A Bayesian Network Modeling For Flood Prediction In The Chaophraya River Basin

by

Peerapol Moemeng

Submitted in Partial Fulfillment of
the Requirements for the Degree of
Master of Science
in Computer Science
Assumption University

December 2000

Master Thesis

A Bayesian Network Modeling For Flood Prediction In The Chaophraya River Basin

By

Peerapol Moemeng

Submitted in Partial Fulfillment of
the Requirements for the Degree of
Master of Science
in Computer Science
Assumption University

December, 2000

121574

The Faculty of Science and Technology
Thesis Approval

Thesis Title	A Bayesian Network Modeling for Flood Prediction in the Chaophraya River Basin
By	Mr. Peerapol Moemeng
Thesis Advisor	Prof. Dr. Peter Haddawy
Academic Year	1/2000

The department of CS, the Faculty of Science and Technology of Assumption University had approved this final report of the **twelve** credits course, **SC7000 Master Thesis**, submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Approval Committee:

(Prof. Dr. Peter Haddawy)
Advisor

(Dr. Tang Van To)
Committee Member

(Dr. Jirapun Daengdej)
Committee Member

(Asst. Prof. Dr. Surapong Auwatanamongkol)
Representative of Ministry of University Affairs

Faculty Approval:

(Dr. Thitipong Tanprasert)
Director

(Asst. Prof. Dr. Pratit Santiprabhob)
Dean

December/2000

ABSTRACT

One of the major problems plaguing the Thailand capital city, Bangkok, is flooding during the rainy season, which is determined by two primary factors: local rainfall and water level in Chaophraya River. This project proposal is an attempt to develop a system to predict the water level in Chaophraya River, based on readings of the water level up-river and downstream as well. Traditional statistical and Bayesian network modeling techniques are used to construct models from daily water level data spanning a period of five years. Some preliminary work has been done and will be discussed in this proposal.

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

1. INTRODUCTION

1.1 MOTIVATION AND BACKGROUND

Flooding during the rainy season in Bangkok has major impacts upon communication, transportation, and public health. The Bangkok Metropolitan District spends large amounts of money in attempts to prevent and minimize the impact of flooding. Many other provinces along the Chaophraya River are also affected by flooding. Thailand loses billions of baht[1] because flooding destroys agricultural crops in these provinces.

Flood prediction is an important component in flood prevention. Many factors affect flooding, such as the flood control system, quantity of rain, water transfer rates to nearby areas, and water level in Chaophraya River. The water level in the river affects the amount of runoff it is able to accommodate, as well as affecting the water levels in the connecting klongs[2], into which the city pumps rain water as part of it flood control measures. Thus the ability to predict the river water level is a key step in overall flood prediction.

[1] Baht *n (Thai)*, Thai currency
[2] Klong *n. (Thai)*, a waterway, a path of water that connects to a river

1.2 EXISTING APPROACHES TO BUILDING FLOOD PREDICTION SYSTEM

The Royal Irrigation Department of Thailand (RID) is responsible to for monitoring the amount of the water that flows in Thailand's rivers and warning any department that is in charge of flood prevention. They predict the water level by using a software package, MIKE-11[3], which predicts the amount of water in the river based on factors such as, soil type, geographical information, current and past amount of water in the river/soil, underground water, and also water level of Thai bay[4]. The most difficult part of the software is calibrating the model, which has many parameters involved. It takes time to adjust the parameters, since some parameters (e.g. dispersion) must be a analyzed in the laboratory (the specification of this hydrology software will be discussed later). After that, they run the model and determine the accuracy by comparing the outcome result with the actual data. Operating the software is not a problem, since the staff can be trained, but calibrating the model takes hydrological modeling knowledge.

The model is fixed with a set of parameters that best describe the hydrological information in the area. Those parameters need to be updated at sometime later when the model begins predicting not precisely.

RID informs that, the water level that is determined to be flooded is 150 cm. With MIKE-11, the accuracy is 4-8 cm, which is very high. But it would be more error

[3] MIKE-11, a system for the 1D modelling of rivers, channels and irrigation systems, including rainfall-runoff, advection-dispersion, morphological, water quality and two-layer flow modules by Danish Hydraulic Institute (DHI)

during the rainy season and flooding period, approximately the accuracy is up to 30 cm. Nowadays, RID is not predicting the water level anymore, because they do not have access to Thai-bay tidal information to input to MIKE-11.

In the past, RID uses expert's prediction[5] together with MIKE-11 to predict the water level in Chaophraya River in Bangkok. Consider Chaophraya River in Bangkok area, there are 3 major factors involved the water level in Chaophraya River; incoming water from upriver, incoming water from small connected rivers, and rain. So expert's prediction is still precise, since the prediction result is only "The level will increase" Or "The level will decrease", but not a number.

1.3 IN THIS THESIS APPROACH

Base on an assumption that the water level on the upstream affects the water level at downstream. In this thesis, we represent the problem of predicting water level. So far, there is no flood prediction system has been done with probabilistic model. Other models were done using mathematic model, neural network [1], and hydrological model [2][3]. In this thesis also represents the ability of Bayesian modeling in predicting continuous value. Water level is measured in numeric unit, while Bayesian results in probability distribution over set of possible states [4]. Using Bayesian network is useful since the technique encodes any kind of relationship, which water level among each station could be related in a non-linear function. It is not practical to discretize the water level into small ranges in a variable, because it could produce very large conditional probability tables (CPT). So we purpose 2 discretization methods and

---

[4] Water level of Thai bay is an input to MIKE-11, informed by RID staff on November 2000.

compare the performance of both methods, which these 2 methods will be described later.

We construct the models by learning Bayesian network from the existing data, which there are 2 datasets which both of them are about water level.

1. RID provided data on daily and hourly water level readings from seven stations along the Chaophraya River.
2. Tide level data was obtained from Marine and Atmosphere Science Information Center, State University of New York, at Stony Brook, shown in table 1.

STATION	Province	DATA POINTS
C3	Singburi	2454
C7A	Angtong	61366
C34	Ayuthaya	7842
C31	Patumtani	12357
C22	Nontaburi	34014
C12, C4	Bangkok	19861
TIDEBKB	Thai Bay	122736
TIDEMKL	Thai Bay	122736
TIDESMS	Thai Bay	122736

Table 1 The seven stations and Thai bay data at which daily water level readings were taken.

[5] As we have interviewed with Mr.Phonchai Klinkhachorn, Engineer, Hydrology and Water Management Office, Royal Irrigation Department.

1.4 DESIGN CRITERIA

Simplified parameters

The parameter set is minimized. Input variables are continuous variables, which are states of water levels. The model will not be difficult to use, since the data input to the model is simplified to the same type of data (the water level only).

Accuracy and Precision

Bayesian network output is a probability distribution over states. We compute the expected value from the probability distribution to get one number as prediction. The prediction should be precise (answer in the right range of water level), and accurate (the different of prediction and actual result should be minimal)

1.5 OUTLINE

After this section, we will discuss about the methodology that we have applied to the available data. Some properties of Bayesian network those are interesting. The prediction in continuous value is described in chapter 2. In chapter 3, we discuss about the overview system, how we implement the system. Chapter 4 is the result of the experiment. Chapter 5 is the analysis of the result obtained from the experiment. Chapter 6, we discuss about other researches that are relevant to this thesis.

2. REPRESENTATION OF FLOOD PREDICTION WITH BAYESIAN NETWORK

2.1 WATER LEVEL DATA

We have several sets of data, but these sets are not smooth. Some of data points are missing. Figure 1 shows the chart that is plotted based on smoothen data using Window Sliding algorithm on intersected data, which we obtained 1102 data points of every station of each time of day.

Figure 1 shows the trend of the C4 and C12 data in the function of time, which are water stations in Bangkok. They are almost the same pattern, which yields in high correlation coefficient.

Figure 2 shows the trend of all stations in the function of time, which also look alike in each station (excluding C4 and C12). The water station at upstream of the river has high water level, and relatively low along downstream of the river.

Figure 3 shows the trend of TIDE data, which also look similar for each dataset.

2.2 SELECTED ATTRIBUTES

We compute the correlation coefficient of each station, to determine the straight of relationship among stations (shown in table 2). The correlation coefficient is computed based on intersected data, because we need the complete data[6] to construct the network. The completeness of data is shown in table 4. The correlation coefficient shows

1. The interested output is the water level at C4 station, so we set output variable to be C4. Because RID is interested in predicting water level at C4 only.
2. Correlation coefficient among TIDE data TIDESMS, TIDEBKB, TIDEMKL are high. Which in the case, we can choose one of the datasets to be a representative of the TIDE data. Here we choose TIDESMS dataset to represent the Thai bay water level, since it has highest correlation coefficient with C4.

In our preliminary work, we have discussed the assumption bases on correlation coefficient between stations in sequence of days, and picked the best couple to design the model. We do not couple the stations on different days, because determining from the chart, shifting the chart by a next day or two, would make the correlation coefficient becomes lower. The physical distance, amount of water, and the speed of water flow between stations are not defined, so we cannot determine the number of days that the water flows from upstream to downstream.

[6] Complete set of input parameter to instantiate the network model.

CORREL

C3								
0.983636	**C7A**							
0.692566	0.742675	**C31**						
0.605941	0.637903	0.91465	**C22**					
0.452741	0.482045	0.770048	0.906534	**C12**				
0.369015	0.391716	0.72879	0.860152	0.859447	**C4**			
0.122928	0.133793	0.316062	0.483173	0.669323	0.540836	**TIDEBKB**		
0.090975	0.090672	0.271978	0.459308	0.656981	0.55993	0.978458	**TIDEMKL**	
0.089103	0.08795	0.256847	0.443316	0.643349	0.567547	0.979053	0.996952	**TIDESMS**

