Error Forecasting Using Linear Regression Model

Lian Guey Ler* / Byung Sik Kim** / Gye Woon Choi*** /

Byung Hwa, Kang****+ / Jung Jae Kwang*****

Abstract : In this study, Mike11 will be used as the numerical model where a data assimilation method will be applied to it. This paper aims to gain an insight and understanding of data assimilation in flood forecasting models. It will start with a general discussion of data assimilation, followed by a description of the methodology and discussion of the statistical error forecast model used, which in this case is the linear regression. This error forecast model is applied to the water level forecast simulated by MIKE11 to produced improved forecast and validated against real measurements. It is found that there exists a phase error in the improved forecasts. Hence, 2 general formula are used to account for this phase error and they have shown improvement to the accuracy of the forecasts, where one improved the immediate forecast of up to 5 hours while the other improved the estimation of the peak discharge.

Keywords : Error Forecasting, Mike11, Linear Regression, WEKA

1. Introduction

Data assimilation, a novel and versatile methodology, has been employed heavily in the numerical models lately. In the past, models are physically-based numerical models and it is only in the late 1980s where scientists and engineers begin to consider data mining as a possible approach for modeling. And with the increasing advancement of the computer processing power and accessibility of observed/ measured data, data mining appears to be a good replacement for the complicated dynamics equations which were unable to fully describe the real environment accurately as it is much more robust and cheaper its counterpart.

Hence, data assimilation is developed where it utilizes the advantages offered by both the numerical model and data mining. Whenever there is observed data available, they will be implemented into the numerical model to generate more accurate forecasts, with the uncertainty of the data and model taken into account. Arising from the application in geosciences, perhaps most importantly in weather forecasting and hydrology, data assimilation has also been extended to the ocean modeling where it is used to estimate and predict the waves' height, biochemical parameters and sediments.

+ Corresponding author : ka12222@korea.kr

* Researcher · International Center for Urban Water Hydroinformatics Research & Innovation, E-mail : lianguey@gmail.com

** Reaserch Fellow · Korea Institute of Construction Technology · E-mail : hydrokbs@kict.re.kr

*** Professor · Inchen University, E-mail : gyewoon@incheon.ac.kr

**** Bureau director, Bureau of Disaster Prevention and Management, National Emergency Management Agency, E-mail : ka12222@korea.kr - Corresponding Author

***** Ph.D. Candidate, Green & Clean Engineering, CEO, E-mail : kill0713@hanmail.net

Nowadays data assimilation methods are applied to models to produce better forecasts. However, the updated correction of the model initial conditions will eventually be "washed out" in time, where it will give results similar to that of the initially uncorrected model. Nonetheless, it is possible to forecast these measurements (model initial conditions), either directly or in the form of the model errors. By using this methodology, the model will be able to give an extended forecast (longer than the "wash-out" period) with much more accuracy, depending on the robust of the error forecast model used.

2. Model Description

EXCEL is used together with VBA scripts for most of the data processing, while WEKA will be used for training of linear regression models. As for the hydrodynamic simulations, Mike11 is used.

2.1 WEKA

Waikato Environment for Knowledge Analysis (WEKA), an open source projects in machine learning, is a comprehensive collection of machine learning algorithms for data mining tasks and is written in JAVA. It has a set of tools for data pre-processing, classification, regression, clustering, association rules and visualization. Since it is licensed under the GNU General Public License (GNU GPL or GPL), it comes as a free program. In this study, WEKA is used for a small part in data pre-processing, and largely in training regression models and visualization. For detailed implementation and user manual, please refer to http://www.cs.waikato.ac.nz/ml/weka/ under the documentation section.

2.2 MIKE11 model

MIKE11 is a 1D engineering tool for modelling conditions in rivers, irrigation systems, lakes/reservoirs and other inland waters. It is also designed and extended for flood risk analysis and mapping, design of flood alleviation systems, real-time forecasting, hydraulic analysis/design of structures like bridges, dam break analysis and water quality issues.

The two governing equations are the continuity equation and the Saint Venant equation.

$$\frac{\delta Q}{\delta t}+\frac{\delta A}{\delta t}=q \tag{1a}$$

$$\frac{\delta Q}{\delta t}+\frac{\delta\left(\alpha\frac{Q^2}{A}\right)}{\delta x}+gA\frac{\delta h}{\delta x}+\frac{g.|Q|.Q}{C^2.A.R}=q \tag{1b}$$

where Q = discharge (m^3/s) , A = cross sectional area (m^2) , q = source / sink term ($m^3/(m.s)$), h = stage above datum (m), C = Chezy resistance coefficient ($m^{0.5}/s$), R = hydraulic or resistance radius (m), a = momentum distribution coefficient.

