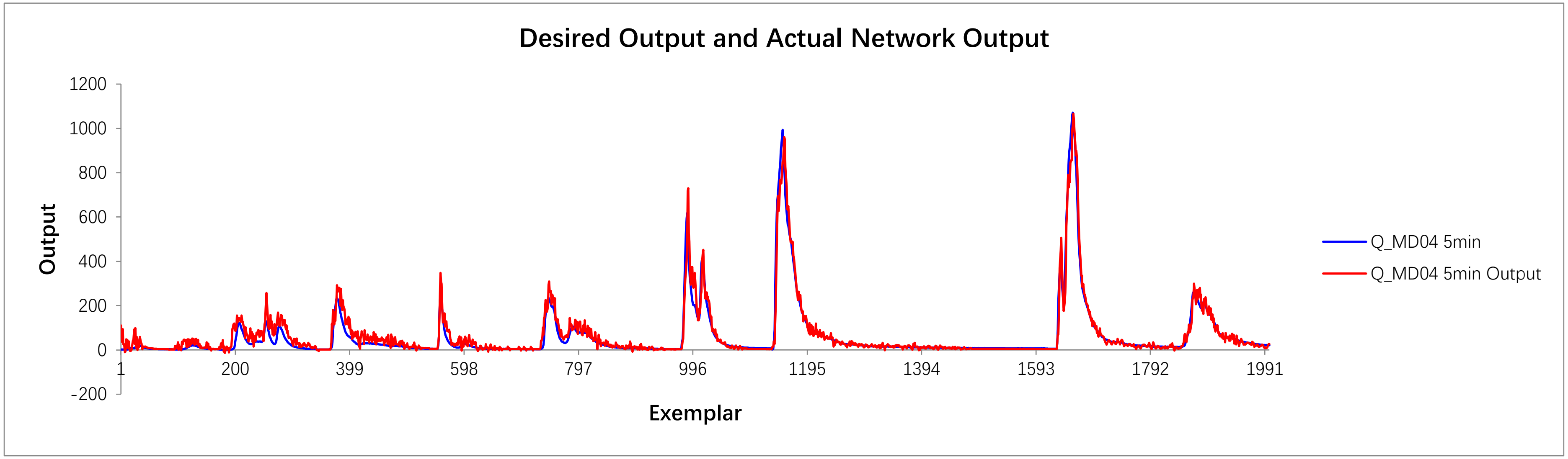
Appendix B: Training and testing results



Straight-Forward Approach Training MSE versus Epoch Graph (Lead time 5 minutes)

Best Networks	Training	Cross Validation
Epoch #	3601	1601
Minimum MSE	0.000407042	0.000183124
Final MSE	0.000407042	0.000228288

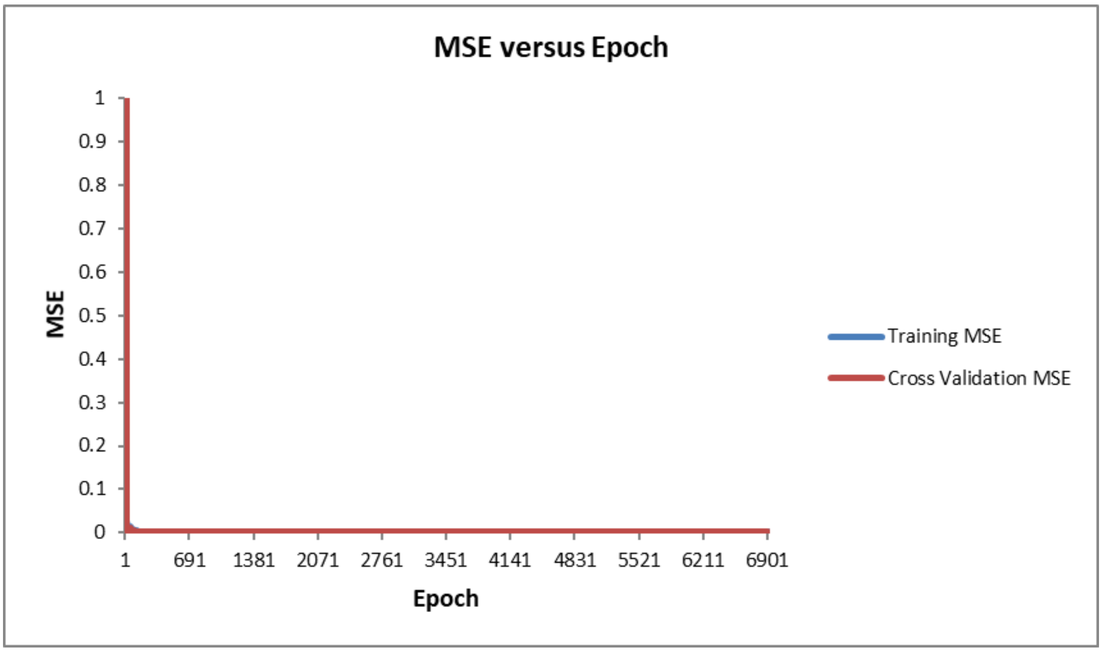
Straight-Forward Approach Training results (Lead time 5 minutes)