Table 2 Correlation Coefficient of all stations

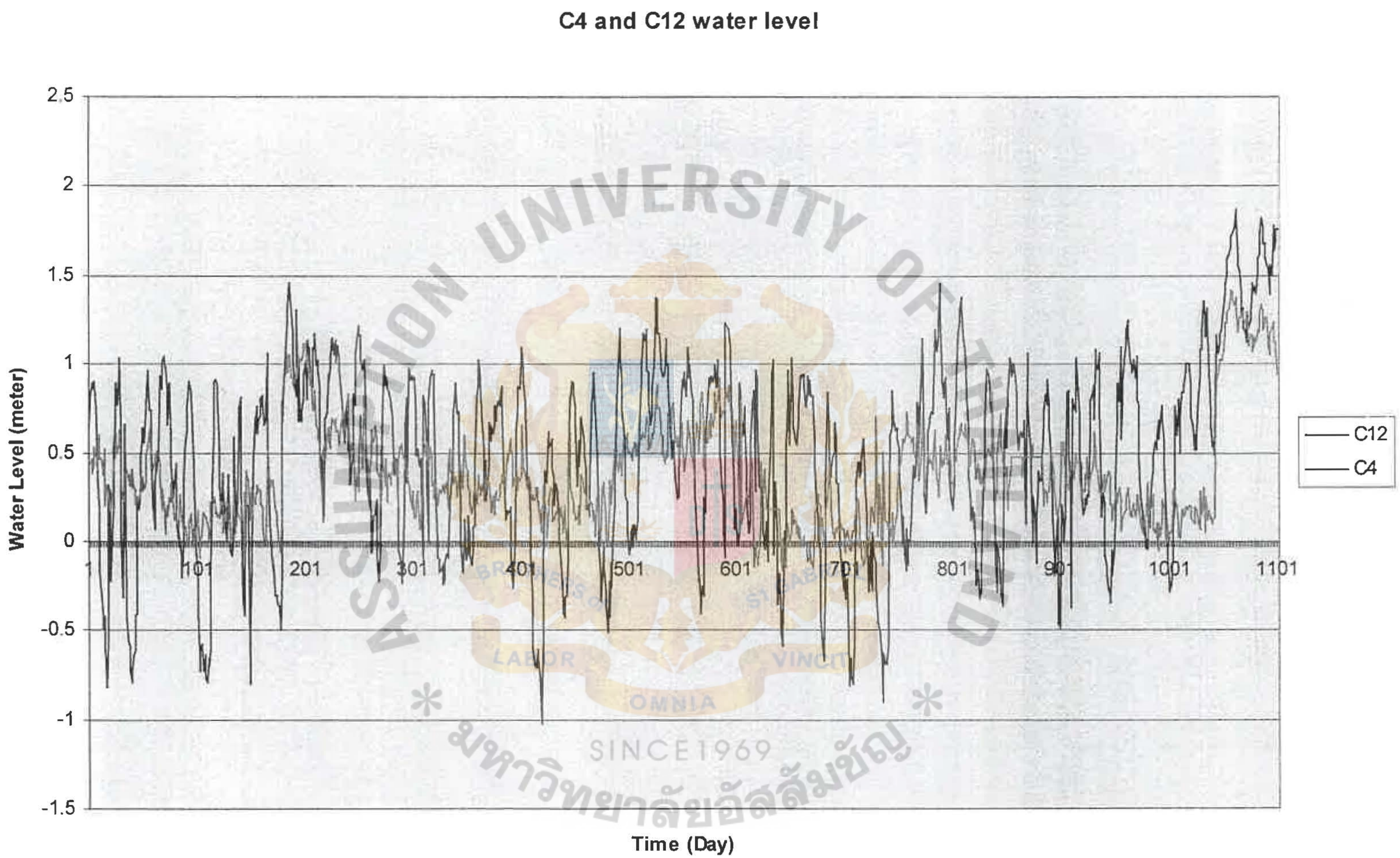


Figure 1 C4 and C12 water level

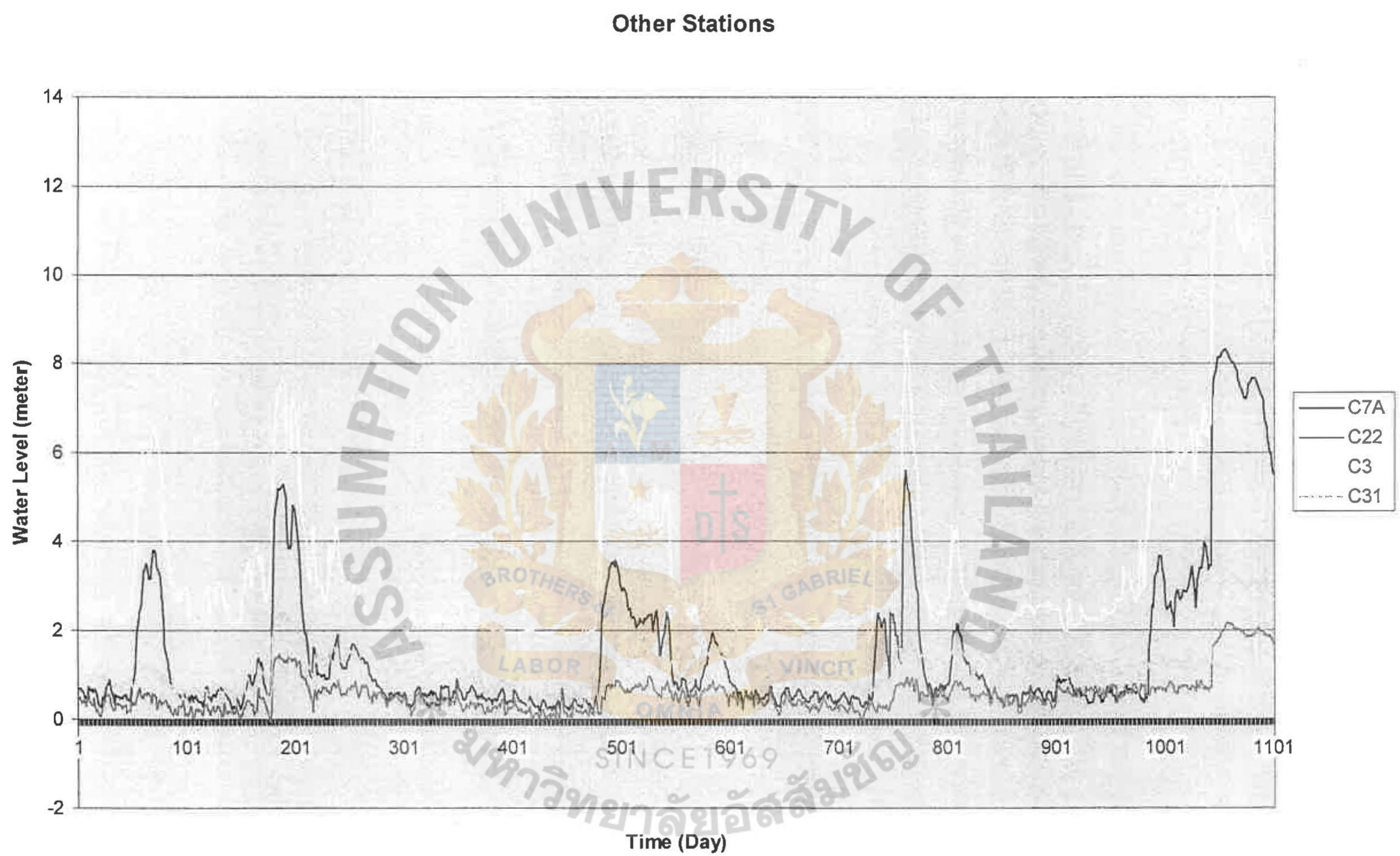


Figure 2 Other stations water level

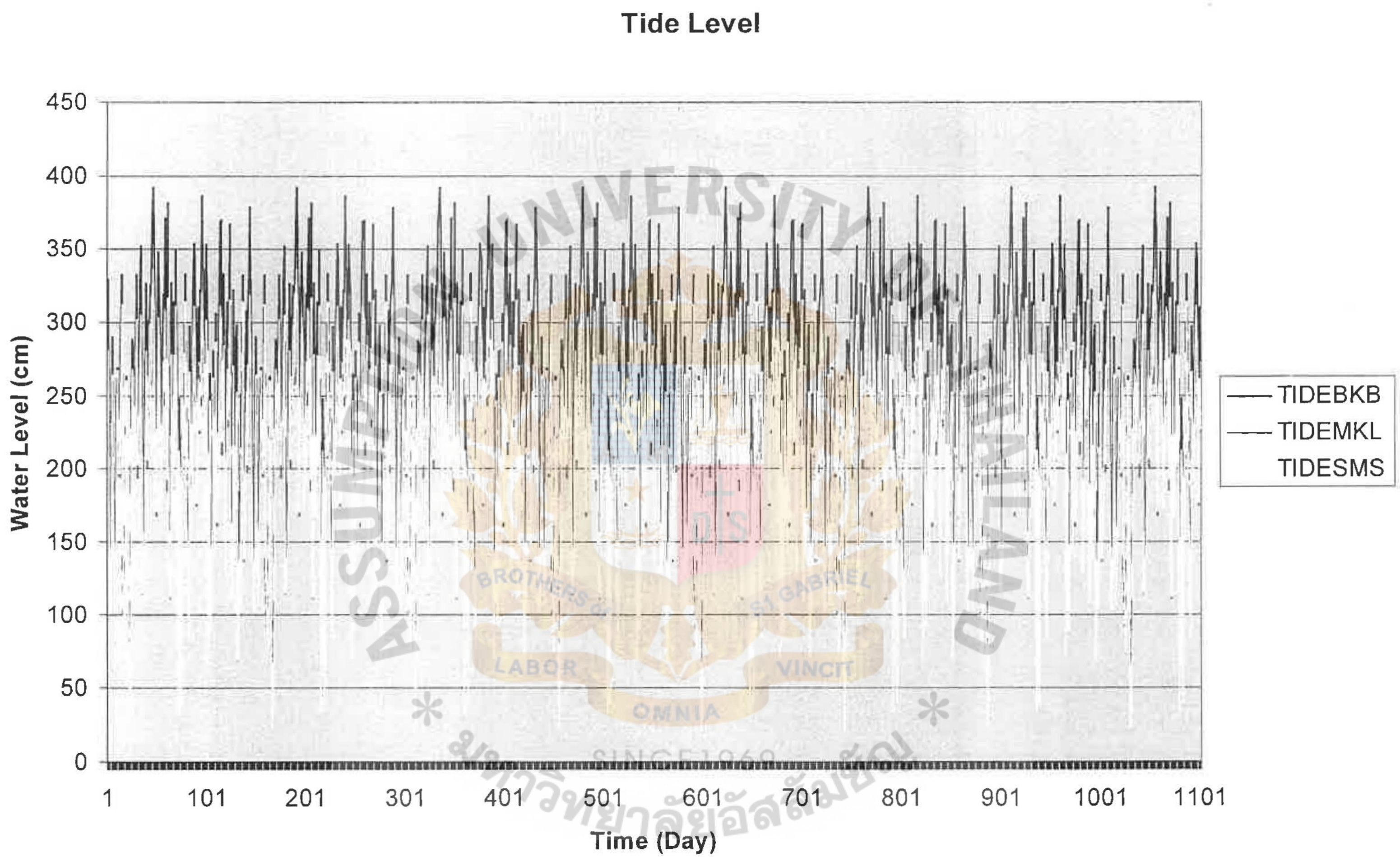


Figure 3 Thai bay water level

2.3 LIMITATION OF BAYESIAN NETWORK PREDICTING THE CONTINUOUS VALUE

One limitation of Bayesian network is Bayesian network results in probability distribution over states. Which each variable must be discretized[7] into set of states. The continuous value is an issue to be talked about here, because the continuous value discretization is defining the range of numbers. When we discretize the continuous value into set of intervals, of course that there will be data lost during discretization. There is no rules to judge the number of intervals and the size of intervals, so we try several discretization, say 10, 15, 20, 25, 30 intervals with 2 methods, which will be discussed in the later section.

[7] A mechanism for defining data into set of states

3. OVERVIEW OF FLOOD PREDICTION SYSTEM

3.1 MODEL DESIGN

Model design, we come up with 2 models for different experiments on different discretization method, which are mixture of METHODS and STATIONS.

1. METHODS, there are 2 methods (describe in next chapter) in discretizing the water level which are discussed in previous section. We will discretize into 10. 15. 20. 25. 30 intervals.
2. STATIONS, as we have to experiment on different models, so one way to do is to simply include every dataset that we have into the model, we name this kind of model as ALL STATION MODEL. We also choose the best couple of datasets from correlation coefficient, which we have chosen C31 ⇨ C22 ⇨ C4 ⇨ TIDESMS, we name this kind of model as BEST STATION MODEL.

There are 2 different models, different 5 interested intervals for 2 discretization methods. So we will come up with 20 experiments.

3.1.1 CONSTRAIN NODES

While we were choosing the discretization of the water levels to avoid conditional probabilities with zero in the denominator, because of insufficient data. This will nonetheless occur due to physically impossible situations. For example, it is not physically possible to have the highest possible water level at station C4 and the lowest possible at C22 on the same day. To prevent our model from assigning positive probability to such physically impossible combinations of variable values, we use constraint nodes [4], denoted as 2-station-name node in the figures. For example, the node labeled C3_7A constrains the combinations of values that the nodes C3 and C7A can assume to rule out impossible combinations. We can then simply assign uniform probabilities to the conditional probability table entries for node C3 conditioned on physically impossible combinations of variable values, since these will not affect computations.

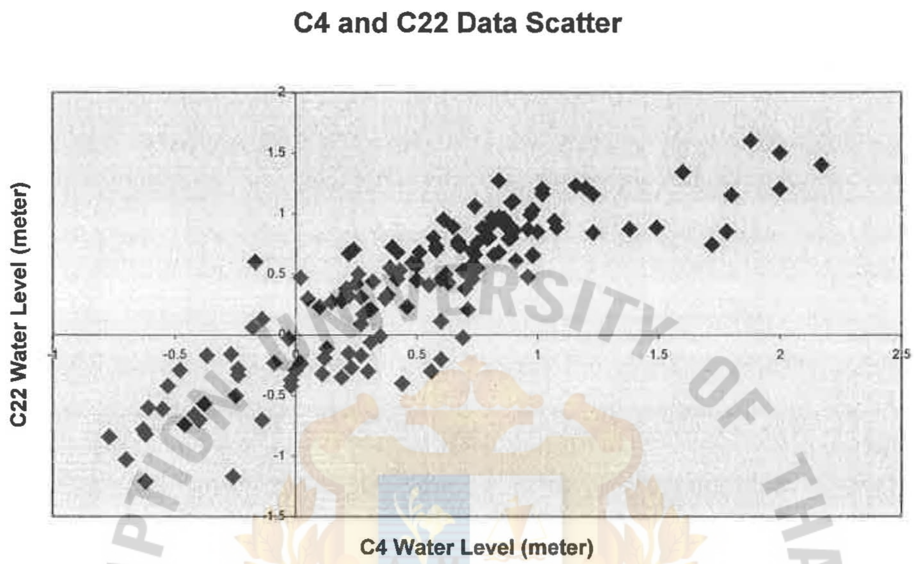


Figure 4 C4 and C22 Data Scatter

Figure 4 shows the data scatter of C4 and C22. The data lies in the diagonal, which represent the relationship between C4 and C22 water level. Other couples of nearby stations are similar to the figure 6, since it is reasonable that water should be related in the same day. Therefore, there should not be data lies at top-left and bottom-right area, since it is impossible to have a very high water level and very low water level of nearby stations in the same day. The following figures, they do not have constrain nodes between the first slice of the network, because these nodes will be instantiated with the real inputs anyway, so there will not be Impossible case happens.

The Bayesian network models we have constructed are called time-sliced Bayesian networks because they represent the state of a system at discrete time slices. Each model is equivalent to a first-order Markov chain. The set of states of the Markov chain is just the cross product of the domains of the variables at a given time, e.g. day zero. The conditional probabilities governing the transition from the state at one time point to the state at the next time point are static, i.e., the same for every transition. The conditional independence semantics associated with the network topology enforces the Markovian assumption that the probability of being in a given state is conditional only on the state at the previous time point.