In the study, the NAM module is used to calculate the rainfall runoff of the catchments. NAM is a tool developed by the Technical University of Denmark to simulate the rainfall-runoff process in a catchment. It is a lumped, conceptual rainfall-runoff model, simulating the overland-, inter- flow, and base-flow components as a function of the moisture contents in four storages.

3. Data Assimilation

Despite the great advancement of the forecast models like Mike and ISIS, there still remain a number of model errors where some can be improved or corrected with the introduction of measure data to update the forecasts produce from the models. As the forecast parameters (eg, water level) are often heavily depended on the initial and boundary conditions; by constantly updating the nowcast, better and more accurate forecast can be obtained.

Data assimilation procedures can be classified into four main categories (Figure 1), through the updating of model's input, states variables, parameters and output.

Updating of the model inputs is justified on the assumption that the dominant error comes from the inputs.

As for the updating of the model state variables, the main advantage stems from the fact that not only does it account for the estimation of the model state; it also gives an estimation of the model prediction uncertainty conditioned on the available measurements. However, it requires a lot of computation power for the uncertainty propagation, which makes it unfeasible for real-time systems. However, recent development has produce more computation-cost-effective methods like the ensemble Kalman filter (EnKF) and reduce rank root filter (RRSQRT).

The updating of output variables, also called error forecasting, uses a model to fit the model errors which is then used to forecast the model errors in the future. This forecast error is then added to the model prediction itself to produce an updated forecast. This method has been employed in hydrological forecasting systems like rainfall-runoff forecasting and use of ANN for error updating of numerical models.

4. Methodology

The first step to building an external error forecasting model is to decide on the approach the model will take. In this study, the Linear Regression method is being investigated.

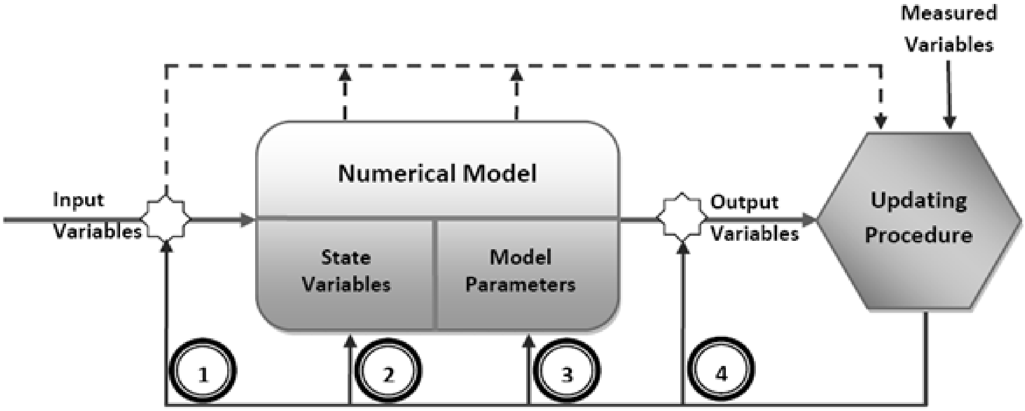


Figure 1. Schematic diagram of data assimilation with updating at (1) input variables, (5) state variables, (3) model parameters, (4) output variables (adapted from Refsgaard, 1997)