Straight-Forward Approach Testing Output Comparison (Lead time 5 minutes)

Performance	Q_MD04 5min
RMSE	32.11084633
NRMSE	0.021767896
MAE	16.96863799
NMAE	0.011503015
Min Abs Error	0.00483469
Max Abs Error	252.2893476
r	0.974397408
Score	96.6448296

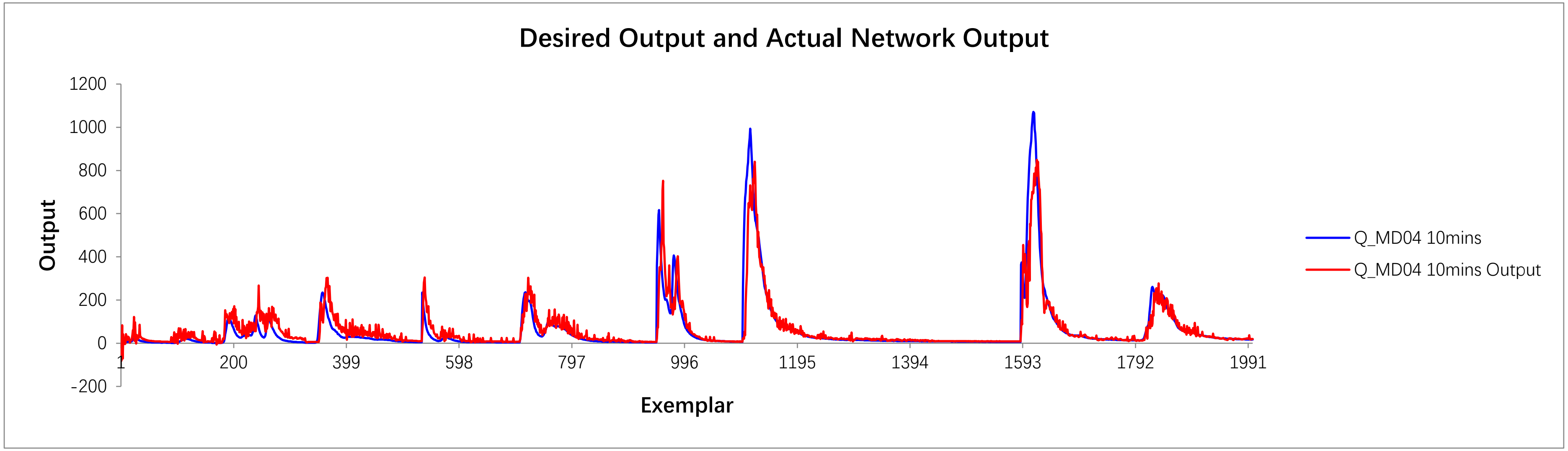
Straight-Forward Approach Testing Results (Lead time 5 minutes)



Straight-Forward Approach Training MSE versus Epoch Graph
(Lead time 10 minutes)

