Statistical Wave Forecasting through Kalman Filtering Combined with Principal Component Analysis

Noriaki Hashimoto[1], Toshihiko Nagai[2] and Masanobu Kudaka[3]

Abstract

Statistical wave forecasting methods have been applied because of their convenience. Most of them, however, include some drawbacks from the statistical or numerical viewpoints. In this paper, these drawbacks are discussed and a new statistical wave forecasting method utilizing the Kalman filter technique combined with Principal Component Analysis (PCA) is proposed in order to mitigate the drawbacks. The applicability and reliability of the proposed method is examined for five wave observation stations around Japan through simulations based on 5-years of wave data and weather charts.

Introduction

Adequate wave forecasting is indispensable for the safe operations of cargo handling, optimum management of port construction projects, and navigating and/or mooring vessels. There are two kinds of wave forecasting methods. One is a numerical model describing the physical process between winds and waves. The other is an empirical model based on a statistical relationship between the weather and the wave data obtained in the past.

The former method has been widely used for wave hindcasting for estimating design wave conditions. The reliability of the method has been discussed in several papers so far. Practical computation with this method, however, requires special knowledge of both the atmospheric and wave systems. Also, a large investment in computations is sometimes required.

On the other hand, the latter method commonly utilizes simple statistical relationships among criterion variables and predictor variables using obtained data.

---

[1] Chief, Hydrodynamics Laboratory, Marine Environment Division, Port and Harbour Research Institute, Ministry of Transport, 1-1 Nagase 3-chome, Yokosuka 239-0826, Japan

[2] Chief, Marine Observation Laboratory, Hydraulic Engineering Division, ditto.

[3] Chief, Wave Information Laboratory, ECOH Co., Ltd., 2-6-4 Kitaueno, Taito-ku, Tokyo 110-0014, Japan.

The statistical method has advantages in that it is easy to handle and does not require special knowledge in practical computations. Because of these advantages, several statistical models have been proposed so far. The most commonly used model is the multiple regression model in which the wave characteristics, such as significant wave height and period at a specific point, and the atmospheric pressure data and/or wind data at several points are interrelated. However, from numerical viewpoints, most of the existing models contain several drawbacks such as multicollinearity among predictor variables, over-fitting of the criterion variable to the data, and over adoption of predictor variables in the model.

For these reasons, it necessitates the development of a reasonable statistical model in which the dynamical behavior of each variable and the statistical relation among variables are properly taken into consideration to eliminate the above drawbacks. Before applying the model, it is necessary to examine the model in detail for the real data obtained in various sea conditions for practical applications.

In this paper, we propose a new statistical wave forecasting model utilizing the Kalman filter technique combined with Principal Component Analysis (PCA) in order to mitigate the above-mentioned drawbacks of the conventional statistical wave forecasting models.

Drawbacks of the conventional statistical wave forecasting methods

In the normal procedure of wave forecasting by using a numerical model, first, the wind field is calculated from the weather charts. Once the wind field is calculated, the generation, development and attenuation of the wave field can be calculated from the wind data. Usually, in numerical computations, a proper grid size is adopted to obtain accurate and reliable results. However, if the same grid size is applied to the statistical wave model, and the data for the predictor variables are given on the same grid points, such fine grid size sometimes cause inferior prediction. This is because of the high correlation among predictor variables themselves, which causes the multicollinearity in the correlation matrix of the predictor variables. In such situations, the data on such fine grids are no longer proper predictor variables. However, if a rough grid size is adopted for predictor variables to eliminate the above multicollinearity problem, important small scale atmospheric pressure patterns will be overlooked and the accuracy of the wave forecasting will be reduced. This is one of the problems for the conventional statistical wave forecasting method.

When analyzing the time series of atmospheric pressure data, the spectrum of the atmospheric pressure includes energies in a considerably wide range of frequencies including seasonal and yearly changes. If the atmospheric pressure data are directly applied for establishing the equations for short-term wave forecasting, unnecessary long-term components of the energy included in the data may cause biases in the relation between the criterion variable and the predictor variables. This leads to the deterioration of the prediction accuracy.

