A STUDY ON APPLICABILITY OF FLOOD FORECASTING METHOD BASED ON TIME SERIES ANALYSIS USING OBSERVED WATER LEVEL AND RAINFALL DATA

NAOKI KOYAMA
Graduate School of Science and Engineering, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan,

koyama@civil.chuo-u.ac.jp

TADASHI YAMADA
Faculty of Science and Engineering, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan,

yamada@civil.chuo-u.ac.jp

ABSTRACT

The purpose of this study is to verify the accuracy of water level prediction in flood disasters using a multivariate autoregressive model, which is one type of time series analysis. In recent years, serious flood disasters have frequently occurred in all over Asia almost every year. the forecasting information on the water level is the most important information for residents to evacuate. The major problem is to decide how long the lead time (Time needed to evacuate before a disaster occurs) should be. Therefore, flood prediction was performed using a multivariate autoregressive model with water level and rainfall data during heavy rains. In the present study, we studied two cases, in the first case, when calculating using only the water level, it was possible to predict the time of concentration at observation station of used observed water level data and the reference station. The second case, when calculating using the water level and rainfall, it can predict with high accuracy several hours ahead compared to using only water level.

Keywords: Flood forecasting, Time series analysis, Multivariate auto regressive model, Water level, Rainfall

1. INTRODUCTION

In recent years, floods have been occurring every year in all over Japan, causing great damage. Similarly, in Asia, human and economic damage has occurred by floods. Generally, such disaster countermeasures are taken from both structural measure and non-structural measure. There is concern that heavy rains will become severe and frequent due to climate change. According to conventional thinking, “Flood damage is prevented by the development of facilities”, in other words, flood disaster countermeasures by structural measures. But, in the future, there is a need to change the attitude of society as a whole, “the facility capacity is limited, and a flood that cannot be prevented will necessarily occur”, in Japan. Thus, the importance of non-structural measures will increase even more.

Among the non-structural measures, water level prediction is important information that encourages residents to evacuate to save lives. Generally, prediction methods can be divided into two types: runoff models and statistical models. In runoff analysis models which is a physical model, rainfall accuracy is important because rainfall is used as input data. However, hydrological observations are not sufficiently prepared in regions such as emerging countries, and rainfall data with the required accuracy is often not available. In regions where such rainfall observations are insufficient, satellite global precipitation maps might be used. However, one of the satellite global precipitation maps “GsMAP” is very effective in areas without rainfall data, but it is known that rainfall tends to be underestimated (S. SETO., 2008). For this reason, rain gages are still a typical rainfall observation method. We quantitatively evaluated the effect of spatial resolution of rainfall data on peak discharge when performing rainfall runoff analysis. We concluded that if the controlled area (basin area / number of rain gauges) of each rain gauge is within 10km^2, the error of flood peak discharge is within 10%. Also, when controlled area of rain gauge is about 100km^2, it indicates that there is a flood peak discharge error of about 5% to 50% depending on the size of the basin area (N. KOYAMA,2019). For rainfall data that depends on the status of rainfall observation and maintenance in the basin, the river water level representing that point can be observed relatively accurately and is highly representative. Therefore, we considered constructing a flood forecasting model using a statistical model mainly based on water level data. Furthermore, there is a possibility that sufficient rainfall data may be available for some basins, so this flood prediction model can also be expanded by incorporating rainfall information.

In recent years, research on predicting water levels by statistical methods using observation information obtained in real time is frequently performed. In a statistical method, the expression of the rainfall-runoff process is a black box, it is often relatively simple compared to a physical methods, and it has the advantage of not including the error due to the H-Q equation. Time series analysis, such as AR and ARMA have been used as statistical flood forecasting methods (Box and Jenkins, 1970). Since around 1990, flood forecasting research using neural networks (ANN: Artificial Neural Network) has been active, and there have been many cases so far. However, it is said that machine learning requires a more significant number of data when performing prediction by machine learning such as a neural network and time series analysis. In addition, it has been reported that in the general problem of time series prediction, time series analysis has higher accuracy

than machine learning. Therefore, in this study, we used a multivariate autoregressive model of time series analysis for the purpose of constructing a flood forecasting method that can be applied even in the watershed where the hydrological data is scarce.