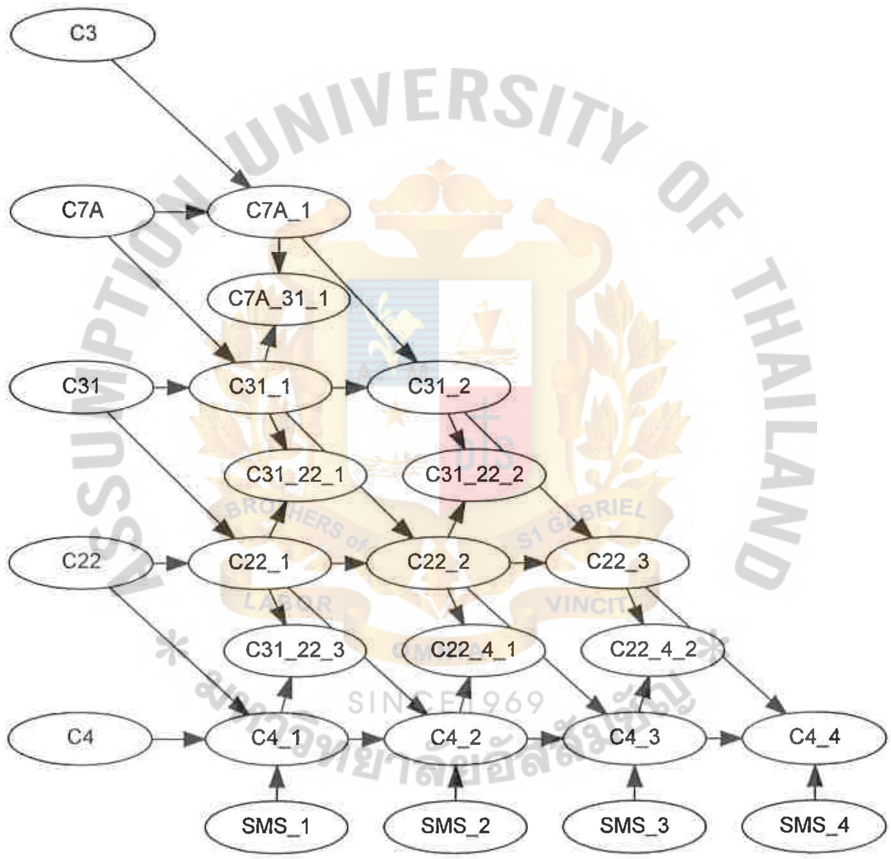


Figure 5 ALL STATION model

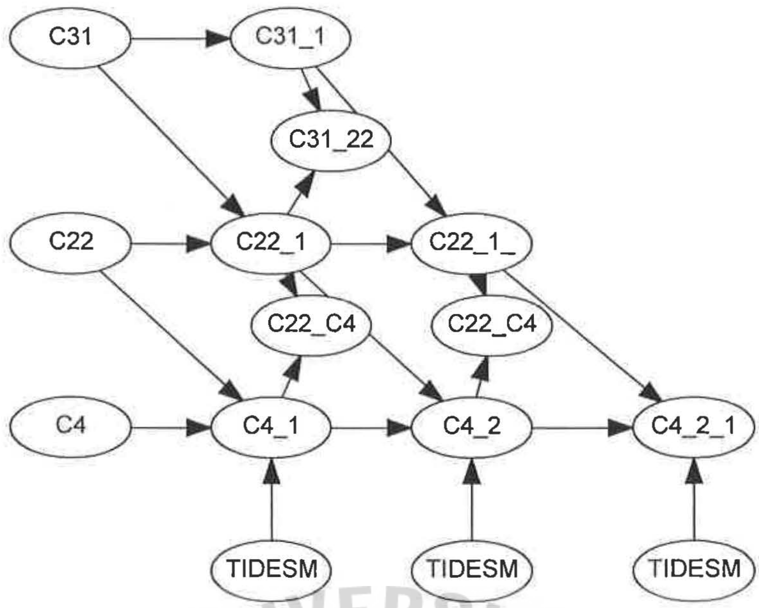


Figure 6 BEST STATION model

Figure 5, ALL STATION models have chain node up to number 4, which it could use to predict for 4 days in advance, but in order to compare the result with BEST STATION model (figure 6), we scope down to 3 days prediction only.

There is no TIDE data influent the C4 node because the C4 is to be instantiated anyway, which this causes the TIDE data that influent C4 be d-separated.

The TIDE level data does not influent the next day TIDE level, because the TIDE level that we have is a predicted result, not the actual tide level, so the predicted value would not influent the next day predicted value.

One consideration to include month into the model, we have reviewed the possibility and found that, it is good idea to have month to cooperate with the model to tell the season and time that water level should begin to increase or decrease. But including month data will enlarge the CPT by 12, which there is not enough data to fill in this large CPT.

3.2 DATA PREPARATION

With enhancement of database software, we execute the SQL statement for different sets of data as required for each model. Since the validity of water level data are not completely same format in each station and also in each day, so we average the water level in a day, so we will have 1102 data points per station. We use this set of data to compute the CPT and also for testing. In this thesis, the supervise learning is used. We apply the same train and test data set to design and test result.

3.3 DISCRETIZATION METHODS

In this thesis, we use 2 discretization methods to compare the result of prediction.

3.3.1 METHOD I

The interval sizes are evenly equal. We compute the size of intervals from number of intervals. The lowest and highest water levels are left as opened ranges. The interval size is the division of range of water level in each station with number of interested intervals. For example,

> C31 highest water level is 3.26 meters, and lowest water level is 0.04 meters. So the range is 3.3. For 10 interested intervals, the size of each interval is 0.33 meter. We add extra size for the ended range, say 1 meter. Example of discretization is shown in table 3.

INVERVAL	LOWER BOUND (OVER)	UPPER BOUND (NOT OVER)	PROBABILTY
1	-0.96	0.04	0.056
2	0.04	0.37	0.070
3	0.37	0.7	0.116
4	0.7	1.03	0.138
5	1.03	1.36	0.192
6	1.36	1.69	0.167
7	1.69	2.02	0.150
8	2.02	2.35	0.076
9	2.35	2.68	0.028
10	2.68	4.26	0.007

Table 3 Method I Discretization for C31

With this discretization method, the probability distribution is a fine bell-shape, since each class is equal size, and the water level data is naturally distributed.

3.3.2 METHOD II

The interval sizes are not equal, but we expect the number of data entries to fit into each interval to be equal instead. We begin with the lowest water level, and then try to adjust the size of intervals that each interval has similar number of population. So the interval sizes are not equal and the probability distribution is almost uniform.

INVERVAL	LOWER BOUND (OVER)	UPPER BOUND (NOT OVER)	PROBABILTY
1	-1.72	0.63	0.11
2	0.63	0.8	0.1
3	0.8	0.94	0.1
4	0.94	1.08	0.09
5	1.08	1.18	0.1
6	1.18	1.26	0.1
7	1.26	1.34	0.1
8	1.34	1.44	0.09
9	1.44	1.62	0.1
10	1.62	5.34	0.11

Table 4 Method II Discretization for C31

3.4 CONSTRUCTING THE CONDITIONAL PROBABILTY TABLE FOR THE MODELS

Since we have data for every node in each of our models, we calculate the CPT entries for each node by simply using the definition of conditional probability. For example, if a node A has parents B and C, then the conditional probability table entries are computed as

$$P(A \mid B,C) = \frac{P(A,B,C)}{P(B,C)}$$

In our preliminary work, we suggested to prevent the zero denominator in computing CPT, by setting up constrain nodes. In this thesis, we do not have to set up constrain nodes to prevent the impossible case and insufficient evidence to fill in CPT, because we initialize every cell in CPT to be a small number, say 0.00001, hence there is no zero. In this thesis, the numbers represent are 4 decimal points, so the fifth decimal points number should have very less effect to the result.

4. EXPERIMENT RESULT

4.1 EVALUATION CRITERIA

Prediction is the result of computing and compares to the actual data, which there are 2 issues to think of, and we also take these 2 issues as our criteria.

1. Accuracy,

   Predicted value is a middle point of the expected range (interval) that we obtain from probability distribution. For example, the expected value shows in a distribution as following table 5. Network model that has discrete value of state intervals. After we get the expected value, we compute the accuracy by finding the difference between the value that we predict and the actual value in the data file. If the actual value lies in the expected interval, then the error is zero. Otherwise, the error is the distance between the actual value and the closest end point. Predicted value is computed to get a predicting number, as the following formula

$$e = \frac{\sum_{i=1}^{n}(Pi * UBi) + \sum_{i=1}^{n}(Pi * LBi)}{2}$$

e = predicted value

P = Probability of

n = number of states

UB = Upper bound value

LB = Lower bound value

C22				
Lower	**Upper**	**Probability**	**Lower * Prob.**	**Upper * Prob.**
-10	-0.77	0.006944444	-0.069444444	-0.005347222
-0.769	-0.6	0.034722222	-0.026701389	-0.020833333
-0.599	-0.43	0.027777778	-0.016638889	-0.011944444
-0.429	-0.26	0.041666667	-0.017875	-0.010833333
-0.259	-0.09	0.055555556	-0.014388889	-0.005
-0.089	0.08	0.055555556	-0.004944444	0.004444444
0.081	0.25	0.104166667	0.0084375	0.026041667
0.251	0.42	0.118055556	0.029631944	0.049583333
0.421	0.59	0.090277778	0.038006944	0.053263889
0.591	0.76	0.145833333	0.0861875	0.110833333
0.761	0.93	0.145833333	0.110979167	0.135625
0.931	1.1	0.076388889	0.071118056	0.084027778
1.101	1.27	0.027777778	0.030583333	0.035277778
1.271	1.44	0.006944444	0.008826389	0.01
1.441	1.61	0.013888889	0.020013889	0.022361111
1.611	1.78	0.013888889	0.022375	0.024722222
1.781	1.95	0.013888889	0.024736111	0.027083333
1.951	2.12	0.013888889	0.027097222	0.029444444
2.121	10	0.006944444	0.014729167	0.069444444
Expected Interval (average)			0.342729167	0.628194444
Predicted Value (average)				**0.485461806**

Table 5 Sample of computing expected interval and value

2. Precision,

   The precision is the expected range that we could predict. Precision can be computed by computing probability distribution for each interval and average them.

4.2 RUNNING TEST

With supports of HUGIN Professional 5.4[8], we can interface with the HUGIN propagation engine through set of API that allow us to load the model into HUGIN engine and run the test by iterating entering information to the models and produce the output. The program we develop for this testing purpose was written in Visual Basic, interface with ORACLE database, and interface with HUGIN API through ActiveX server (new feature in HUGIN Professional 5.4). We develop separate sub program for each models and methods to run the test.

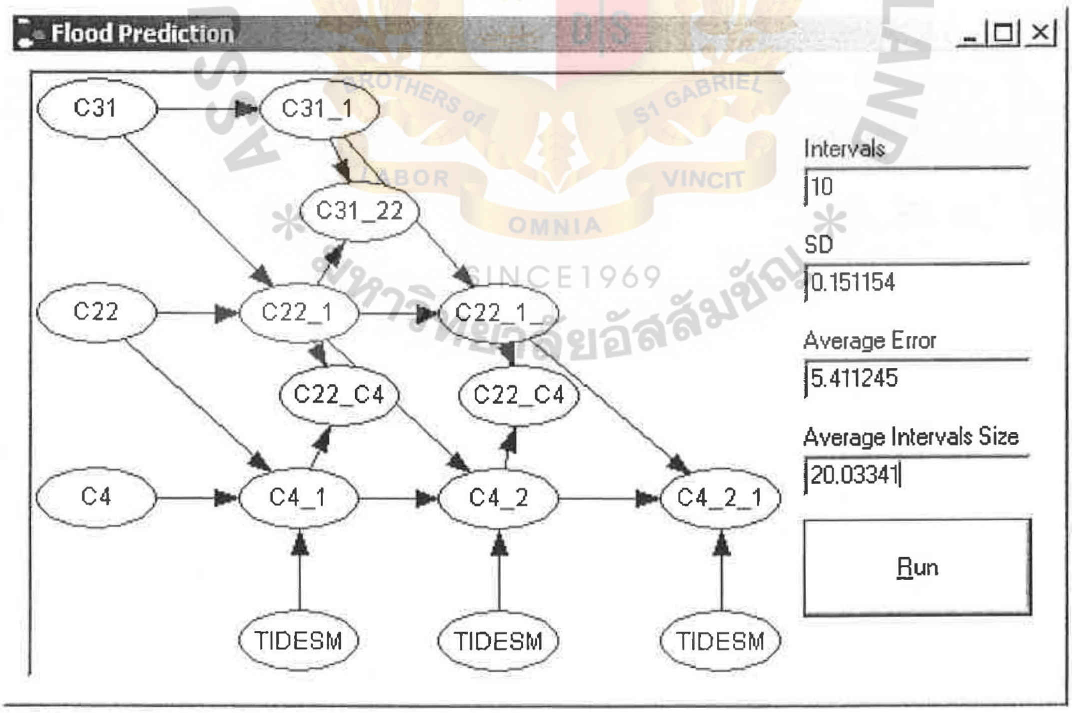


Figure 7 Sample program for testing the model

[8] Thank you to Dr.Peter Haddawy for supporting this software with his Ph.D. license.

4.3 *TEST RESULTS*

Explanation of the figure

The labels 10, 15, 20, 25, and 30 on the figures 8-19 are the numbers of intervals co responding to the data point on the chart.