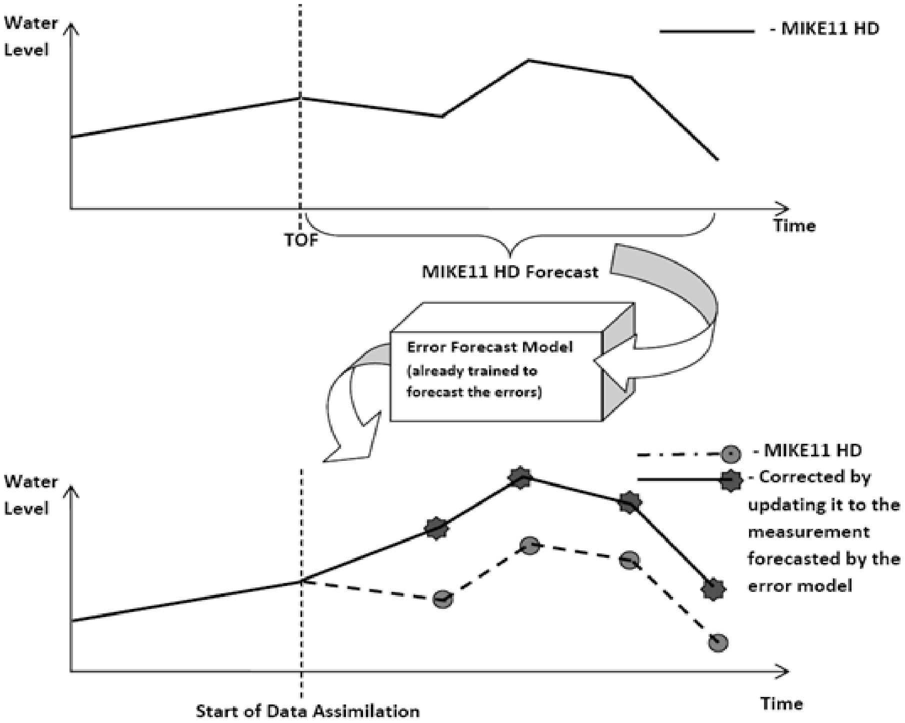

Water Level
- MIKE11 HD
TOF
Time
MIKE11 HD Forecast
Error Forecast Model (already trained to forecast the errors)
- MIKE11 HD
- Corrected by updating it to the measurement forecasted by the error model
Water Level
Start of Data Assimilation
Time

Figure 2. Diagram of the offline error correction process in MIKE11 data assimilation module

In order to build the error forecast model, the uncorrected calibrated Mike11 model must first be simulated into the future in order to obtain a time series for all the measurement locations. This time series will then be processed and fed into the trained error predicting models to produce the error forecasts. These forecasted errors can then be added to the uncorrected model output to obtain the state variables (the water level in this case).

4.1 Statistics

In this study, the root-mean-squared error (RMSE) will be used as the primary indicator for goodness of fit to the measurement, where it measures the differences between forecast values (estimates) and the measured data (observed data). It is defined by:

$$\mathrm{RSME} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \qquad (2)$$

where, N = length of the time series, y_i = observed values, y_i = estimates

4.2 Accounting for Errors

The calculation of the error at Time (T) =0 is the difference between the measured data and the uncorrected forecast from mike11 simulation of the same location point at T=0.

When calculating the errors for the historical period, it is the difference between the measurement and the uncorrected forecast from mike11 simulation at the same time frame of interest and same location as shown in the table 1 below.

Table 1. Table of error calculation

Measurement*	Mike11^	E(T=0)	E(T=0)	E(T=-1)	E(T=-2)
X1	Y1	X1 -Y1 = E1	E1	-	-
X2	Y2	X2 -Y2 = E2	E2	E1	-
X3	Y3	X3 -Y3 = E3	E3	E2	E1
X4	Y4	X4 -Y4 = E4	E4	E3	E2
X5	Y5	X5 -Y5 = E5	E5	E4	E3

* where measurement is the measured data ^ where mike11 is the uncorrected forecast by mike11

5. Case Study

The area of interest is the Po River basin is located in the North-west of Italy, at the Piemonte region. The Piemonte region has an area of 25,399 km^2 and a population of about 4.4 million. 73% of its area consists of alpine with surrounded by mountain chains with Ligurian Appenies in the south and the Alps in the north. Due to the extensive area of the Po River basin, the analysis will focus on a smaller part of it – the upper Po River basin, an area which has similar hydrological and hydraulic characteristics as the whole Po River basin.

5.1 Overview of the River Networks

The main river Seisa starts at an attitude of 4.6m (Mount Rosa) with a length of 131km. Its gradient decreases as it progresses downstream. The tributaries Elvo, Cervo, Sessera and Mastallone have a length of 4.7km, 5.1km, 7.4km and 2.6km respectively.

5.2 Rainfall Runoff Model

There are in total 14 sub-catchments in the model, where each catchment is physically linked to one of the nodes of the river networks. The area of each catchment can be seen in the table 6-1 below.