2. Methodology

2-1 Deviation of water level prediction equation by time series analysis using water level

Time series analysis is one of the statistical models that relate the current states of a system to its past movements, and this method is used for prediction and controlling the system. In time series analysis, the most basic formula is the autoregressive model (AR Model). It is shown in equation (1).

$$h_n = \sum_{i=1}^{N} a_i h_{n-i} + \varepsilon_n \tag{1}$$

Where h_n is the investigated time series; ε_n, a white noise, i.e. a non-correlated, zero-mean random variable that is also not correlated with the past values of h_n; a_i, the auto-regressive parameters; N is the order and represents the number of past data. However, in the case of a river, the water level is not determined at one point but is determined by the river water of various tributaries upstream. Therefore, equation (1) expanded to equation (2).

$$\begin{bmatrix} h_n^1 \\ h_n^2 \\ \vdots \\ h_n^p \end{bmatrix} = \sum_{i=1}^{N} \begin{bmatrix} a_{11}^i & a_{12}^i & \cdots & a_{1p}^i \\ a_{21}^i & a_{22}^i & & a_{12}^i \\ \vdots & & \ddots & \vdots \\ a_{p1}^i & a_{p2}^i & \cdots & a_{pp}^i \end{bmatrix} \begin{bmatrix} h_{n-i}^1 \\ h_{n-i}^2 \\ \vdots \\ h_{n-i}^p \end{bmatrix} + \begin{bmatrix} \varepsilon_n^1 \\ \varepsilon_n^2 \\ \vdots \\ \varepsilon_n^p \end{bmatrix} \tag{2}$$

Equation (2) is expressed in vector form, and equation (3) is expressed in the tensor format.

$$\mathbf{h}_n^p = \sum_{i=1}^{N} \sum_{l=1}^{P} \mathbf{a}_{pl}^i \mathbf{h}_{n-i}^l + \varepsilon_n^p \tag{3}$$

Where P is the investigated number; $\mathbf{q}_n^p$ is water level at station p at n time.

Equation (2) and (3) is a more generalized equation (1), and this is called a multivariate autoregressive model or vector autoregressive model. In addition, equations (2) and (3) expressed that the effect of the upstream of tributaries can be transmitted downstream. This multivariate autoregressive model has been pioneered by Akaike and Nakagawa in the control of cement kiln processes in the engineering field. These equations use past states of a system to calculate current states. Since the water level data can be known in real-time, it is possible to calculate the value one hour ahead by using the formula with the current value. Further, the next time can be calculated by using the predicted value. If we calculate the prediction up to x hours later, equation (6) is obtained.

$$\hat{\mathbf{h}}_{n+x}^p = \sum_{j=1}^{x-1} \sum_{l=1}^{P} \mathbf{a}_{pl}^i \hat{\mathbf{h}}_{n+x-j}^l + \sum_{i=1}^{N} \sum_{l=1}^{P} \mathbf{a}_{pl}^i \mathbf{h}_{n+1-i}^l + \varepsilon_n^p \tag{3}$$

$\hat{h}$ is a predicted value. From this equation, prediction is made using past values and predicted values. Here, prediction intervals were up to 8 hours later to be useful for evacuation information during floods.

2-2 Deviation of water level prediction equation by time series analysis using water level and rainfall

Considering the process of rainfall-runoff, which infiltrates into the soil, flows into the river, and gathers at a certain point. The water level information is the last information in this rainfall-runoff process, and the first is the rainfall information. Therefore, it is conceivable that by adding rainfall information to the model, information before the water level data can be added, and a longer-term prediction can be performed. Equation (4) is obtained by adding rainfall information to equation (2).