Figure 8, Figure 9, Figure 10 ALL Model Overall Performance

The figures are result of ALL models prediction in day III, II, and I respectively. These figures show the trade-offs comparison between method I and II in predicting water level on overall data.

Figure 11, Figure 12, Figure 13 ALL Model Prediction During High Water Level

The figures are result of ALL models prediction in day III, II, and I respectively. These figures show the trade-offs comparison between method I and II in predicting high water level on high water level data only.

Figure 14, Figure 15, Figure 16 BEST Model Overall Performance

The figures are result of BEST models prediction in day III, II, and I respectively. These figures show the trade-offs comparison between method I and II in predicting water level on overall data.

Figure 17, Figure 18, Figure 19 BEST Model Prediction During High Water Level

The figures are result of BEST models prediction in day III, II, and I respectively. These figures show the trade-offs comparison between method I and II in predicting high water level on high water level data only.

Figure 20 BEST Model Performance

The figure shows the prediction performance of BEST Model. It shows the upper and lower bound of the predicted intervals compare to the actual water level at C4.

Figure 21 BEST Model Performance (ZOOM) Predicted Interval VS Actual Water Level

This figure is a zoomed version of the figure 20, excluding predicting water level. It shows the coverage of the predicted intervals on the actual water level at C4.

Figure 22 BEST Model Prediction VS Actual Water Level

The figure compares the predicted water with the actual water level.

Figure 23 ALL Model Performance

The figure shows the prediction performance of ALL Model. It shows the upper and lower bound of the predicted intervals compare to the actual water level at C4.

Figure 24 ALL Model Performance (ZOOM) Predicted Interval VS Actual Water Level

This figure is a zoomed version of the figure 23, excluding predicting water level. It shows the coverage of the predicted intervals on the actual water level at C4.

Figure 25 ALL Model Prediction VS Actual Water Level

The figure compares the predicted water with the actual water level.

DAY	INTERVALS	[redacted]						METHOD II					
					HIGH WATER LEVEL PREDICTION						HIGH WATER LEVEL PREDICTION		
		AVG ERROR	AVG INTERVAL SIZE	SD	AVG ERROR	AVG INTERVAL SIZE	SD	AVG ERROR	AVG INTERVAL SIZE	SD	AVG ERROR	AVG INTERVAL SIZE	SD
DAY III	10	7.651	21.843	0.147	9.425	21.873	0.306	7.619	21.833	0.562	6.240	23.674	0.310
	15	9.884	18.502	0.831	10.523	20.851	0.974	8.736	19.009	0.066	7.514	22.834	0.707
	20	12.269	16.353	0.112	13.508	19.363	0.183	12.410	16.022	0.688	15.498	19.098	0.646
	25	14.210	14.689	0.760	19.421	15.244	0.807	15.234	14.366	0.164	16.545	18.737	0.491
	30	18.054	12.399	0.685	22.034	14.216	0.936	19.750	12.013	0.321	21.715	16.035	0.595
DAY II	10	7.308	21.395	0.421	9.136	21.469	0.042	7.195	21.304	0.318	5.487	23.131	0.209
	15	8.939	18.222	0.952	9.958	20.491	0.408	8.351	18.810	0.486	6.965	22.279	0.375
	20	11.979	15.650	0.170	13.420	19.239	0.889	12.023	15.671	0.488	14.531	18.400	0.979
	25	14.024	13.872	0.250	19.317	14.970	0.227	14.573	14.142	0.140	15.754	17.899	0.261
	30	17.067	11.569	0.414	21.731	14.068	0.133	18.985	11.186	0.581	21.032	15.960	0.693
DAY I	10	6.678	21.143	0.241	8.220	20.893	0.820	6.347	20.805	0.028	5.284	22.907	0.480
	15	8.029	17.992	0.038	9.689	19.492	0.232	8.048	17.985	0.755	6.464	21.390	0.084
	20	11.769	14.773	0.910	12.623	18.601	0.232	11.595	15.504	0.860	13.942	18.238	0.679
	25	13.492	13.165	0.808	19.304	14.667	0.139	13.589	13.324	0.892	15.341	17.196	0.028
	30	16.968	11.301	0.491	20.969	13.612	0.701	18.403	10.353	0.374	20.040	15.334	0.925

Table 6 Test Result for ALL Model

		[redacted]						METHOD II					
					HIGH WATER LEVEL PREDICTION						HIGH WATER LEVEL PREDICTION		
DAY	INTERVALS	AVG ERROR	AVG INTERVAL SIZE	SD	AVG ERROR	AVG INTERVAL SIZE	SD	AVG ERROR	AVG INTERV AL SIZE	SD	AVG ERROR	AVG INTERVAL SIZE	SD
DAY III	10	5.411	20.033	0.151	6.075	20.032	0.094	3.991	22.046	0.194	3.031	22.037	0.246
	15	6.641	17.068	0.093	8.059	19.094	0.088	7.340	17.017	0.139	7.361	20.013	0.111
	20	8.151	15.736	0.122	11.007	17.052	0.087	12.940	12.060	0.149	10.965	18.089	0.133
	25	10.240	14.029	0.110	16.035	15.008	0.089	16.221	9.552	0.117	16.289	16.017	0.191
	30	17.251	9.016	0.139	20.092	14.084	0.091	19.320	8.605	0.148	18.348	15.026	0.209
DAY II	10	4.495	19.940	0.155	6.027	20.000	0.132	3.426	21.968	0.100	2.953	21.948	0.146
	15	6.094	17.064	0.183	7.972	19.048	0.114	6.606	16.957	0.155	7.281	19.934	0.189
	20	7.191	15.657	0.147	10.910	16.973	0.117	12.896	11.984	0.183	10.963	18.035	0.175
	25	9.798	13.958	0.084	16.007	14.937	0.124	15.594	9.537	0.125	16.271	15.921	0.071
	30	16.906	8.917	0.144	20.040	14.082	0.122	18.638	8.524	0.165	18.290	14.968	0.189
DAY I	10	3.908	19.845	0.183	6.003	19.923	0.215	2.937	21.893	0.171	2.905	21.903	0.108
	15	5.494	16.971	0.130	7.903	18.998	0.205	5.671	16.886	0.193	7.241	19.880	0.169
	20	6.866	15.567	0.113	10.640	16.889	0.208	11.987	11.963	0.126	10.436	17.976	0.160
	25	9.703	13.918	0.160	15.440	14.841	0.114	15.504	9.513	0.139	16.192	15.877	0.171
	30	16.696	8.870	0.162	19.518	14.062	0.137	18.328	8.430	0.155	17.558	14.879	0.178