In the NAM model, 22 rainfall stations of hourly data are used in the calculation of the area-average hourly rainfall. This is done by overlaying a 2-D inverse-distance-weighted surface map, derived from the rainfall stations' locations, onto the sub-catchments. As for the temperature, hourly data are taken from 15 stations where similar area-average hourly temperature is calculated.

5.3 Data

The data used in this study is from the downstream of the Sesia River (Sesia 200). The data from the period from April to October 2000 is used for the training data set of the error forecast model while data for the months of November and December 2000 will be used for the validation of the error forecast model.

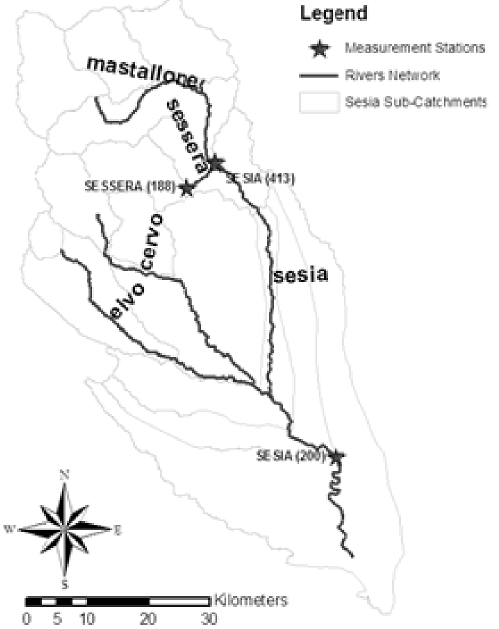


Figure 3. Diagram of the River network and its sub-catchments in lower Po River basin

Table 2. Table of sub-catchments area

Catchments	Area (km^2)
Sermenza A Balmuccia	131.11
Mastall0ne A Ponte Folle	146.82
Elvo A Sordevolo	32.03
Sesia A Borgosesia	247.72
Sesia A Romagnano	162.73
Sesia A Campertogno	170.01
Sessera A Pray	126.19
Elvo A Carisio	229.88
Strona A Cossato	103.46
Cervo A Biella	124.33
Sesia A Vercelli	284.84
Sesia A Palestro	317.89
Cervo A Quinto V.Se	508.87

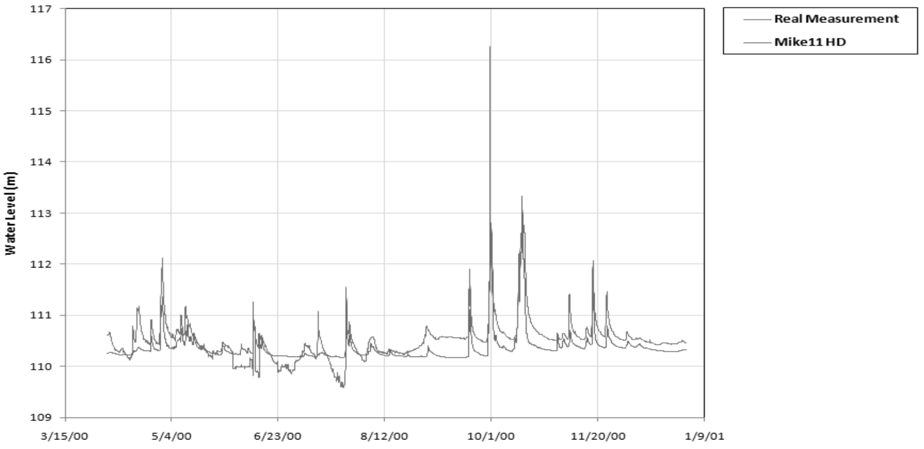


Figure 4. Graph of Real Measurements and Simulated Mike11 Results

5.4 Analysis Results

In this study, the error model, which is the linear regression model, is run with sets of input data, from hourly data of 6 hours, 12 hours, 24 hours, 48 hours and 72 hours. For the analysis below, the error forecast model of 72 hours is used.

Figure 5 shows the peak event which is used to validate the accuracy of the error forecast model. As evident in the graph, one can see that the recursive forecast of the real measurement (purple line) appears to inherit the phase error of the MIKE11 forecast.