$$\begin{bmatrix} h_n^1 \\ \vdots \\ h_n^p \\ r_n^1 \\ \vdots \\ r_n^s \end{bmatrix} = \sum_{i=1}^{N} \begin{bmatrix} a_{11}^i & \cdots & a_{1p}^i \\ \vdots & \ddots & \vdots \\ a_{p1}^i & \cdots & a_{pp}^i \\ b_{11}^i & \cdots & b_{11}^i \\ \vdots & \ddots & \vdots \\ b_{11}^i & \cdots & b_{11}^i \end{bmatrix} \begin{bmatrix} h_{n-i}^1 \\ \vdots \\ h_{n-i}^p \\ r_{n-i}^1 \\ \vdots \\ r_{n-i}^s \end{bmatrix} + \begin{bmatrix} \varepsilon_{qn}^1 \\ \vdots \\ \varepsilon_{qn}^p \\ \varepsilon_{rn}^1 \\ \vdots \\ \varepsilon_{rn}^s \end{bmatrix} \tag{4}$$

This equation does not only predict water level, but also predicts the rainfall while performing the water level prediction.

3. TARGET BASIN

The target basin is the Tone River upstream basin. The basin area is 5100km^2, and the downstream end of this basin is the Yattajima observation station. Figure-1 is the basin map, and each points represent water level stations. In addition, the basin divided into water sheds for each observation station, and the colors are color-coded according to the average value of the time difference between the flood peak water levels obtained from past data based on Yattajima point. It is about 5~6 hours at the maximum. Water level data at each observation point used hourly data.

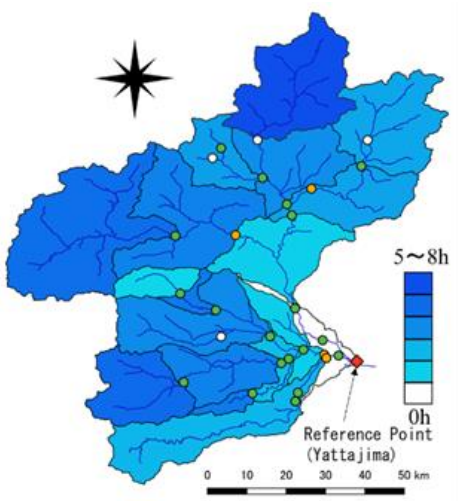

Figure 1. Upper Tone river basin with the location of Water Level Obs. Station, river network and Yattajima points.

4. RESULT and DISCUSSION

4.1 Structure of the Model and Verification of the model reproducibility

In order to develop this model, the number P of stations and the order N to consider the past investigation data must be determined. Therefore, we firstly confirmed the reproducibility with the number N of stations. The number of stations to develop the model was 4 (including only large tributaries), 6 (including medium tributaries), 21 (all stations), in three patterns, the reproducibility of the hydrograph was verified. Fig-2 shows the observation points were used. The September 2015 flood event was used for the verification. This event exceeded the designated water level and is one of the major floods in recent years. Fig-3 shows the result of the reproduction calculation. The blue, green, and red lines are the

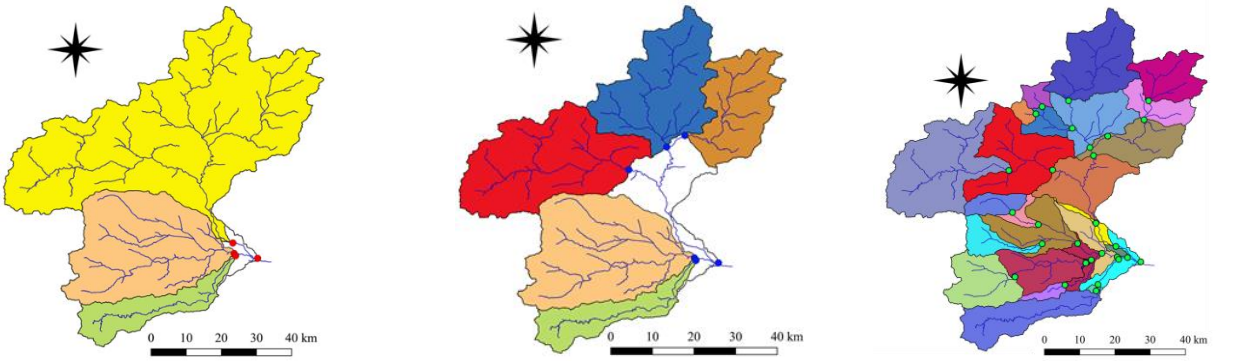

Figure 2. Basin map divided by water level stations(4,6,21 points)