Table 7 Test Result for BEST Model

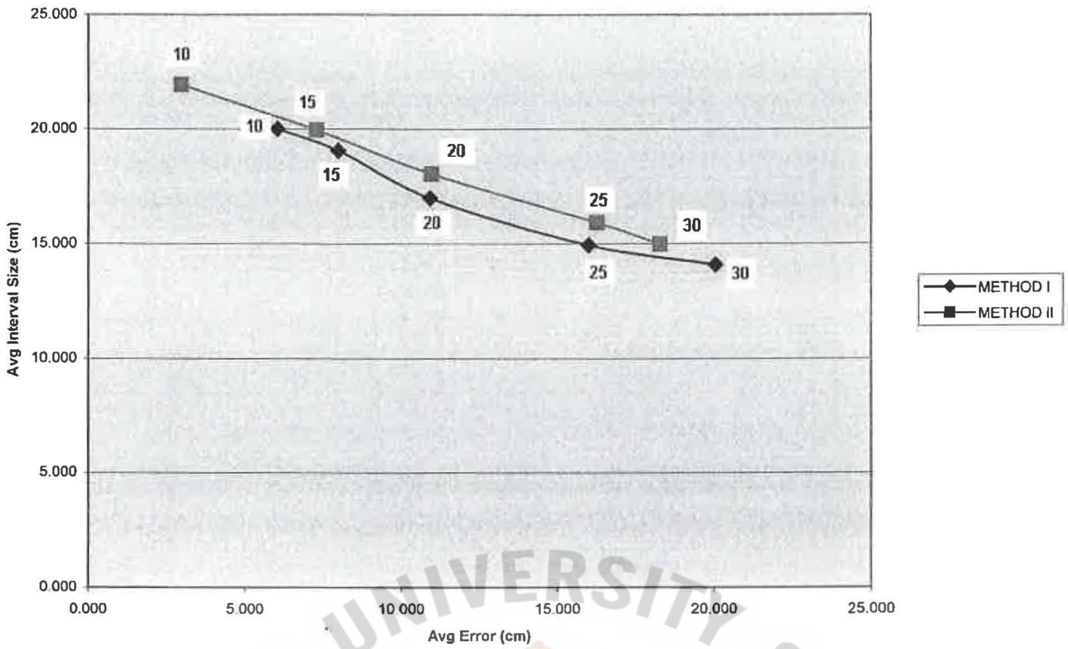


Figure 18 BEST Model Prediction During High Water Level Day II

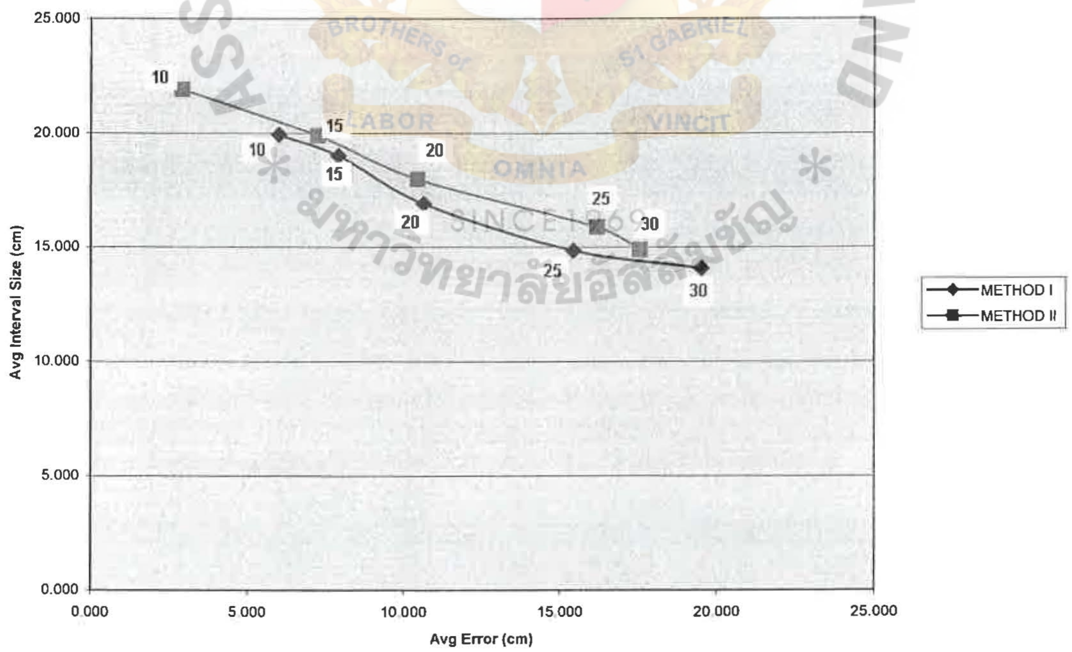


Figure 19 BEST Model Prediction During High Water Level Day I

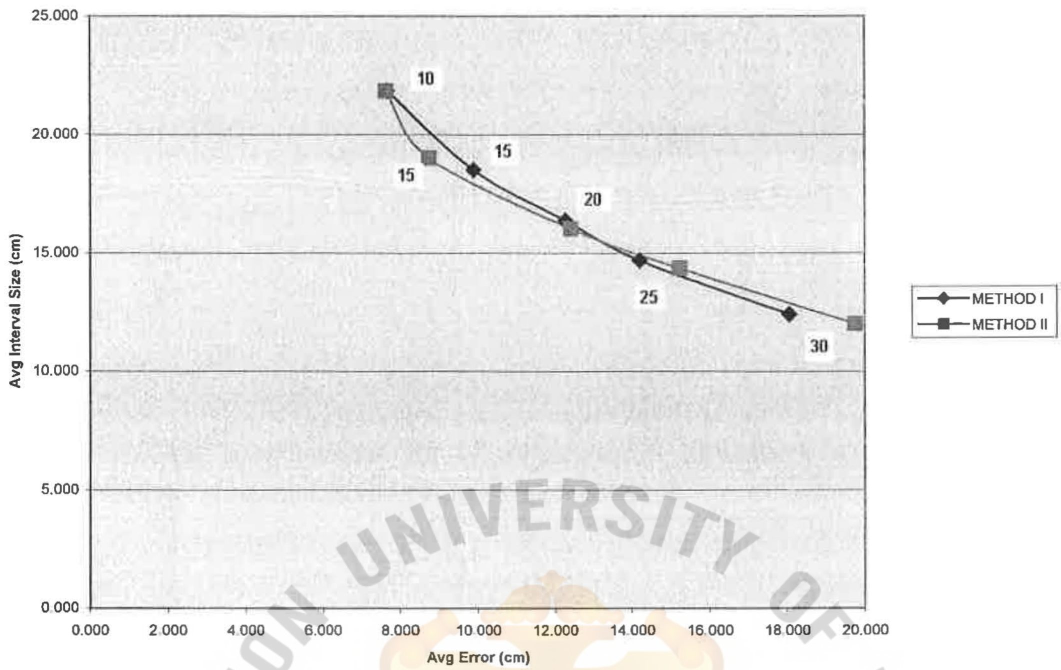


Figure 8 ALL Model Overall Performance Day III

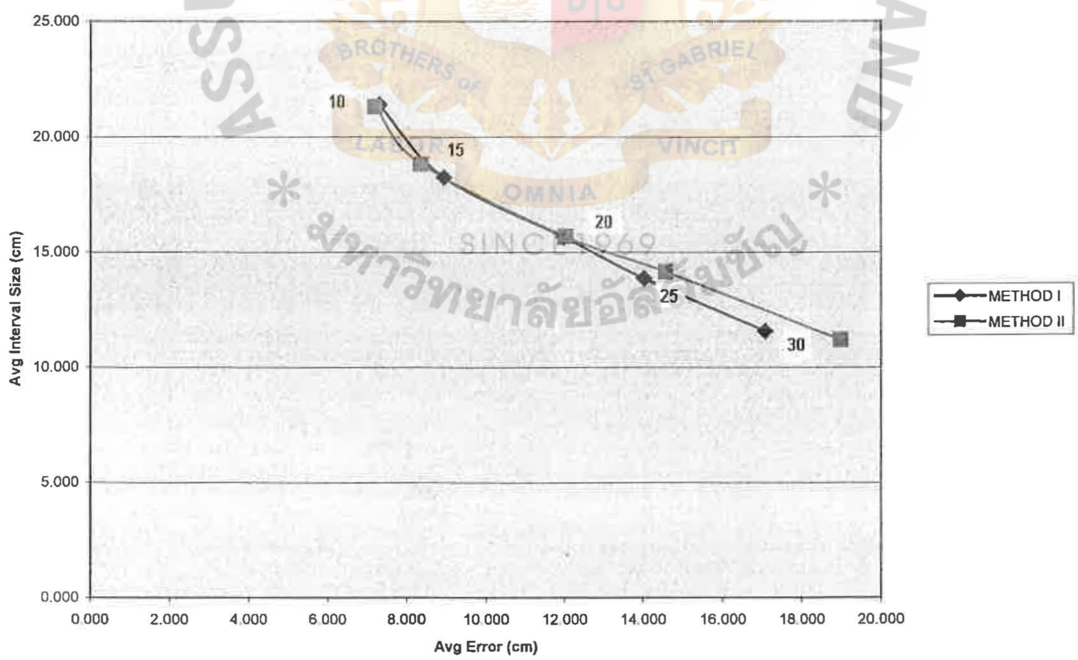


Figure 9 ALL Model Overall Performance Day II

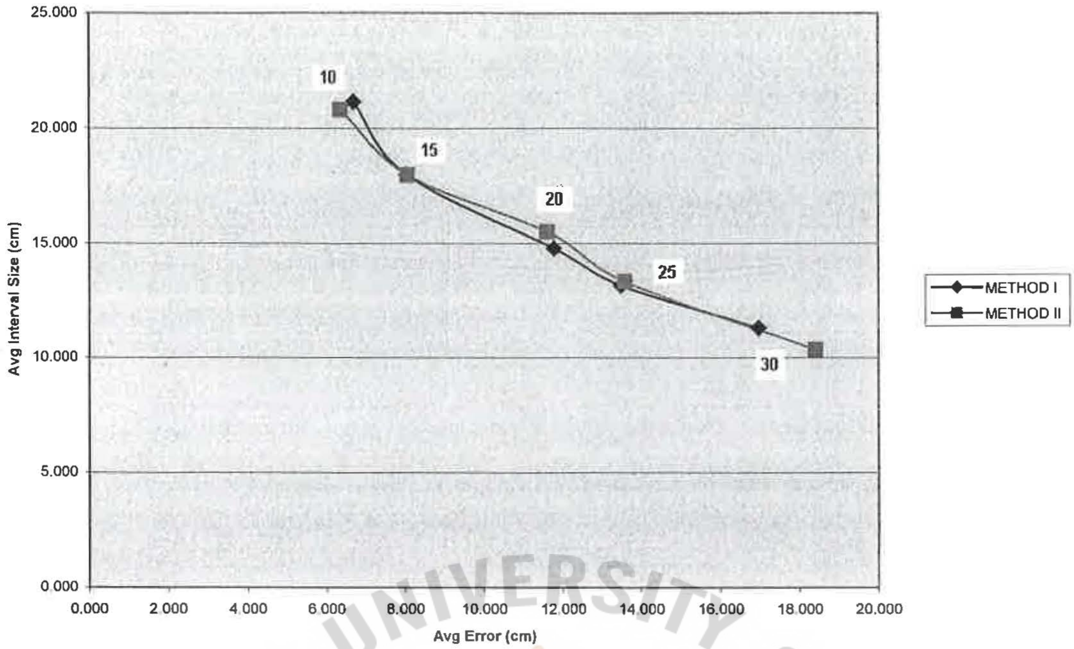


Figure 10 ALL Model Overall Performance Day I

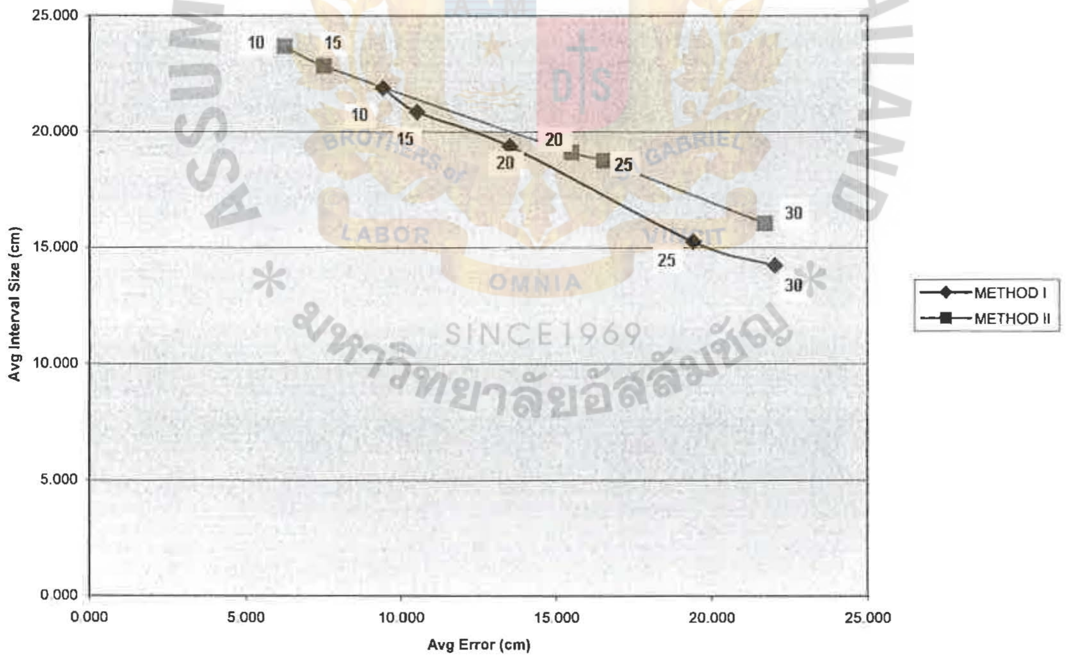


Figure 11 ALL Model Prediction During Flood Day III

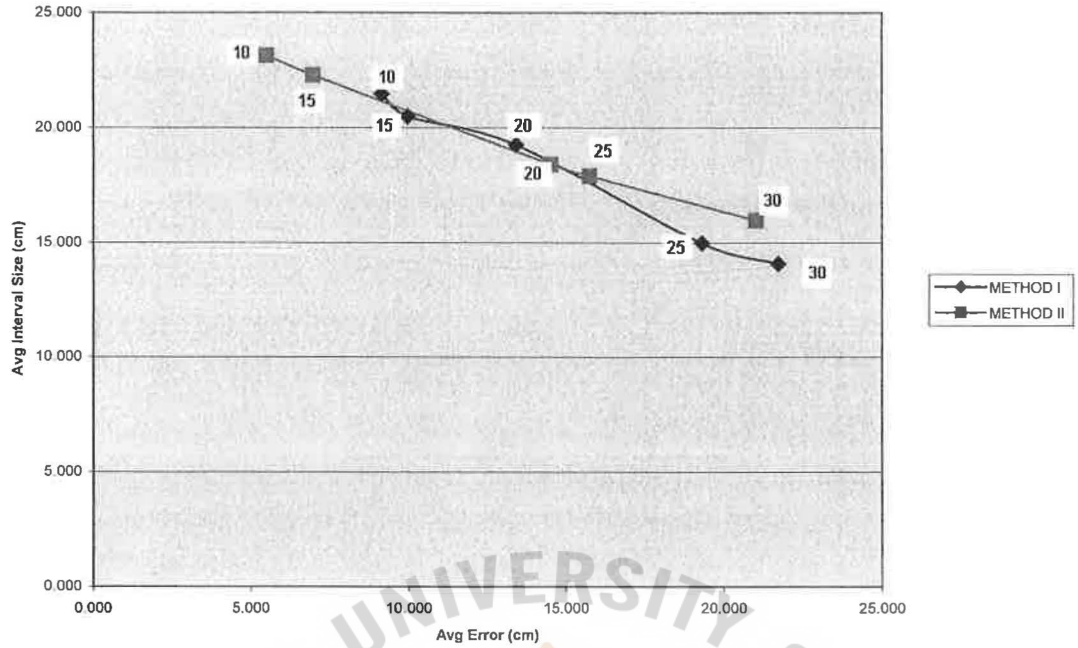


Figure 12 ALL Model Prediction During High Water Level Day II

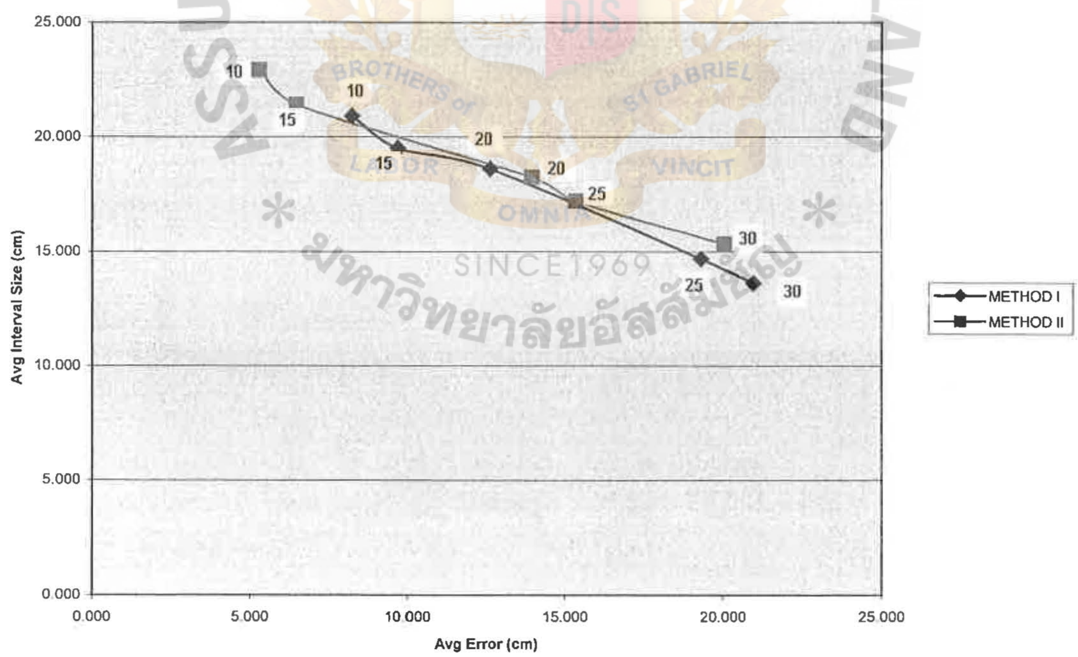


Figure 13 ALL Model Prediction During High Water Level Day I

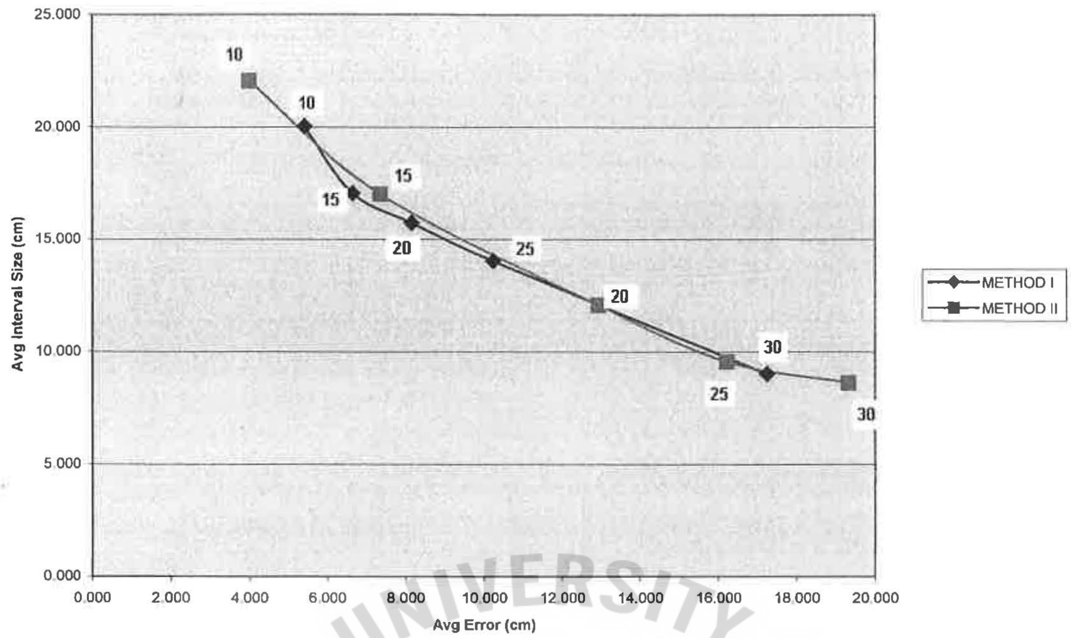


Figure 14 BEST Model Overall Performance Day III

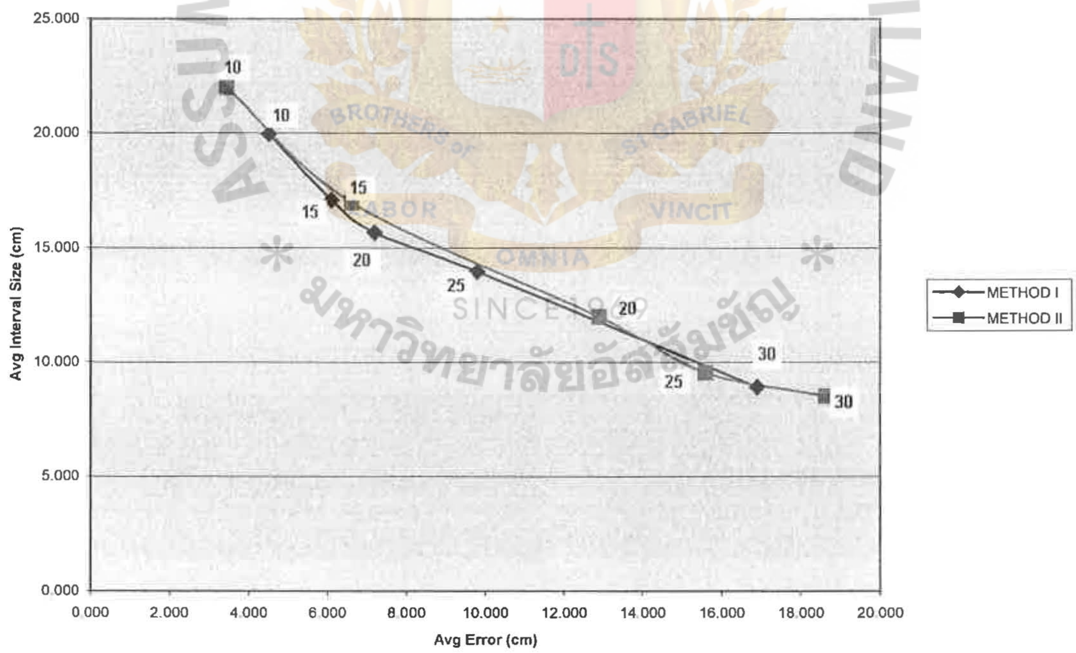


Figure 15 BEST Model Overall Performance Day II

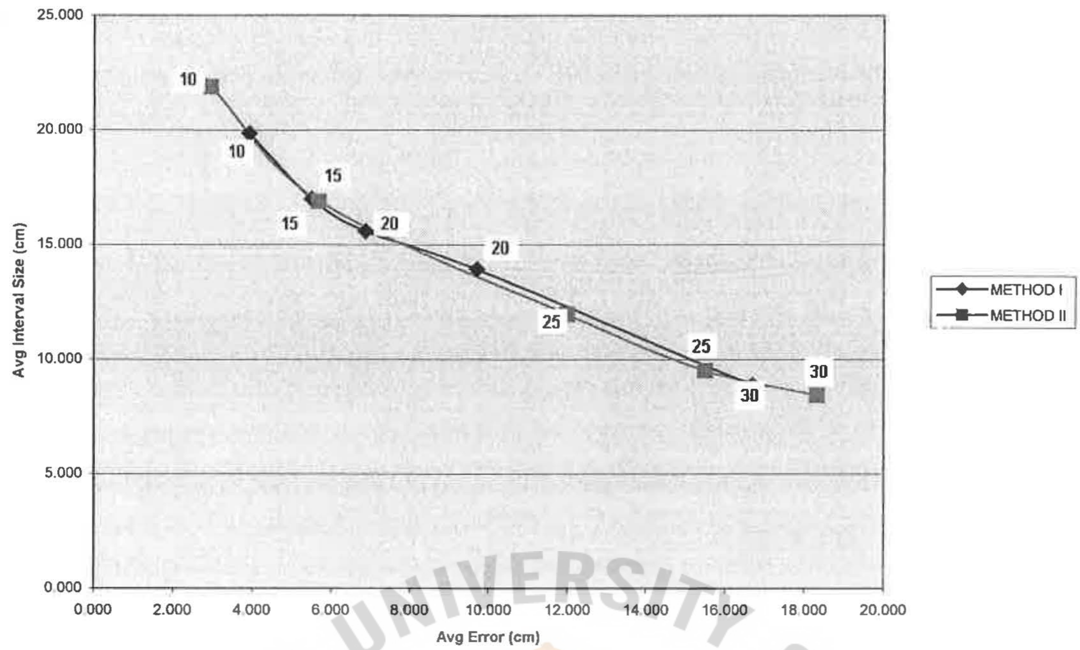


Figure 16 BEST Model Overall Performance Day I

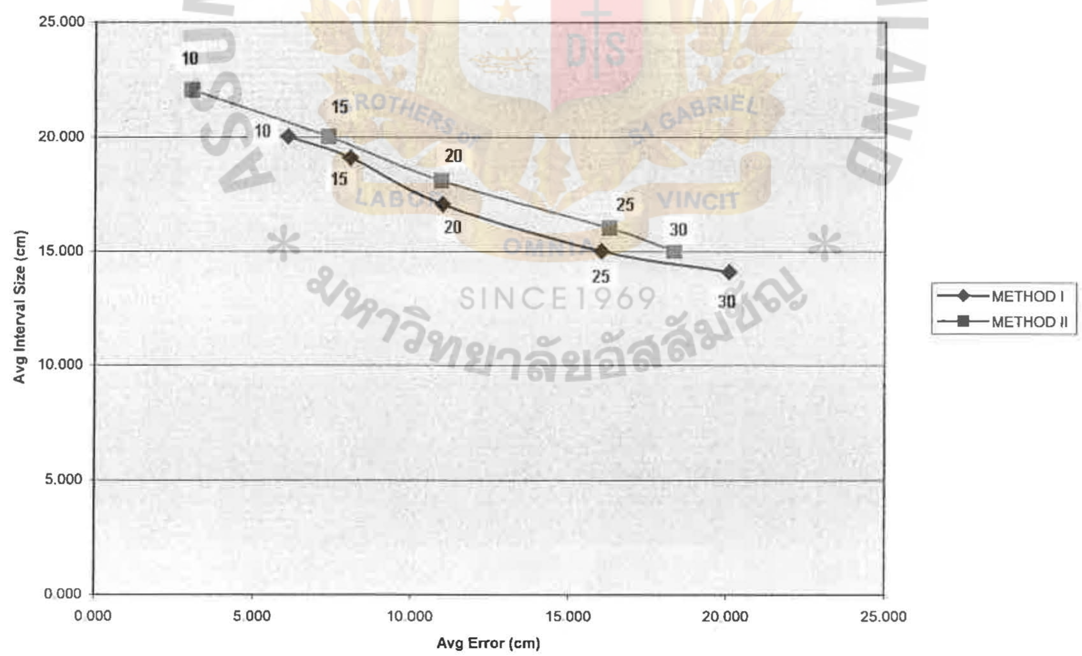


Figure 17 BEST Model Prediction During High Water Level Day III

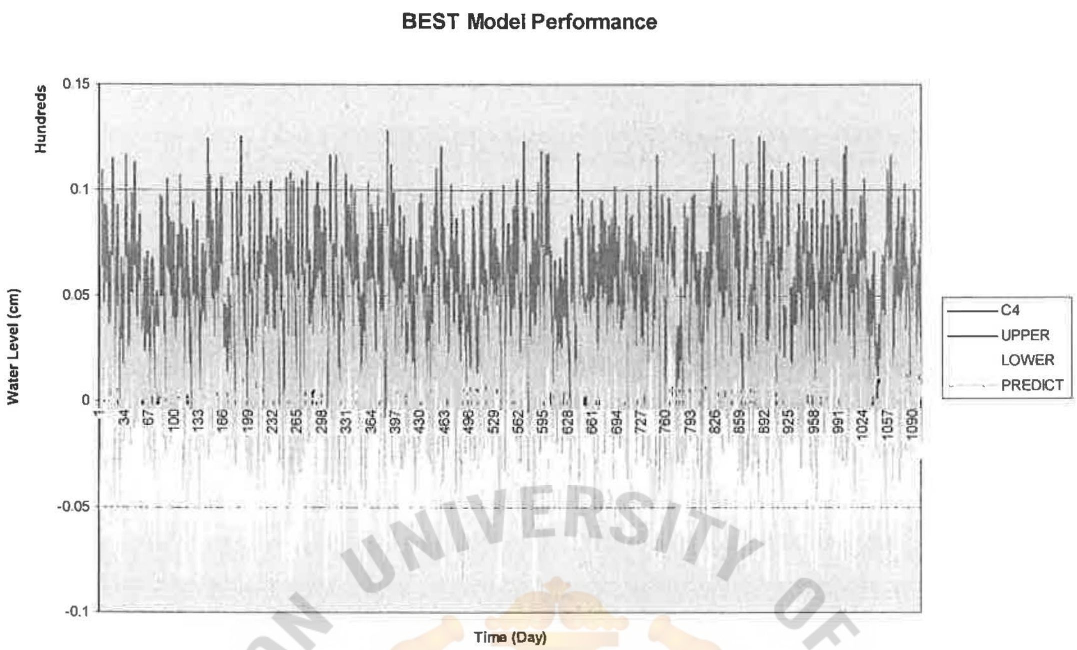


Figure 20 BEST Model Performance

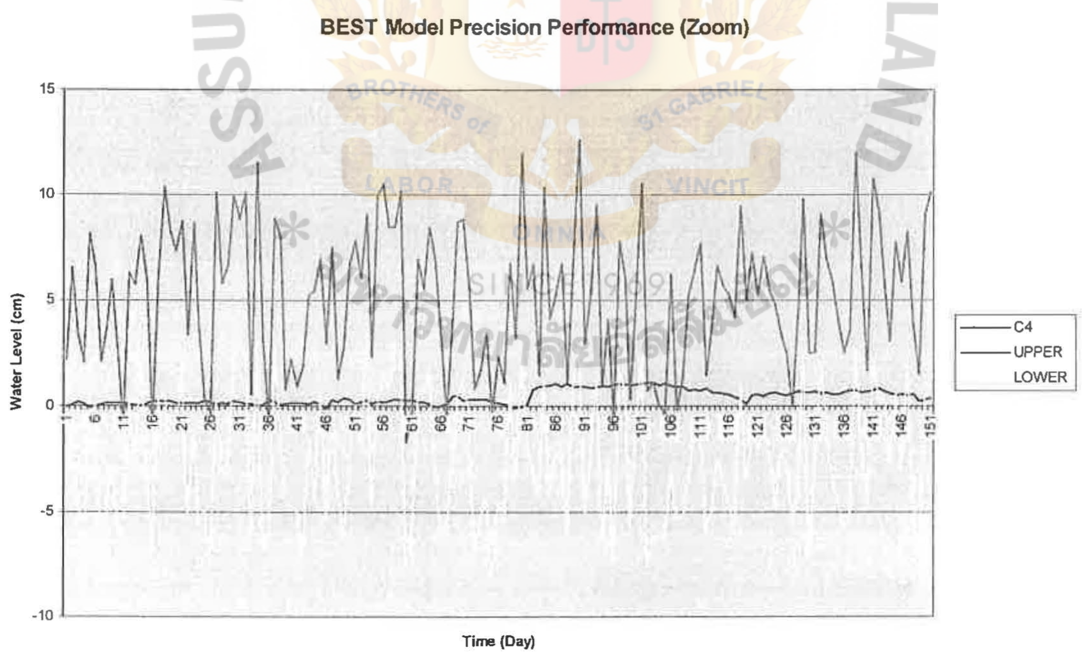


Figure 21 BEST Model Performance (ZOOM) Predicted Interval VS Actual Water Level

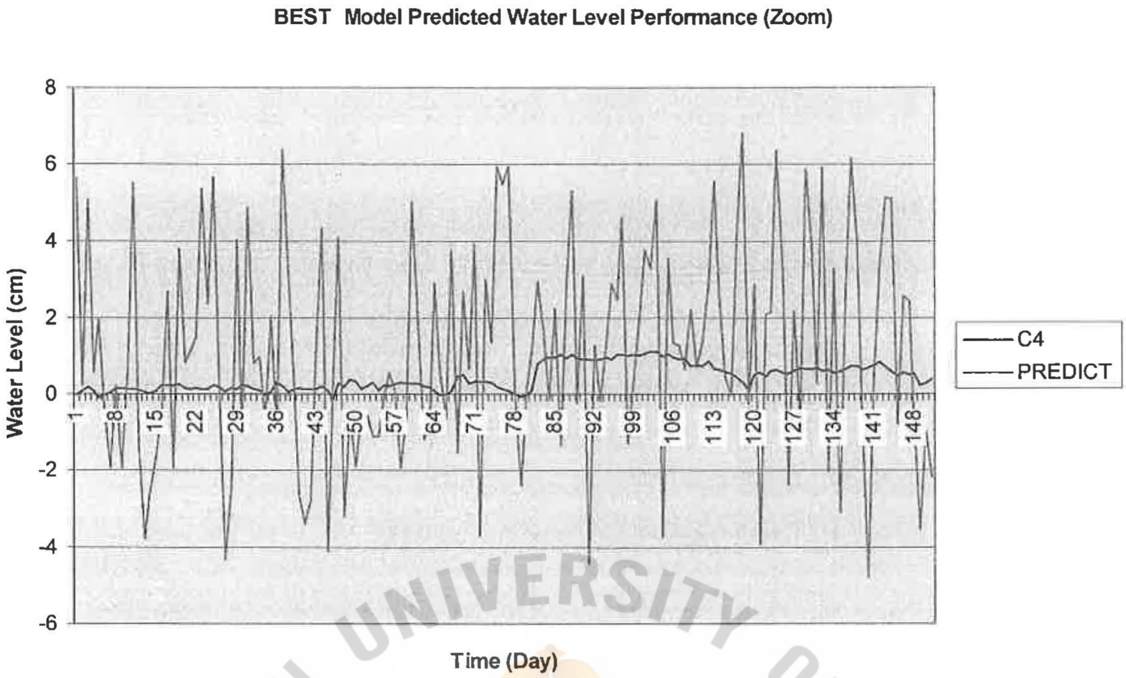


Figure 22 BEST Model Prediction VS Actual Water Level

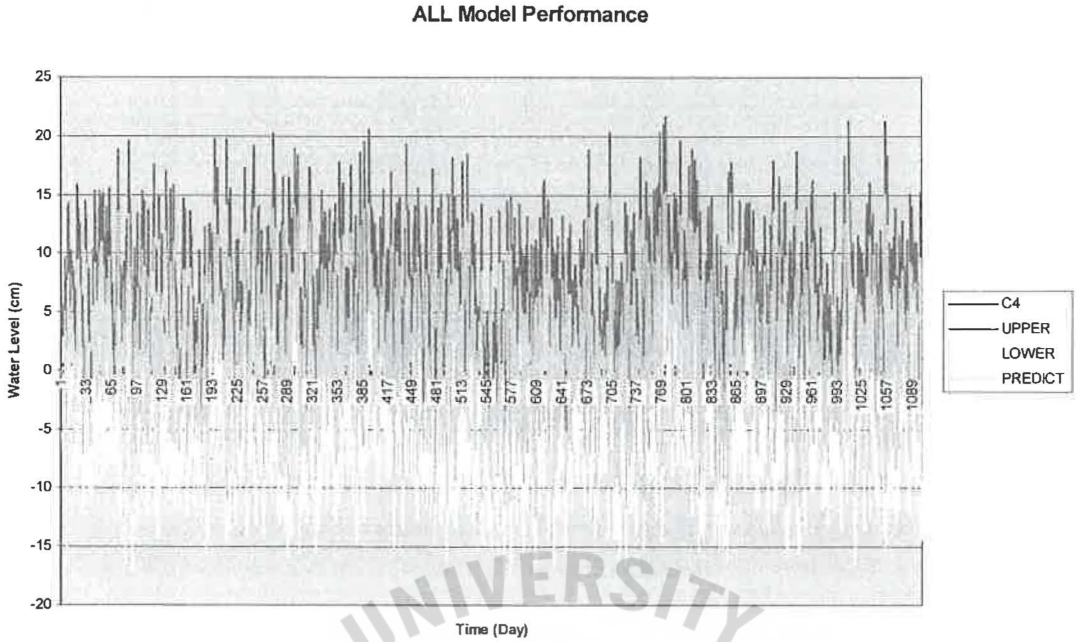


Figure 23 ALL Model Performance

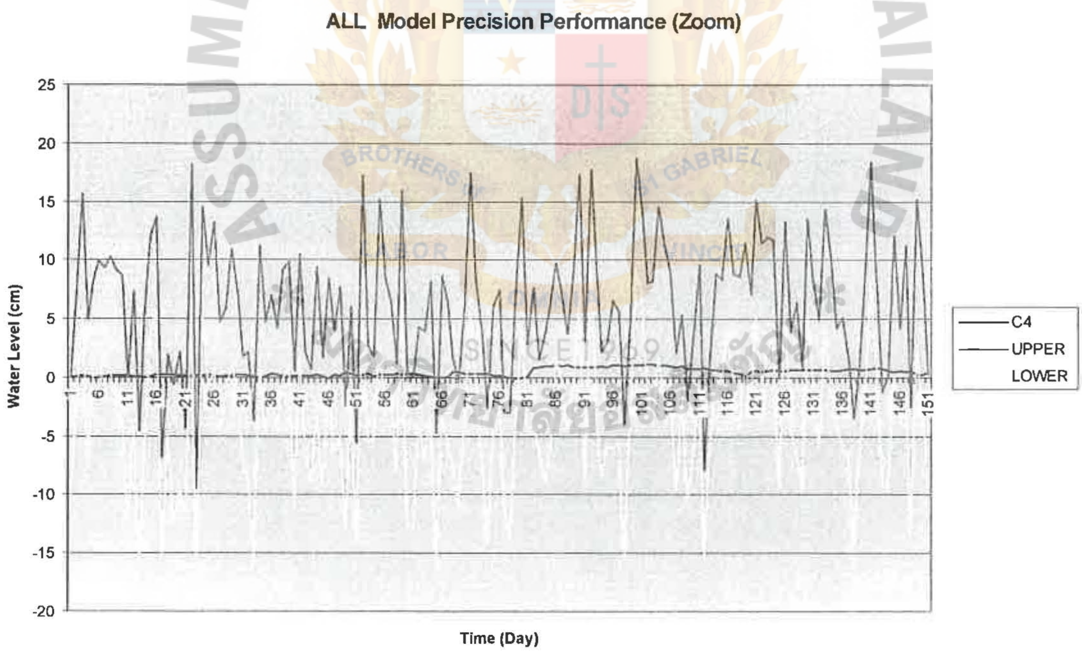


Figure 24 ALL Model Performance (ZOOM) Predicted interval VS Actual Water Level

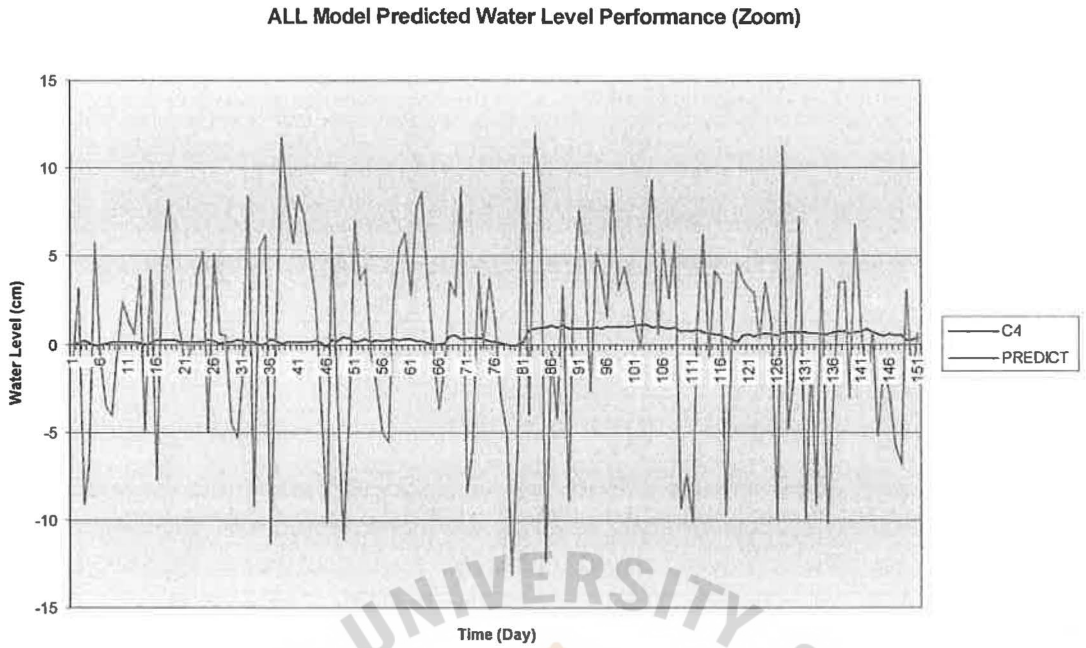


Figure 25 ALL Model Prediction VS Actual Water Level

5. EMPIRICAL ANALYSIS

We test the model in 2 interested datasets.

1. Test on overall data. Which this could tell the overall performance of the model in predicting the water level compare to the actual water level.
2. Test on flood situation data. The flood situation is considered by water level 1.5 meters at C4 station. In this case, we sort the data we have, and take the high water level, approximately 15% of the overall dataset to test as the high water level, or flood situation.