Best Networks	Training	Cross Validation
Epoch #	6907	4908
Minimum MSE	0.002025743	0.001362867
Final MSE	0.002025744	0.001409998

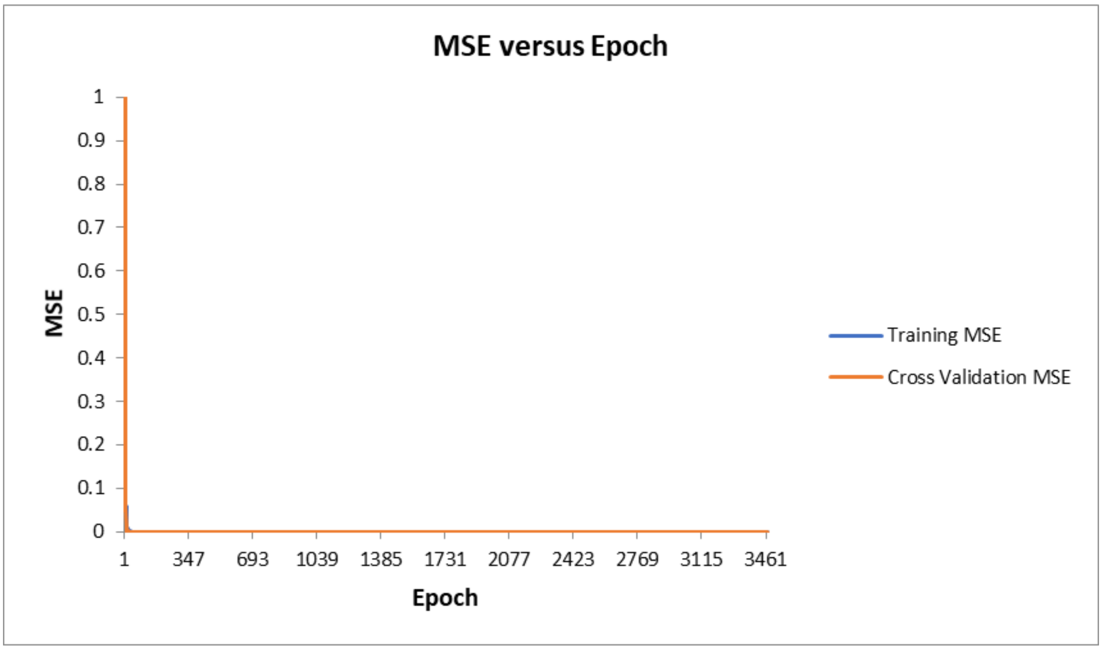
Straight-Forward Approach Training results (Lead time 10 minutes)



Straight-Forward Approach Testing Output Comparison (Lead time 10 minutes)

Performance	Q_MD04 10mins
RMSE	66.68490527
NRMSE	0.04520842
MAE	26.88503983
NMAE	0.018226466
Min Abs Error	0.013048231
Max Abs Error	636.6691346
r	0.874578946
Score	90.80735097