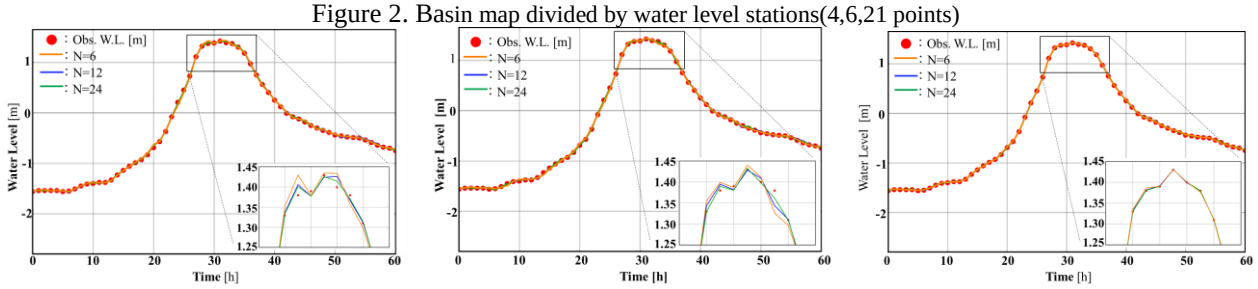

Figure 3. Reproduction results of flood events (4, 6, 21 point)

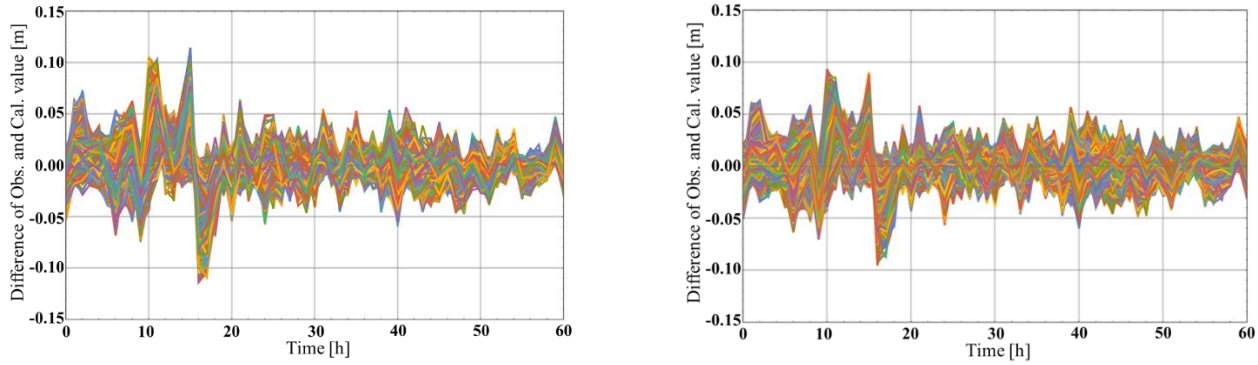

Figure 4. Differences between observed and calculated water level
(Left: using 4 points including the reference point,
Right :6 points including the reference point)

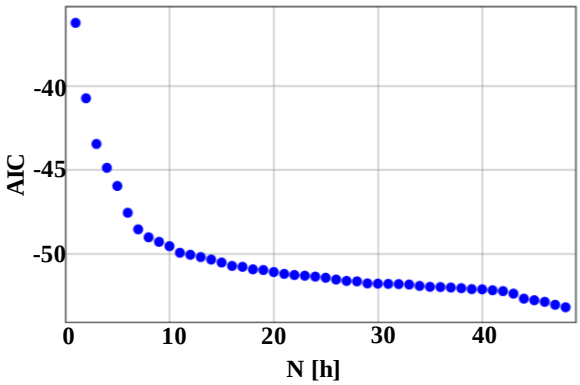

Figure 5. Relation of N and AIC
(AIC value does not change significantly after order 10.)