Therefore, an assumption is made that for the recursive forecast of the water level based on the MIKE11 simulation results; there exist an additional phase error which is not present during the training of the error forecast model. Using an equation from a research study "Improvement of updating routine in MIKE11 modelling system for real-time flood forecasting" (Morten. R, Jens Chr. R, Karsten. H), where it states that the best fit forecast is the

$$\text{Minimum of} \quad \sum_{i=1}^{N}\left[F_i\left(M_i-\left(S_i+A_e-\frac{S_i-S_{i+1}}{\Delta t}.P_e\right)\right)\right]^2 \qquad (3a)$$

Where Ae = amplitude error (m^3/s), Pe = phase error time (s), M = measured discharge (m^3/s), S = simulated discharge (m^3/s), F = weighting factor, n = number of values taken into account, Δt = time-step (s).

Thus, using Equ (3) as a guideline; 2 general phase error equations are obtained:

$$\textit{Phase Error Equation } 1 = \frac{S_i - S_{i+1}}{\Delta t}.P_e \qquad (3b)$$

$$\textit{Phase Error Equation } 2 = S_i - S_{\left(i+\frac{P_e}{\Delta t}\right)} \qquad (3c)$$

Where, Pe = phase error time (s), S = simulated discharge (m^3/s), Δt = time-step (s)

Therefore for the recursive forecast of the water level, the phase error equation will be included in the error forecast model where:

Final (Real) Error = Error Forecast Model (Amplitude Error) + Phase Error

From figure 5, the phase error is estimated at 6 hours. And the peak event is forecasted recursively 7 times between 6th November 2000 2pm to 8pm. The figure 6 shows the RMSE of the 3 different approaches of the forecast, (i) the normal error forecast model, (ii) the updated error forecast model with Phase Error Equation 1 and (iii) the updated error forecast model with Phase Error Equation 2.

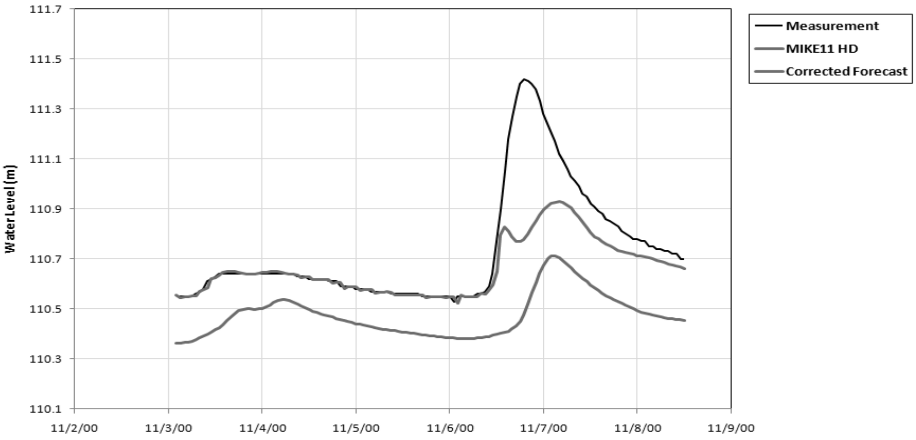


Figure 5. Graph of a flood peak on 7th November 2000

One can draw the conclusion that with the inclusion of the phase error into the train error forecast model, one does get significant improved results, with up to 70% better forecast of the real measurement. Another interesting observation from the results is that the Phase Error Equation 1 gives a better immediate forecast (up to 6 hours) while the Phase Error Equation 1 gives a better estimation of the maximum peak for the water level.