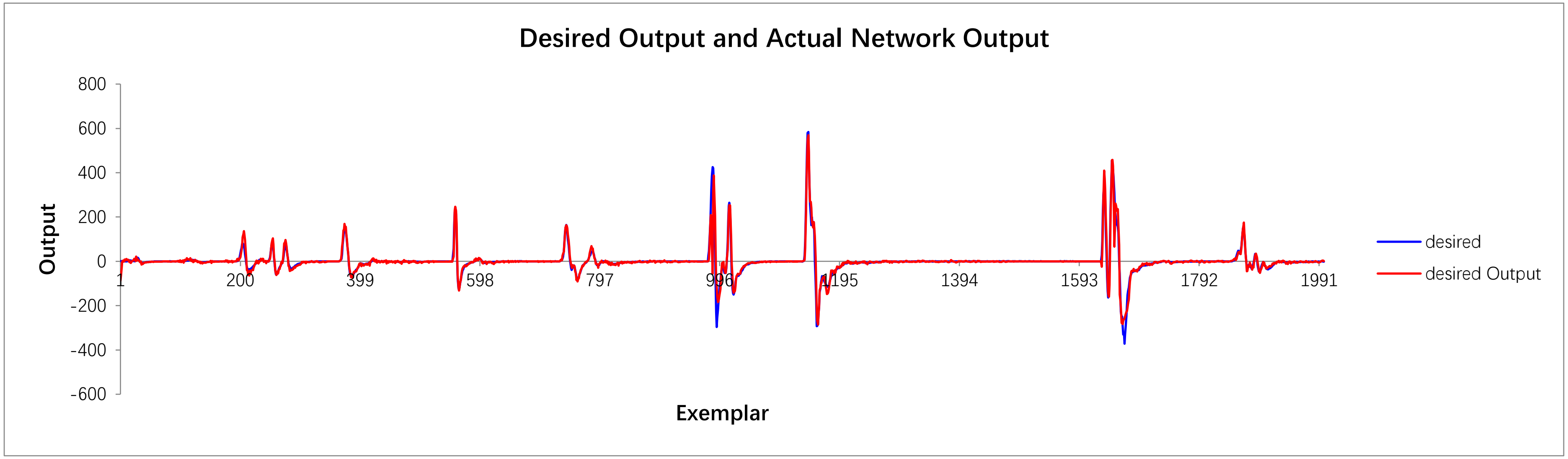
Straight-Forward Approach Testing Results (Lead time 10 minutes)



dQ Method Training MSE versus Epoch Graph (Lead time 5 minutes)

Best Networks	Training	Cross Validation
Epoch #	3471	1471
Minimum MSE	6.99225E-05	4.8834E-05
Final MSE	6.99225E-05	5.42509E-05

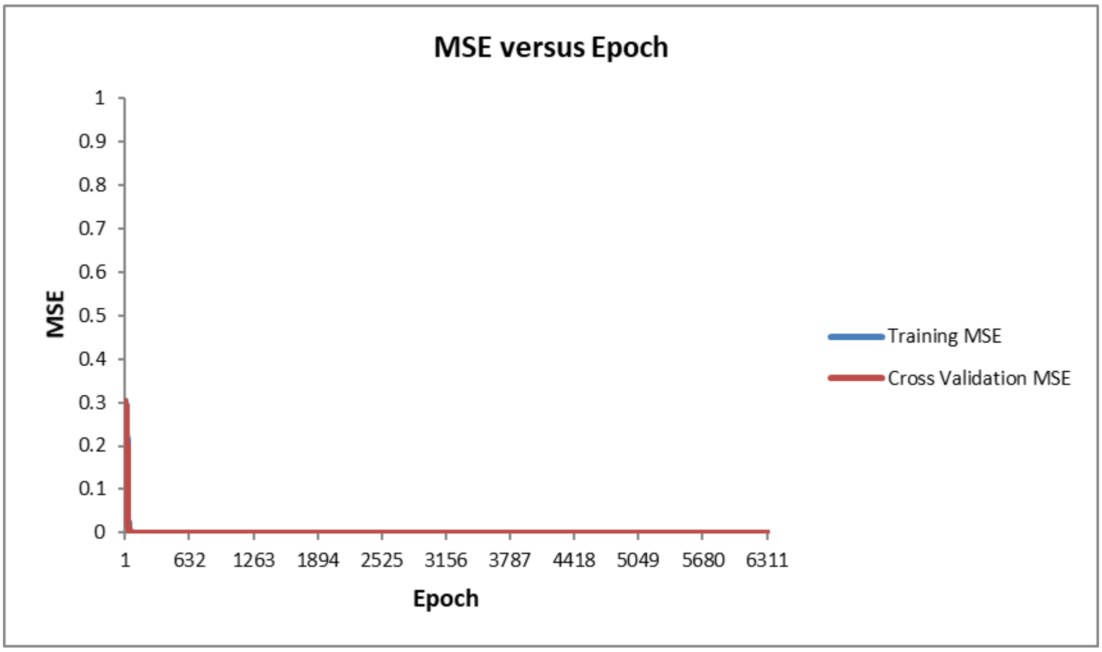
dQ Method Training results (Lead time 5 minutes)



dQ Method Testing Output Comparison (Lead time 5 minutes)