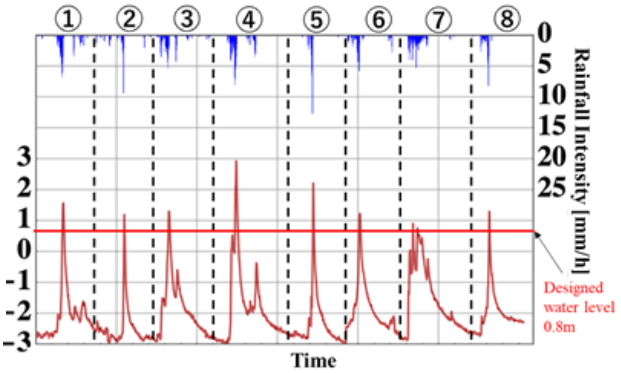

Figure 6. Floods event that exceeded the designated water level at reference point from 2002 to 2018(all 8 events)

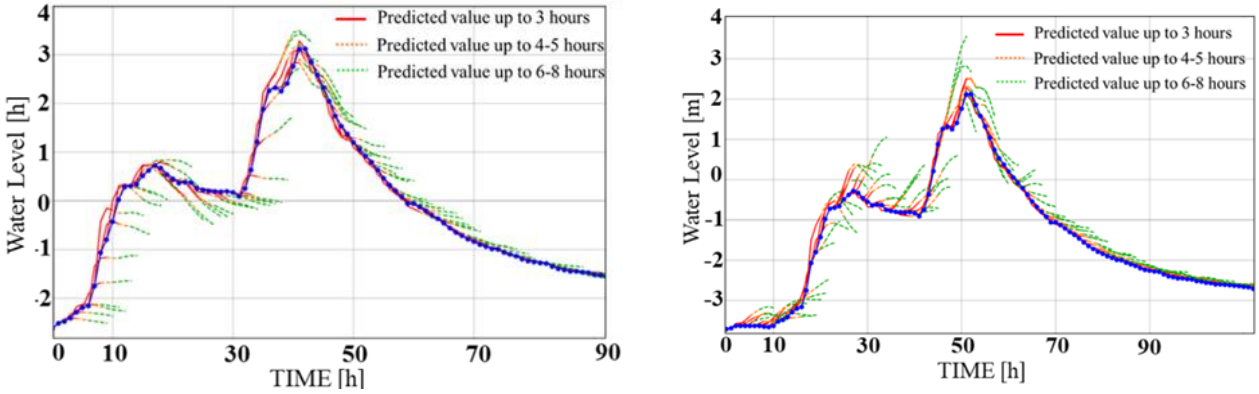

Figure 7. Result of water level forecast in flood event4 at reference point.
(Left : only water level, Right : water level and rainfall)