Figure 8, 9,and 10 show trade-offs between error and interval size, and also compares two methods trade-offs.

5.1 EVALUATION OF MODELS

In overall performance measurement, BEST model results better compare with the ALL model, as well as the flood situation prediction. The reason to this result could be that the prediction in BEST model involves only stations that have high relationships, or we may say, BEST model involves only the stations that influents the water level in C4. In another saying, it could be that, if we are interested in other station that is not C4, the stations involved in the model may not be only 4 nodes, but could be more or less. Involving many nodes into the model makes the prediction not good, compare ALL and BEST model result. It could be that the noise of data in low correlation coefficient interfere the propagation of the model, so the precision of the ALL model is wider than BEST model.

The prediction in the sequence of days becomes inaccurate for the later days. The precision drops in day 2 and 3, so it could be that the noise of the data could interfere the model. Since the model compute the probability of the water level in each status only, but it may not cover every case of the water level status. It could be unexpected case like, immediate storm encounters in Thai bay, high water level from the ocean directly increases water level in Bangkok.

5.2 EVALUATION OF DISCRETIZATION METHODS

As the result of the experiment, method I and II are giving similar performance. But during the flood situation, the prediction of model using method I is a little bit better than method II both precision and accuracy. The reason to this could be that the discretization in method II has wider interval size in low and high water level, so probability distribution that distribute to all the intervals, although low in low and high water level, affects the computation for expected water level and intervals. While the discretization method I has equal interval size, so the probability distribution distributes over these intervals does not affect much on computing the expected water level and intervals, compare to the wider intervals.