Performance	dQ 5mins
RMSE	20.94367043
NRMSE	0.01420505
MAE	6.188470897
NMAE	0.004197332
Min Abs Error	3.92309E-05
Max Abs Error	480.992967
r	0.942325741
Score	95.4559074

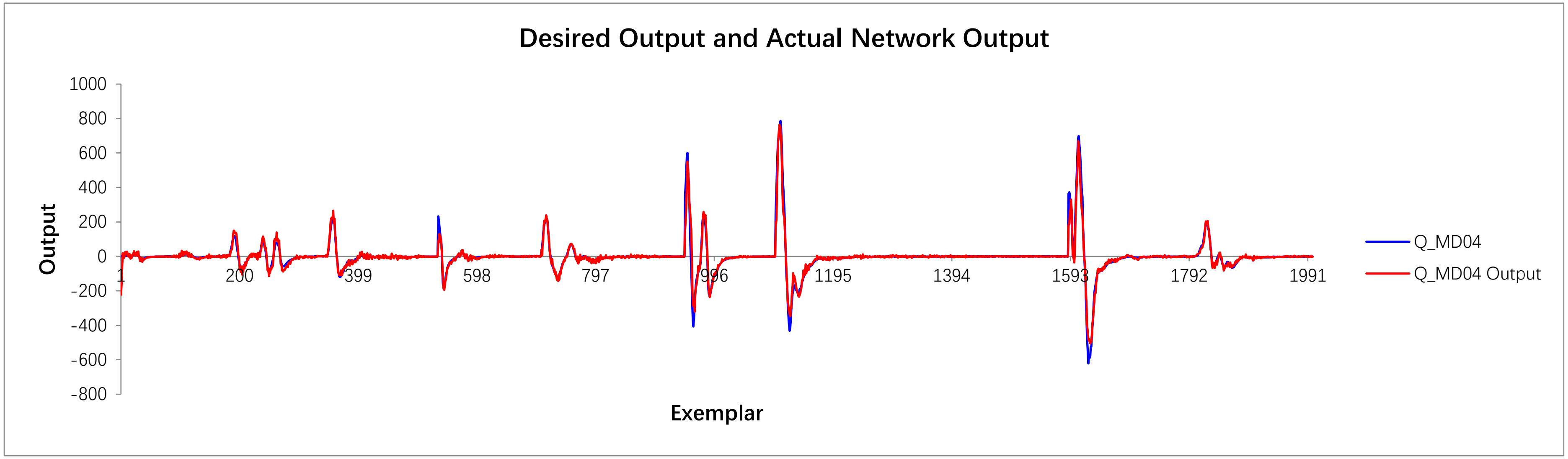
dQ Method Testing Results (Lead time 5 minutes)



dQ Method Training MSE versus Epoch Graph (Lead time 10 minutes)

Best Networks	Training	Cross Validation
Epoch #	5479	4322
Minimum MSE	6.55272E-05	8.71113E-05
Final MSE	6.55276E-05	8.74086E-05