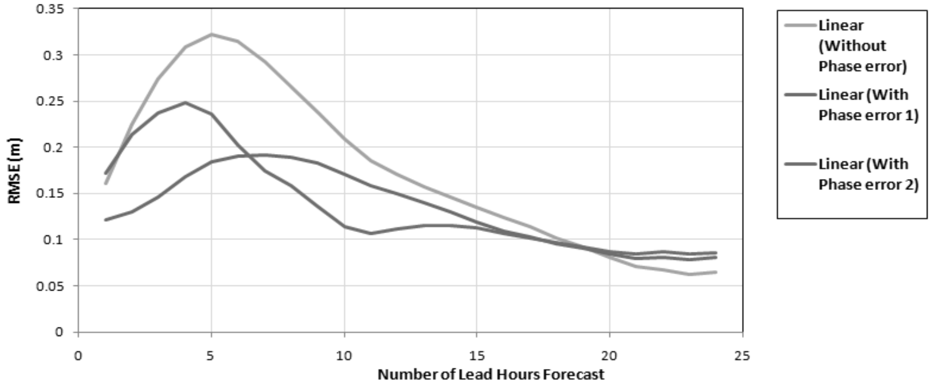


Figure 6. Graph of RMSE of the forecasts with and without Phase Error accounted

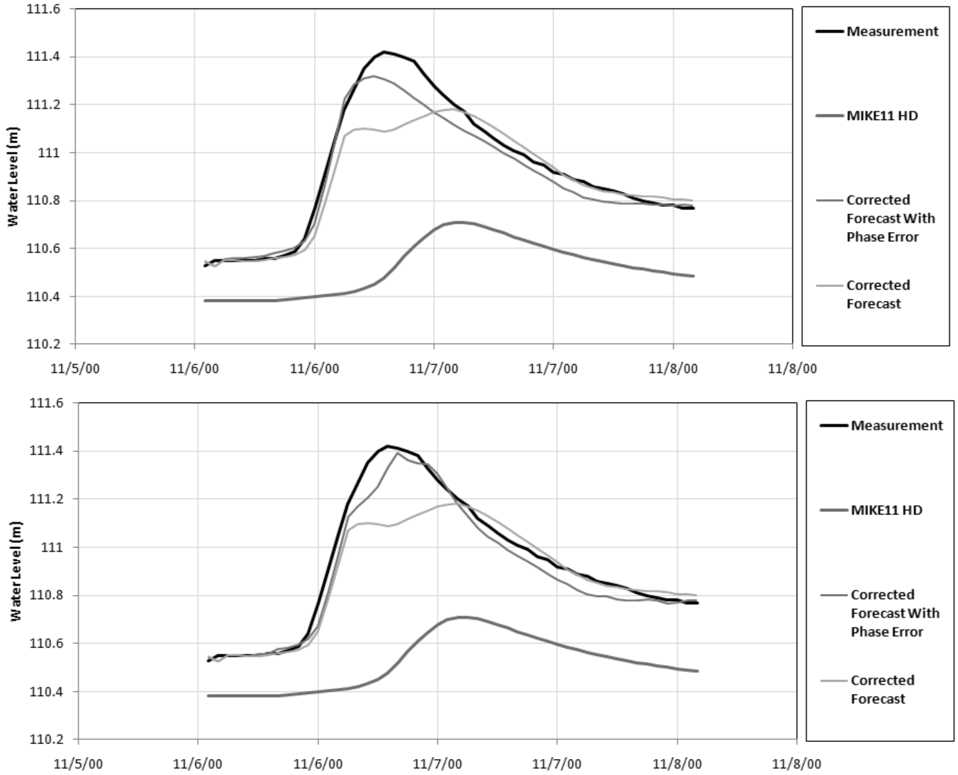


Figure 7. Graphs of Forecast with and without Phase Error accounted
(Top – Phase Error Equ 1, Bottom – Phase Error Equ 2)

Now that it has been established that the inclusion of the phase error term gives a better forecast, different rainfall event will be tested using the estimated phase error time (Pe) to see if that this Pe will hold for other events as well.

From Figures 8a, 8b and 8c, 2 out of the 3 events produce better estimation when compared to the original error forecast model. The event on 18th November 2000 shows that the original error forecast model gives an almost 100% accurate forecast where