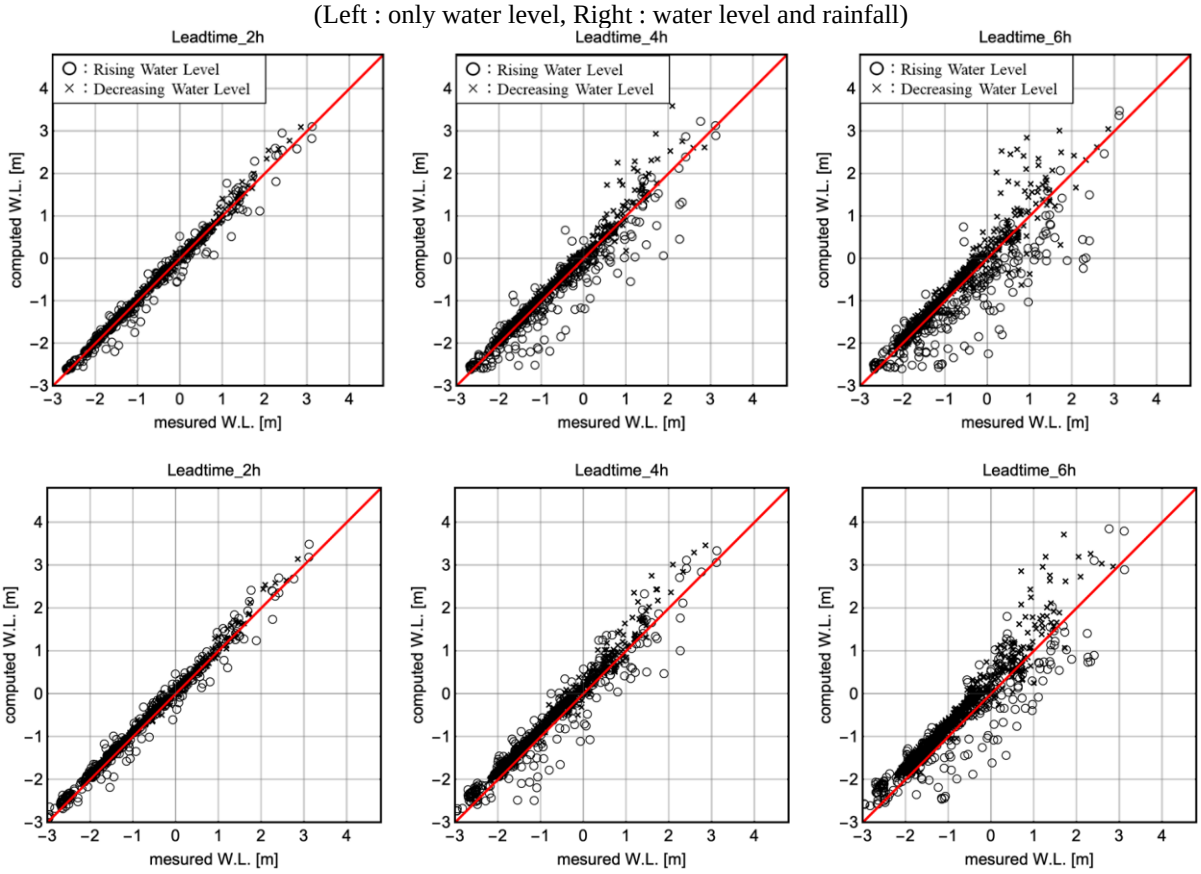

Fig.- 8 Comparison of predicted values and observed values for 2, 4 and 6 hours
(Upper : only water level , Below : water level and rainfall)

results of changing the order N. In each case which the number of N is 4, 6, and 21, the observed value and each line overlapped. Thus, it was found that it could be reproduced sufficiently regardless of the number of stations. The lower right of Fig-3 shows an enlarged view of the area near the flood peak water level. When the number of stations used in the model

is 4 or 6, there are some points that deviate from the observed value near the peak, but it is several cm

and able to be reproduced sufficiently. In addition, it is able to be seen that the accuracy was improved when all the stations were used. The calculations used observations in the major and middle tributaries of the basin. However, in some basins, there is a possibility that hydrological data is insufficiently prepared and data equivalent to the target basin cannot be obtained. Therefore, 4 and 6 points including the reference point were selected from 21 observation stations without duplication, and 1140 and 15504 kinds of reproduction calculations were performed respectively. The result is Fig-4. From this figure, the maximum difference between the calculated value and the observed value is 8% (11cm) and 6% (9cm) at the peak water level, good reproduction was obtained. From this result, it was found that it was possible to reproduce with high accuracy by using multiple observation stations in the target basin. Comparing the result of 4 and 6 stations, the accuracy is higher when 6 stations are used. Therefore, 6 stations will be use as P in the analysis.
Next, the order N in Equation (3) was determined using Akaike's Information Criterion (AIC). AIC is expressed by equation (5)

$$AIC = -2\log(L) + 2K \tag{5}$$

Where L is the Maximum log-likelihood; K is the number of model parameters. Fig-5 shows the relationship between N of order and AIC. This shows that the value of AIC does not change significantly even when considering more than 10 hours ago. Therefore, N was set to 12 in the analysis.
Parameter estimation for VAR models can be performed in several ways. We applied here the least-squares method based on householder transformation. This estimation method can assume that there is a time lag in the response between the variables, and it estimates the model assumed that certain coefficients are zero.