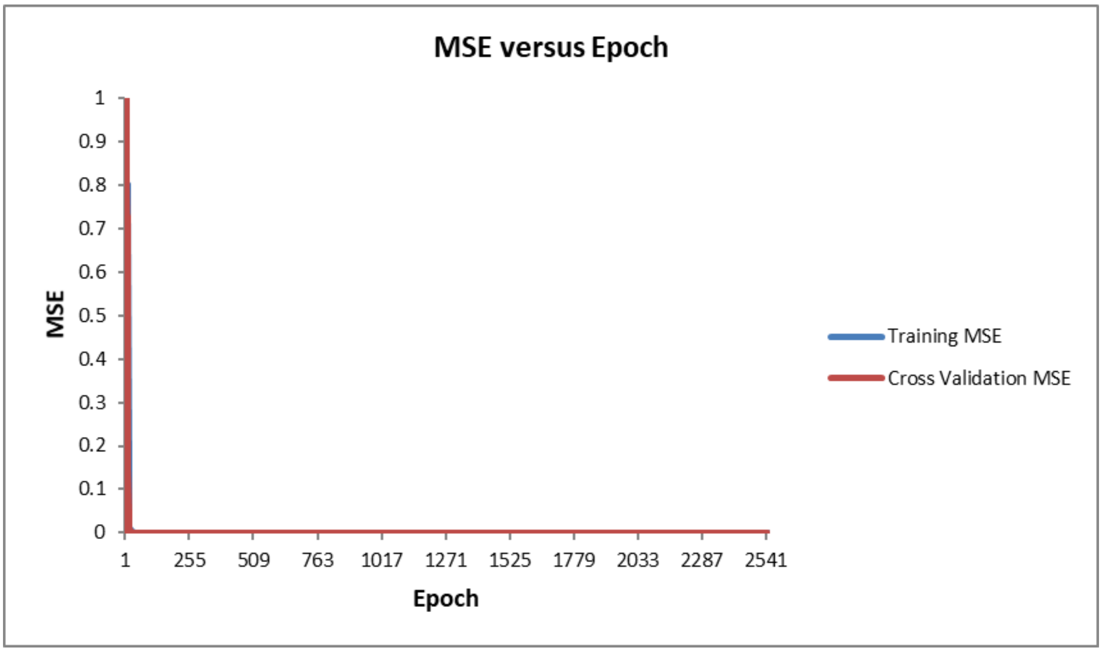
dQ Method Training results (Lead time 10 minutes)



dQ Method Testing Output Comparison (Lead time 10 minutes)

Performance	Q_MD04
RMSE	23.57424181
NRMSE	0.011975605
MAE	9.152484964
NMAE	0.00464942
Min Abs Error	0.000100591
Max Abs Error	224.5343607
r	0.975070177
Score	97.21656249

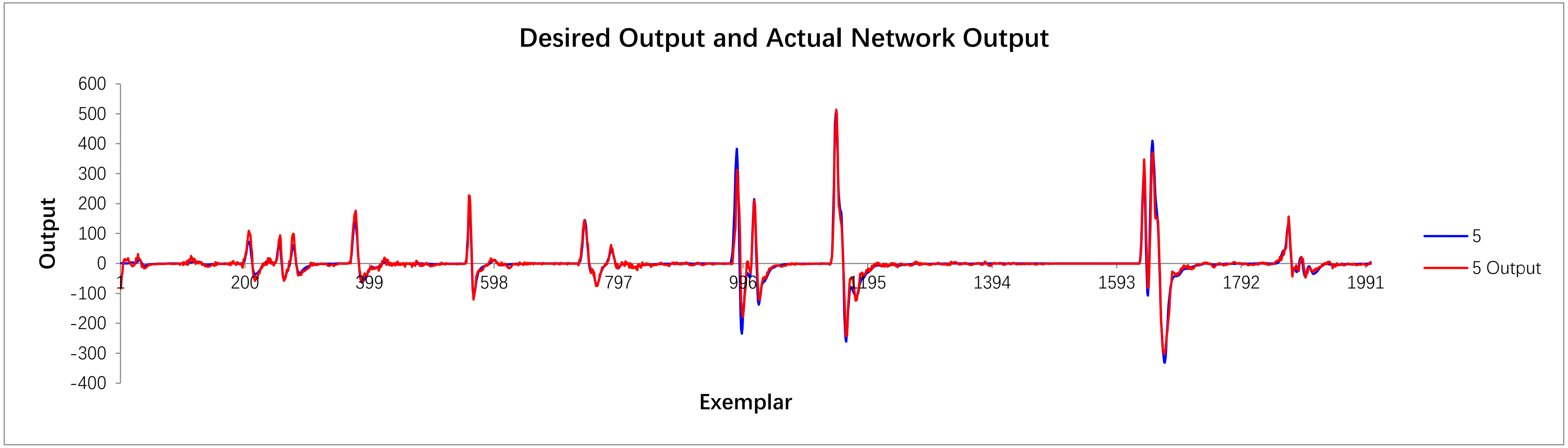
dQ Method Testing Results (Lead time 10 minutes)