Here we can summarize that, 2 discretization methods give the similar performance, but in specific situation, if we are interested at very high and very low water level, the discretization of method II does not result so precise because of the wide range.

6. RELATED WORK

6.1 Introduction to Bayesian Network

Bayesian networks have become the most popular technique for representing and reasoning with probabilistic information. A Bayesian network [2], [3] is a directed acyclic graph that represents a probability distribution. Nodes represent random variables and arcs represent probabilistic correlation between the variables. The types of paths (and lack thereof) between variables indicates probabilistic independence. Quantitative probability information is specified in the form of conditional probability tables (CPT). For each node the table specifies the probability of each possible state of the node given each possible combination of states of its parents. The tables for root nodes just contain unconditional probabilities.

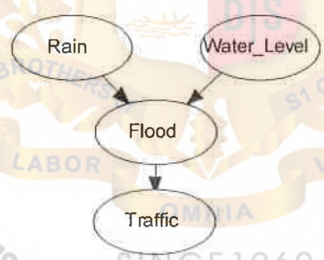


Figure 26 A simple Bayesian network

Figure 26 shows a simple Bayesian network representing the fact that local rainfall and river water level influence the degree of flooding in central Bangkok, with in turn influences the degree of traffic congestion. (For simplicity of the example, the many other factors influencing traffic have been left out.) The network is quantified by specifying two unconditional probabilities and two sets of conditional probabilities:

P(Rain),

P(Water_Level),

P(Flood | Rain, Water_Level),

P(Traffic | Flood).

The key feature of Bayesian networks is the fact that they provide a method for decomposing a probability distribution into a set of local distributions. The independence semantics associated with the network topology specifies how to combine these local distributions to obtain the complete joint probability distribution over all the random variables represented by the nodes in the network. In our example, the network topology encodes the assumption that Rain and Water_Level are probabilistically independent and that Traffic is independent of Rain and Water_Level given Flood. The joint probability distribution is

P(Rain, Water_Level, Flood, Traffic) =

P(Rain) * P(Water_Level) *

P(Flood | Rain,Water_Level) *

P(Traffic | Flood)

Bayesian networks have three important advantages.