4.2 forecasting and verification of prediction accuracy

In this chapter, we verified the accuracy of flood forecasting. For the parameter estimation data, water levels from May to November from 2002 to 2018 were extracted. The flood event that exceeded the designated water level was defined as one flood event. During that period, there were eight flood events. Figure 6 lists all eight flood events. The numbers in the figure represent each flood event. When estimating parameters of a target event, a method of dividing learning and verification data is generally used, in order to avoid over learning (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000; Dawson and Wilby, 2006; Maier et al.,2010). we performed split cross-validation using all data of limited flood events. The procedure of split cross-validation is as follows. ① Divide all flood events into eight datasets as shown in Fig.-6. ② Use one flood events verification data. ③ Use the remaining seven events as a data set for parameter estimation. ④ Parameter estimation and accuracy verification are performed using the set data. These steps ②, ③ and ④ were repeated to evaluate the prediction accuracy.

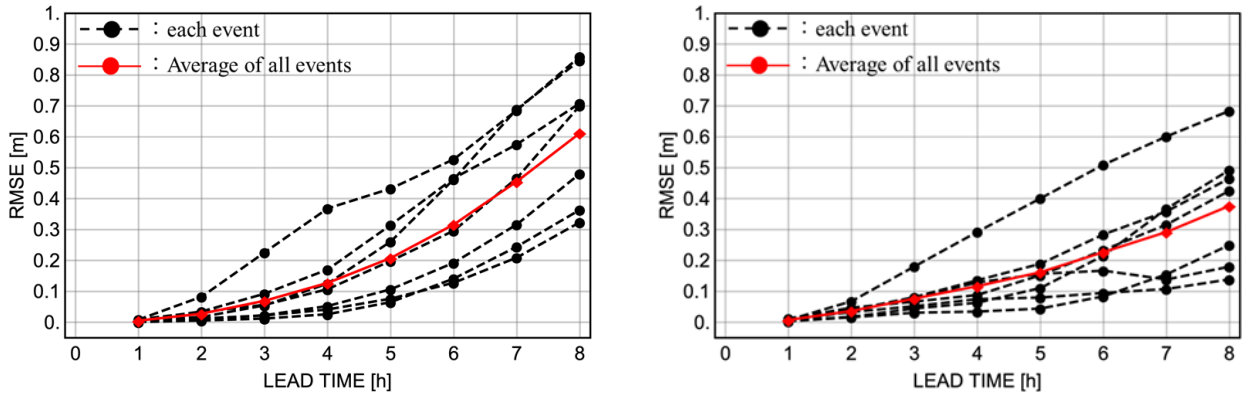

Fig.-9 Difference between observed and predicted water level for each predicted time by RSME
(left: only water level, Right: water level and rainfall)

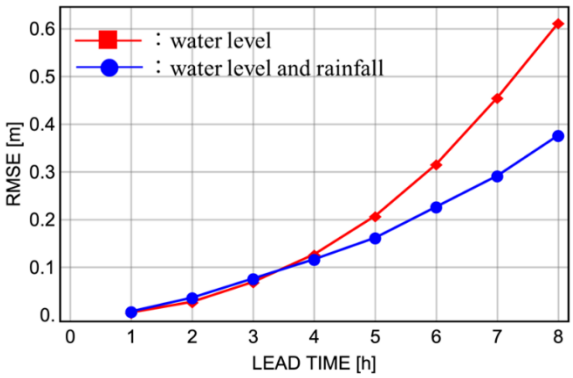

Fig.-10 Comparison of RMSE average value of 8 events using only water level and using water level and rainfall

Fig-7 shows the flood event 4 with the highest water level as a representative of the results of the water level prediction up to eight hours ahead. The water level station used for the calculation is orange as shown in Fig-1. From this result, the predicted values up to 3 hours (red line) can reproduce the hydrograph of the observed values throughout. However, the predicted values after 4 hours tend to be underestimated at the part where the water level rise, and it tends to be overestimated at the part where the water level reduces. Fig.-8 shows the comparison of the observed and predicted values of the hydrograph at Event 4. The upper figure is a calculation using only water level data, and the lower figure is a calculation using water level and rainfall data. From this figure, predicted value of two hours later, both were well predicted. However, it can be seen that the dispersion between the observed value and the predicted value is larger after 4, 6 hours. Especially at the rise of the flood, predicted values tend to be underestimated. Comparing the two calculations, the variability is smaller when including rainfall than when using an water level.