SMA dQ Method Training MSE versus Epoch Graph (Lead time 5 minutes)

Best Networks	Training	Cross Validation
Epoch #	2547	548
Minimum MSE	8.94239E-05	5.2564E-05
Final MSE	8.9426E-05	5.36957E-05

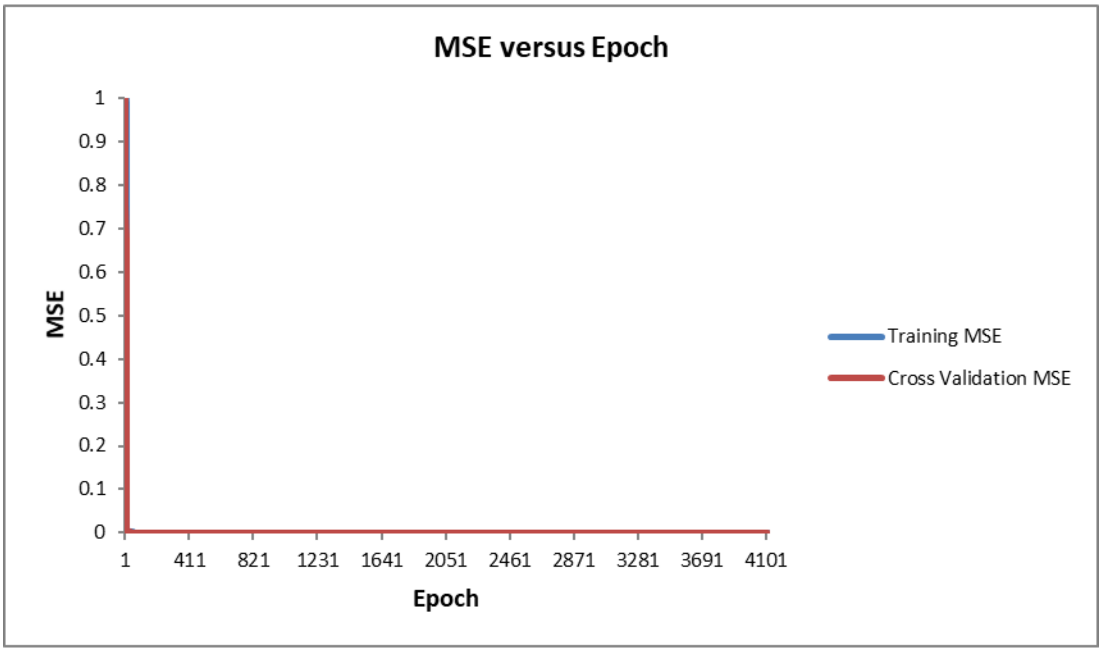
SMA dQ Method Training results (Lead time 5 minutes)



SMA dQ Method Testing Output Comparison (Lead time 5 minutes)

Performance	Q_MD04
RMSE	14.78095853
NRMSE	0.012179199
MAE	6.28020593
NMAE	0.005174758
Min Abs Error	0.000321534
Max Abs Error	193.4613219
r	0.967082275
Score	96.80130163