1. First, naively specifying a joint probability distribution with a table requires a number of values exponential in the number of variables. In systems in which interactions among the random variables are sparse, Bayesian networks drastically reduce the number of values required.

2. Second, efficient inference algorithms exist that work by transmitting information between the local distributions rather than working with the full joint distribution.
3. Third, the separation of qualitative representation of the influences between variables from the numeric quantification of the strengths of the influences has a significant advantage for knowledge engineering.

In building a Bayesian network model, one can first focus on specifying the qualitative structure of the domain and then focus on quantifying the influences. When finished, one is guaranteed to have a complete specification of the joint probability distribution. Many commercial and free software packages exist for building and performing computations with Bayesian networks. For a list of these see [4].

For this section we discuss related works that have been done earlier and the temporal probabilistic modeling technique.

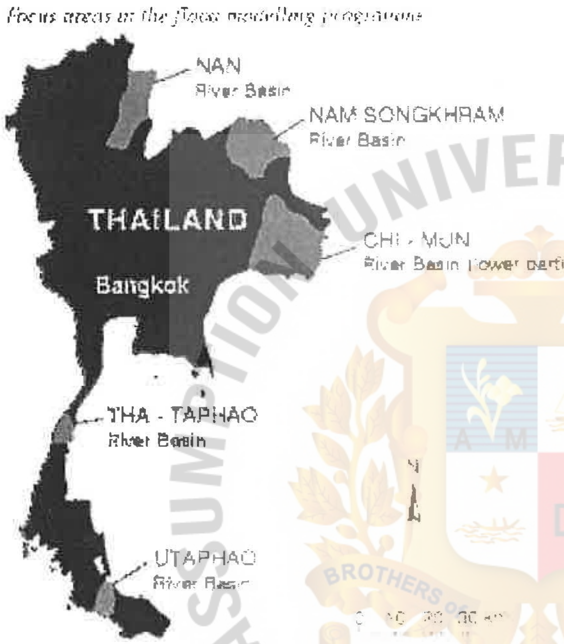
In first 2 related works, we take a look at how MIKE-11 is applied in Thailand. MIKE-11 is a product of Danish Hydraulic Institute (DHI). DHI is international provider of specialized constancy and software within coastal, marine and water resources engineering. DHI has solved complex problems for clients in over 120 countries within a wide range of water related fields. Thailand is one of the countries that DHI has involved.

6.2 Flood Modeling Program, Thailand (1993-96)

Location: Five river basins in Thailand

Type of Project: Transfer of MIKE-11 modeling technology

Client: Cooperation project between six government institutions in Thailand and DHI/Asian Institute of Technology (AIT). Sponsored by DANIDA [2]



Description During this project, the MIKE-11 modeling system has been set up in five river basins in Thailand for the purpose of planning, design and operation of flood control measures as well as real-time flood forecasting. The project is carried out by DHI and AIT in cooperation with six government institutions in Thailand (DOLA, DEPD, DPW, EGAT, MED and RID).

The objective is to enhance the capabilities of the institutions to plan, design and operate flood mitigation and preparedness programs by introducing effective flood modeling tools for river basins. The project involves the transfer of computer hardware and software and comprehensive training of 18 government officers from the participating institutions.

The project activities are centered around the selected five river basins in Thailand and serve the dual purpose of training and establishment of operational facilities for real-time flood forecasting as well as for the planning, design and operation of flood control measures for high-priority areas. The models for Chi-Mun, Nan and Utaphao river basins, where telemetric systems have been established, are now being operated for real-time flood forecasting on a routine basis.

6.3 Flood Modelling In The Chi-Mun River Basin, Thailand (1993-96)

Location: Chi-Mun River basin in northeastern Thailand

Type of Project: Transfer of MIKE-11 modelling technology and training

Client: A collaboration between six government institutions in Thailand and DHI/AIT. Sponsored by DANIDA [3]

Description During the 3-year flood modeling program in Thailand a comprehensive MIKE-11 model has been set up for the lower part of the large Chi-Mun River basin. In 1994 the Pak Mun dam, close to the confluence between the Mun and Mekong Rivers, was completed and the Electricity Generating Authority of Thailand (EGAT) needed a model for inflow forecasting and operation of the dam in order to ensure that strict operation criteria were fulfilled requiring water levels just upstream the new reservoir not to exceed a certain limit.

The hydraulic conditions in the lower Mun River are very complicated due to a number of rapids (small water falls). Due to the flexibility of MIKE-11 it has, however, been possible to accommodate for these, and the modeling results are excellent. The model has been operated in real-time by EGAT at the dam site since July 1996 using real-time data on rainfall and water levels from a newly established automatic telemetric system in the 13,000 km² model area upstream the dam.

The forecasted inflow hydrographs are very reliable (within +/-10% of the subsequently observed ones) and the built-in structure operation module provides useful advance guidance on how to operate the spillway gates during flood periods to fulfil the water level criteria upstream.

6.4 MIKE-11 Specification

MIKE-11 [9] is a hydrology software package for the simulation of flows, water quality and sediment transport in estuaries, rivers, irrigation systems, channels and other water bodies. MIKE-11 is a fully dynamic tool for the detailed analysis, design, management and operation of both simple and complicate rivers and channel systems.

The hydrodynamic (HD) module is the core engine of the MIKE-11 modeling system and forms the basis for most modules including Flood Forecasting. In order to predict water level, the hydrodynamic module takes several input parameters and most of them have to be collected from the environment of that area.

The example parameters for hydrodynamic module are shown as follow,

- Hydrodynamic Module Parameter
- Initial condition
  - The initial water level and discharge at the start time of the computation.
- Wind
  - The wind field boundary condition consists of specification for wind direction and the wind velocity.
- Groundwater Flow-rate
- Wave Approximation
  - To specify which wave approximation should be used in the computation, viz Kinematic, Diffusive or one of two fully dynamic wave approximations.
- Water Loss
- Parameters for the water loss are:

Smax(mm) :Capacity of retention storage.

INF(mm/hr) :Infiltration capacity of flood plains.

K(hr) :Time constant for return flow,

Wmin(m) :Minimum river width for water loss.

- Dispersion
  - The dispersion can be specified as a function of the flow velocity calculated by the following expression:

$$D = F * V^{ex} \quad ,$$

Where:

D is the dispersion coefficient (m^2/s)

V is the flow velocity (m/s)

F is the dispersion factor

ex is a dimensionless exponent.

- Decay

The hydrological modeling, normally, gives accurate results. However, most of the parameters are the data that we have to collect from the environment of the predicted area. This makes the difficulty in gathering data for predicting the water level since we need some special equipment to collect the data.

6.4 Discretizing Continuous Attributes While Learning Bayesian Networks

N. Friedman and M. Goldszmidt [7] present a method for learning Bayesian network that handles the discretization of continuous variables as an integral part of the learning process. They formally derive a criterion based on the *Minimal Description Length* principle for choosing the threshold values for the discretization. This new metric embodies a tradeoff between the complexity of the learned discretization, the complexity of the Bayesian network, and the fitness of the network as a model of the training data. The metric has the attractive property of *decomposition*: the discretization of each variable depends only on the interactions between the variable and its local neighborhood in the network. They examine other properties of this metric that are relevant to the computation of a discretization policy and propose an iterative algorithm for learning a policy.

The work discussed about the interaction of nearby nodes, which only some interested information will be interacted. The iterative algorithm for learning policy could be applied to the learning policy in this thesis problem area.

6.5 Graphical Decision Models, Planning and Control

Thomas L. Dean and Michael P. Wellman [8] discuss about the Markov process and the graphical model that is used in planning and control. The Markov model is a temporal model that each state is influenced by the previous state. The Markov model can be in multiple orders. In this thesis, we design model as a Markov model order one. The reason we do not involve multiple orders into the model design, which we could describe the water level in current day with the 1 or 2 or even 3 previous days. But involving multiple orders cause the conditional probability table to be much more larger. Example, one node is discretize to 10 states, Markov order 2. The consequence time slice, the node must have CPT of joint between current day, previous 2 days, which the CPT size is 10^3. Together with constrain nodes and influence from other stations, the CPT size would increase to be $10^3*10*10$, which there will not be enough data to fill in the CPT. The probability for each cell in CPT would be very tiny, that will not have much significant in prediction.

But the multiple orders Markov chain model is an interesting idea to try. When we have more data, we would try in multiple orders.

6.6 Prediction of the Water Level during Storm Situations using Neural Networks

M.C. van de Wag [1] develop Neural network framework for the flood prediction in river in Hoak van riven, Holland. He compared the 2 training techniques, radical basis training with training with a feed-forward back propagation network, which the multi-layer feed-forward network gives better result.

This is an attempt to develop flood modeling using other technique, which the development approach is similar to this thesis, which is the work based on an assumption to construct an initial model, and then what he was interested in is the methodology in learning the network.

7. CONCLUSION AND FUTURE RESEARCH

7.1 THEORICAL ISSUE

This thesis is an attempt to improve the flood prediction done with Bayesian network technique. From our preliminary work that has been done before, we improve techniques and implementations. Many points that we pinned down ourselves weak points form the preliminary work, and prove that we have improve the work, although the result is not that impressive. The model needs to be calibrated, but due to the limitation of the data we have. But from the limitation, we still can discover some things, for example, one single type of data input is possible to predict the same type of data quite well. But the prediction in open environment, like a river, needs more inputs as factors. MIKE-11 is an example of hydrological modeling software. It takes a lot of inputs to generate the best output. We need to balance the ease of use and the result output, and also the precision and accuracy are invert.

7.2 IMPLEMENTATION ISSUE

The flood prediction system was developed by

Programming tool: Microsoft Visual Basic, SQL, PL/SQL

Database tool: Microsoft Access, ORACLE 8i

Statistical tool: Microsoft Excel, SPSS 8.0

Model implementation and computation: HUGIN Professional 5.4

The software is to be re-done if there is any change in the modeling. But updating the probability to the model does not need to re-build the software, simply update the database and re-run the CPT computation again.

7.3 FUTURE RESEARCH

We are planning to improve Bayesian network prediction that deal witContinuous variables by using Conditional Gaussian variable. The feature is also available in HUGIN Professional 5.4, and also accessible through the HUGIN API.

We have discussed through e-mail with well-known researchers, for this discussion, we are really appreciate for help from Dr.Peter Haddawy.

Robert Dodier suggested an alternative to deal with continuous variables.

"Discretize all the continuous variables

Conditional Gaussian (CG) variables. A key feature of this scheme is that the Bayesian network can be represented using original distributions from the problem domain, and at least some of the results will also have some ordinary parametric form, rather than being uninterpretable approximations."

Discretize all the continuous variables, of course that, the information must be lost some where during the interval chopping.

And, Lars M Nielsen also suggested,

"Conditional Gaussian (CG) variables have a normal distribution for each configuration of their discrete parents where the means depend linearly on the CG parents.

If you want any other kind of continuous variables, you need to discretize them (and lose accuracy). I don't know exactly how much you lose if you discretize. If you use many narrow intervals you can of cause get this down to a minimum."

From his sentence, his idea also supports us to narrow the intervals to a minimum, as what we have done in this thesis.

Alexander V. Kozlov suggested the algorithms for the continuous value,

"He was trying to minimize the KL distance of the answer to a query given BN on discrete and continuous variables. It can be done by propagating information back and forth thorughout a network (in a fashion very similar to LS algorithm).

The initial work was published in UAI-97"

From these comments from these researchers, we need to learn more in other techniques and apply our problem, so the flood prediction could be better.

8. BIBLOGRAPHY

[1] M.C. Van de Wag 1997, *Prediction of the water level during storm situations using Neural Networks*, Department of Computer Science, Lieden University, 1997.

[2] *DHI, Flood Modeling Program, Thailand (1993-1996)*, [Online], 1996, Danish Hydraulic Institute, http://www.dhi.dk/project/thailand/ [November 10, 1999].

[3] *DHI, Flood modeling in the Chi-mun River Basin, Thailand (1993-1996)*, [Online], 1996, Danish Hydraulic Institute, Place of Publication, http://www.dhi.dk/project/thailand/ [November 10, 1999].

[4] Jensen, 1996 F.V. Jensen. *An introduction to Bayesian Networks*, Springer, 1996.

[5] Haddawy 1999 *An Overview of Some Recent Developments in Bayesian Problem Solving Techniques*, AI Magazine, Special Issue on Uncertainty in AI, Summer 1999.

[6] Pearl, 1988 J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988. (Revised second printing 1991.)

[7] N. Friedman and M. Goldszmidt, *Discretizing Continuous Attributes While Learning Bayesian Networks* , Thirteenth Internation Conference on Machine Learning.

[8] Thomas L. Dean, Michael P. Wellman, *Planning and Control*, pg 292-338.

[9] *MIKE11 Specification*, [Online], 2001, Danish Hydraulic Institute,

http://www.dhisoftware.com/mike11/index.htm [January 10, 2001].