Over-fitting of the criterion variable to the data and over-adoption of predictor variables in the model are also common matters to be attended in the statistical model, which sometimes cause the prominent delay of the predicted values to the real values, and the predicted values tend to be more unstable.

In order to mitigate the above-mentioned drawbacks of the conventional statistical wave forecasting models, this new statistical wave forecasting model is developed which properly considers the dynamical behavior of each variable and the statistical relation among.

Definition of the "wave forecast" used in this study

First, to eliminate any possible confusion, the meaning of the word "forecast" used in this study should to be clarified. The word "forecast" is generally used to estimate an unknown situation based on the information obtained at present or in the past. However, in this study, we assume that the accurate weather information at the time for wave forecasting has already been known, which is assumed to be predicted by some other methods. That is, when we try to forecast waves 24 hours ahead, the accurate weather chart 24 hours ahead has already been obtained and wave data observed at the time are also available. Though the accuracy of the wave forecast strongly depends on the accuracy of the weather information, we examine only the accuracy of waves forecasted by the proposed method under the condition that the accurate weather information is given. Though the forecasted weather chart may include some errors, it is beyond our research to examine the accuracy of the weather chart.

Wave data and atmospheric pressure data used in this study

The computation for wave forecasting is carried out using the atmospheric pressure data read every 12 hours at the grid points of 500km×500km shown in **Figure 1**. Five years of atmospheric pressure data, from 1980 to 1984, are used in this study. The locations of the wave forecasting points are denoted by the upper case letters in **Figure 1**. At each location, wave observations have been obtained every 2 hours for many years. Though the grid size is very rough compared to the numerical model, the forecasted wave heights using this grid size is acceptable for practical applications as shown later.

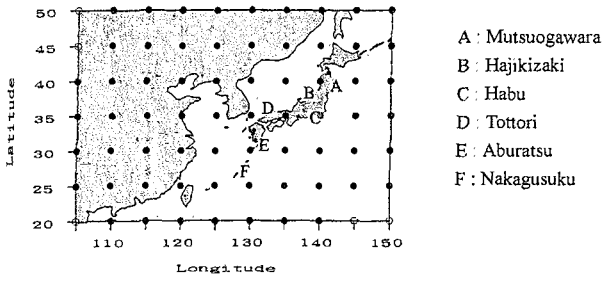

Figure 1 Wave observation stations and grid points of atmospheric pressure data

Statistical wave forecasting model utilizing Kalman filter combined with Principal Component Analysis (PCA)

Figure 2 shows the flow chart of the procedure of wave forecasting by using the Kalman filter combined with Principal Component Analysis (PCA). The model employed here consists of three parts. The first is the real-time filtering by using the Kalman filter. The second step is the PCA. The final step is the wave forecasting by the time-dependent regression model utilizing the Kalman filter. By obtaining new wave information and atmospheric pressure data, these three steps are repeated to update the wave forecasting equation and improve the accuracy of the wave forecasting. The procedure of each step is introduced in detail in the following.

1) Real-time filtering of the atmospheric pressure data by the Kalman filter

The first step in **Figure 2** is the real-time filtering by the Kalman filter. Using their technique, the time series of atmospheric pressure data on the grid points are separated into two components, i.e., a long-term component (longer than one-week period) and the remaining short-term component. The outline of the Kalman filter is as follows.