We describe the method of evaluation prediction accuracy. For the evaluation, we use the time from the rise of the water level to the peak water level in the hydrograph, which are important times for evacuation. The evaluation formula is equation (6).

$$RMSE = \sqrt{\sum_{n=1}^{N_{peak}} (h_{ob}^{i} - h_{pre}^{i})^2 / \mathrm{N}_{peak}} \tag{6}$$

Equation (6) is the root mean square error (RMSE). Figure 9 shows the result of the accuracy evaluation. The left side of Fig.-7 shows the prediction accuracy using water level is used, and left side is using water level and rainfall. Each black line is the result for each event, and the red line is the average of all eight events. From this result, it can be seen that in the case of only the water level, the prediction accuracy becomes worse from around 5 hours. On the other hands when the rainfall data is included, the predicted value with high accuracy tends to be extended for several hours compared with the case of only water level. However, the longer the prediction, the worse the accuracy as the water level. This is because rainfall prediction is based on an autoregressive model like water level prediction, and it is not a prediction that considers meteorological viewpoints such as the occurrence and development of rainfall. Fig.-8 shows the comparison of the average value of the accuracy evaluation. From this figure, it can be seen that the accuracy after 4 hours is significantly improved due to the difference in the data used.

5. CONCLUSIONS

This paper addresses the problem of forecasting the water level on the basis of water level and rainfall data. The twofold: First, we wanted to develop tool to analyze river behavior resulting during heavy rain periods; and second, we attempted to set up a model which would be able to forecast the water level of the river on the basis of water level and rain fall information in order to reduce the consequence of floods.

When the model is based on a multivariate autoregressive model with only water level, prediction can be made within 30 cm up to about 4 hours ahead. The reason for the predictable time is the flood concentration time which the reference point and observation station in the model. Furthermore, when rainfall information is added to the model, it is possible to predict with high accuracy several hours ahead as compared with only the water level. This is because information before the water level information was added.

REFERENCES

S. SETO, Kei YOSHIMURA and Taikan OKI (2008). Simulations of Flood Detection All Over Japan by Using High-Resolution Satellite Precipitation Maps, Annual journal of Hydraulic Engineering, JSCE, Vol. 52, pp.355-360.

N. Koyama and Tadashi YAMADA (2019). A Study on the Effects of the Uncertainty of Rainfall Spatial Distribution on the Discharge and Water Level of River on Runoff, Global Environment Engineering Research, JSCE,

Box, G.E.P. , and Jenkins, G. M. (1970). Time Series Analysis Forecasting and Control, Holden Day, San Francisco

Mark N. French, Witold F. Krajewski and Robert R. Cuykendall, (1992). Rainfall forecasting in space and time using a neural network, Journal of Hydrology, Vol. 137.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology(2000). Artificial neural networks in hydrology. I : Preliminary Concept, Journal of Hydrologic Engineering. Vol5, No.2, pp.115-123

Dawson, C. W., Wilby, R.L. (2001): Hydrological modeling using artificial neural networks, Progress in Physical Geography, Vol.25, No.1, pp.80-108.

Maier, H.R., Jain, A., Dandy, G.C. and Sudheer, K.P. (2010) : Method used for the development of neural networks for the prediction of water resource variables in river system: Current status and future directions, Environmental Modelling & Soft ware, 28(2), pp135-147

T. Otomo, T. Nakagawa and H. Akaike (1972). Statical approach to computer control of cement rotary kilns, Automatica, volume 8, PP.35-48.

Hirotugu Akaike (1974), A New Look at the Statistical Model Identification, IEEE Trans. Automat. Contrl., NO.6, pp.716-723.