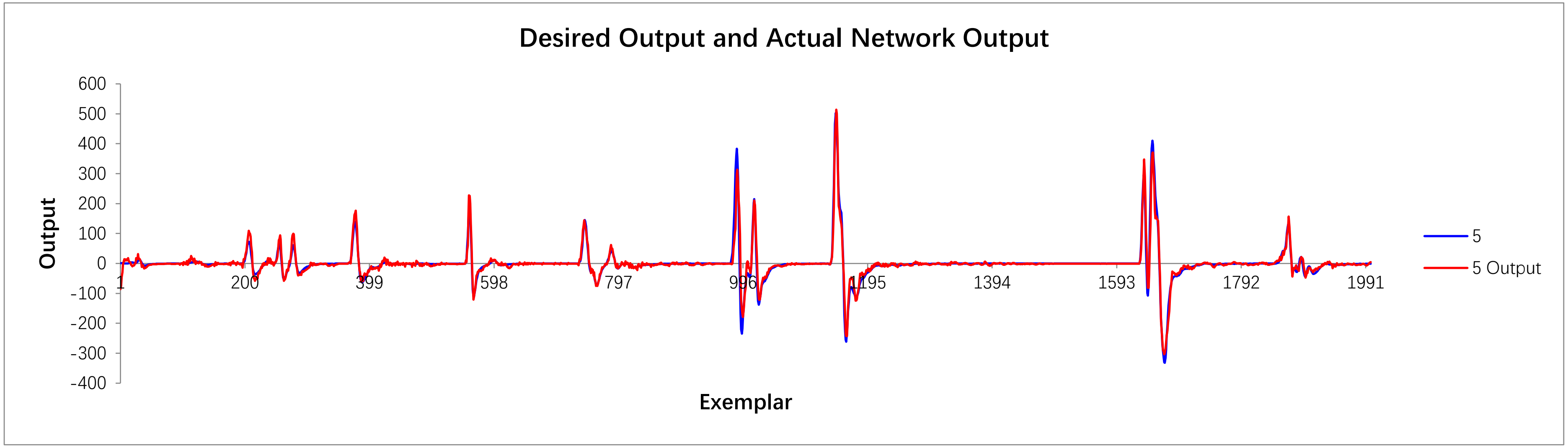
SMA dQ Method Testing Results (Lead time 5 minutes)



SMA dQ Method Training MSE versus Epoch Graph (Lead time 10 minutes)

Best Networks	Training	Cross Validation
Epoch #	2547	548
Minimum MSE	8.94239E-05	5.2564E-05
Final MSE	8.9426E-05	5.36957E-05

SMA dQ Method Training results (Lead time 10 minutes)



SMA dQ Method Testing Output Comparison (Lead time 10 minutes)

Performance	Q_MD04
RMSE	14.78095853
NRMSE	0.012179199
MAE	6.28020593
NMAE	0.005174758
Min Abs Error	0.000321534
Max Abs Error	193.4613219
r	0.967082275
Score	96.80130163

SMA dQ Method Testing Results (Lead time 10 minutes)

References

Artificial Neural Networks in Hydrology. I: Preliminary Concepts. (2000). Journal of Hydrologic Engineering, 5(2), 115-123. doi:10.1061/(asce)1084-0699(2000)5:2(115)

Babovic, V., Cañizares, R., Jensen, H. R., & Klinting, A. (2001). Neural Networks as Routine for Error Updating of Numerical Models. Journal of Hydraulic Engineering, 127(3), 181-193. doi:10.1061/(asce)0733-9429(2001)127:3(181)

Benmahdjoub, K., Ameur, Z., & Boulifa, M. (2013). Forecasting of Rainfall Using Time Delay Neural Network in Tizi-Ouzou (Algeria). Energy Procedia, 36, 1138-1146. doi:10.1016/j.egypro.2013.07.129

Ghumman, A., Ghazaw, Y. M., Sohail, A., & Watanabe, K. (2011). Runoff forecasting by artificial neural network and conventional model. Alexandria Engineering Journal, 50(4), 345-350. doi:10.1016/j.aej.2012.01.005

Govindaraju, R. S., & Rao, A. R. (2011). Artificial neural networks in hydrology. Dordrecht: Springer.

Herawati, H., Suripin, & Suharyanto. (2017). River flow modeling using artificial neural networks in Kapuas river, West Kalimantan, Indonesia. doi:10.1063/1.5011620

Shanmuganathan, S. (2016). Artificial Neural Network Modelling: An Introduction. Artificial Neural Network Modelling Studies in Computational Intelligence, 1-14. doi:10.1007/978-3-319-28495-8_1