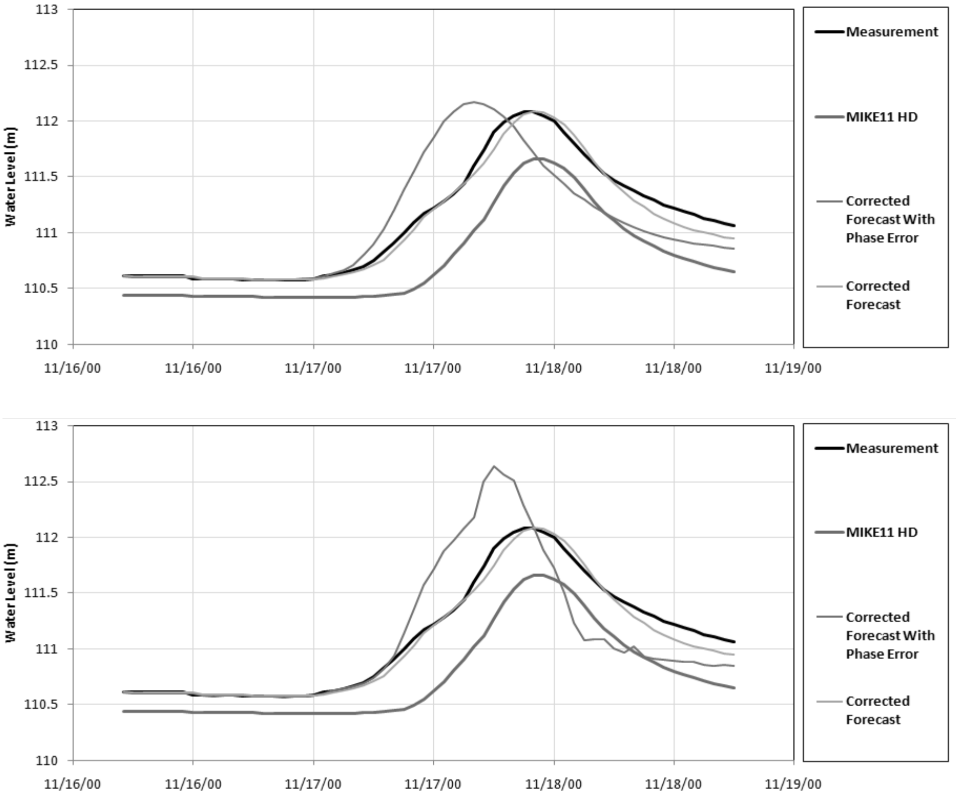
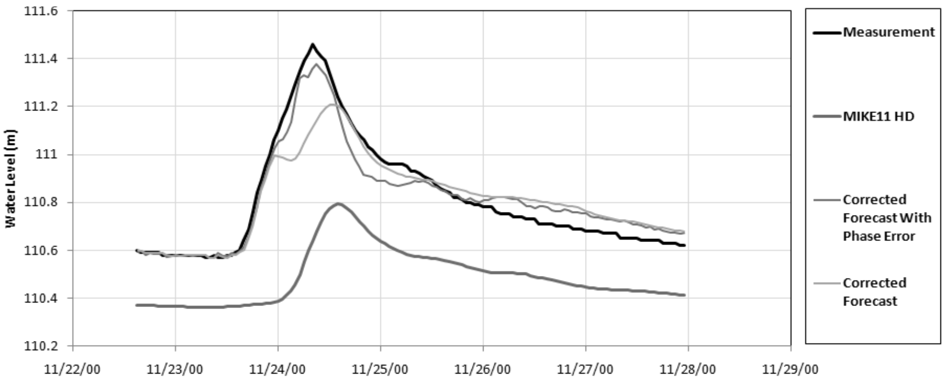


Figure 8a. Graphs of Forecast with and without Phase Error on 18th Nov 2000 (Top – Phase Error Equ 1, Bottom – Phase Error Equ 2)



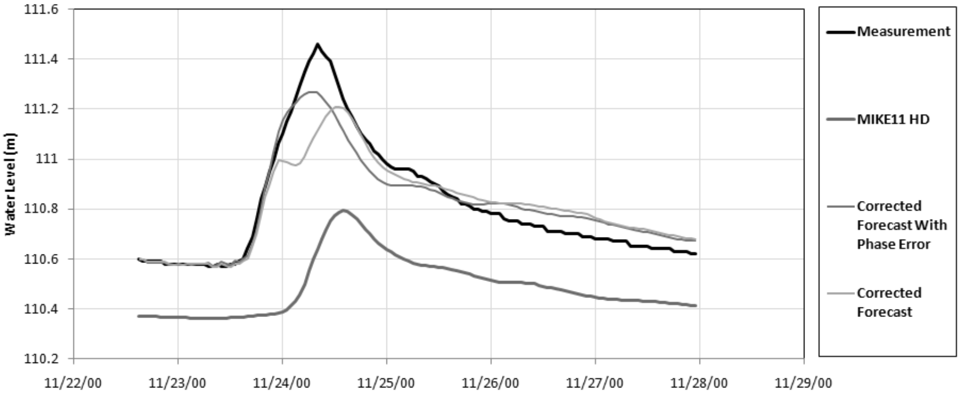


Figure 8b. Graphs of Forecast with and without Phase Error on 25th Nov 2000
(Top – Phase Error Equ 1, Bottom – Phase Error Equ 2)

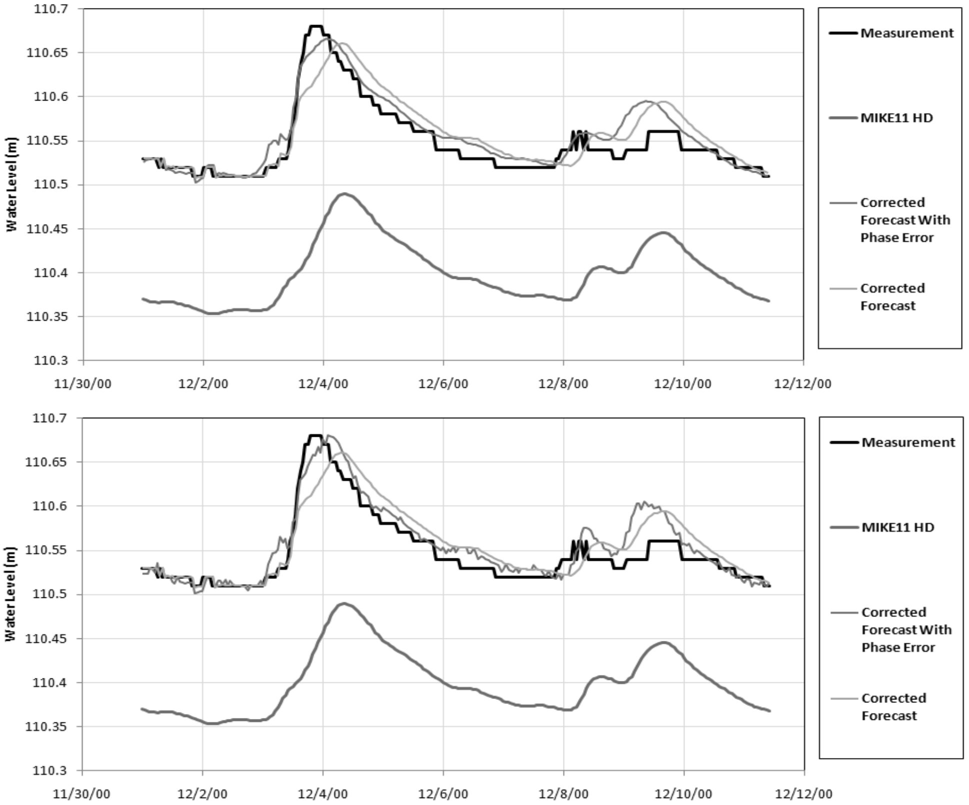


Figure 8c. Graphs of Forecast with and without Phase Error on 4th Dec 2000
(Top – Phase Error Equ 1, Bottom – Phase Error Equ 2)

with the phase error accounted, it actually gives a forecast which is leading the real measurement by 6 hours. This is mainly due to the fact that for this particular event, the recursive forecasts based on the original error model does not contain any phase error.

6. Conclusion

Based on the analysis results from section 5, 67% of the time, the inclusion of the phase error does improve the recursive forecast of the water levels. However, it is important to note that the sample size is too small to draw any just and sound conclusion. Thus, more events have to be tested out to see whether or not the assumption of the existence of the phase error in the recursive forecasts is valid or not. Also, such test should be carried out at different sites too, to determine if this behaviour is location - based or not.

Also, instead of using the assumption of the existence of a phase error, a more sophisticated approach like the Artificial Neural Networks can be used to determine the error between the recursive forecasts and the real measurements.

Lastly, another possible improvement to the accuracy of the recursive forecast is to have a separate relationship equation for the different lead forecast hours instead of using a general one for all the forecast lead hours

Acknowledgements

We would like to thank all the contributors for their valuable research especially Dr. Henrik Madsen of DHI (Denmark) for providing data and guidance during the course of this paper.

References

1. Babovic, V., Canizares, R. ; Jensen, H. R.; Klinting, A., Neural networks as routine for error updating of numerical models, Journal of Hydraulic Engineering, ASCE, 127(3): 181-193, 2001
2. MIKE 11 Reference Manual, DHI Water, Environment and Health (2008)
3. Morten. R, Jens Chr. R, Karsten. H, Improvement of updating routine in MIKE11 modelling system for real-time flood forecasting (1989)
4. Refsgaard, J.C., Validating and intercomparision of different updating procedures for real-time forecasting, Nordic Hydrology, (28): 65-84, 1997

○논문접수일 : 11년 02월 11일
○심사의뢰일 : 11년 02월 18일
○심사완료일 : 11년 02월 28일