The equations of the state space representation are expressed by:

$$\mathbf{x}_n = \mathbf{F}_n \mathbf{x}_{n-1} + \mathbf{G}_n \mathbf{v}_n \quad \text{(System Equation)} \tag{1}$$

$$\mathbf{y}_n = \mathbf{H}_n \mathbf{x}_n + \mathbf{w}_n \quad \text{(Observation Equation)} \tag{2}$$

where $\mathbf{x}_n$: state vector ($k \times 1$), $\mathbf{v}_n$: system noise (Gaussian white noise with mean vector O and covariance matrix $\mathbf{Q}_n$) ($m \times 1$), $\mathbf{y}_n$: observation vector ($l \times 1$), $\mathbf{w}_n$: observation noise (Gaussian white noise with mean vector O and covariance matrix $\mathbf{R}_n$) ($l \times 1$) and $\mathbf{F}_n$, $\mathbf{G}_n$, $\mathbf{H}_n$: respectively ($k \times k$), ($k \times m$) and ($l \times k$) matrix.

To estimate the state vector $\mathbf{x}_n$ in Equation (1) on the basis of the observation vector $\mathbf{y}_n$ in Equation (2), the following one-step prediction and filtering are recursively applied through Equations (3) - (7).

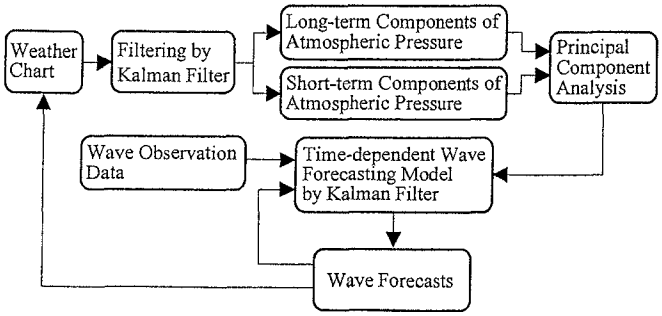

Figure 2 Flow chart of the proposed wave forecasting procedure

[One-step prediction]

$$\mathbf{x}_{n|n-1} = \mathbf{F}_n \mathbf{x}_{n-1|n-1} \tag{3}$$

$$\mathbf{V}_{n|n-1} = \mathbf{F}_n \mathbf{V}_{n-1|n-1} \mathbf{F}_n^t + \mathbf{G}_n \mathbf{Q}_n \mathbf{G}_n^t \tag{4}$$

[Filtering]

$$\mathbf{K}_n = \mathbf{V}_{n|n-1} \mathbf{H}_n^t (\mathbf{H}_n \mathbf{V}_{n|n-1} \mathbf{H}_n^t + \mathbf{R}_n)^{-1} \quad : \text{Kalman gain} \tag{5}$$

$$\mathbf{x}_{n|n} = \mathbf{x}_{n|n-1} + \mathbf{K}_n (\mathbf{y}_n - \mathbf{H}_n \mathbf{x}_{n|n-1}) \tag{6}$$

$$\mathbf{V}_{n|n} = (\mathbf{I} - \mathbf{K}_n \mathbf{H}_n) \mathbf{V}_{n|n-1} \tag{7}$$

where $\mathbf{x}_{n|j} = E(\mathbf{x}_n \mid \mathbf{Y}_j)$ and $\mathbf{V}_{n|j} = E(\mathbf{x}_n - \mathbf{x}_{n|j})(\mathbf{x}_n - \mathbf{x}_{n|j})^t$

For separating long-term and short-term components of the atmospheric pressure, equation (8) is assumed as the observation equation of equation (2).

$$y_n = t_n + w_n \quad \text{(Observation equation)} \tag{8}$$

where y_n : the observed value, t_n : the long-term component and w_n : the short-term component.

Since the long-term component is a slowly varying value, Equation (9) is assumed as the system equation of equation (1).

$$\Delta^k t_n = v_n \quad \text{(System equation)} \tag{9}$$

where Δ^k : k-th order difference operator, and the second order difference operator is applied in this study as $t_n - 2t_{n-1} + t_{n-2} = v_n$.

Figure 3 shows the example of the spectra of the logarithm of the significant wave height measured at Mutsuogawara port and the atmospheric pressure data at a point. The spectra of the separated time series data of long-term component and short-term component are shown in the figure. As seen in the figure, appropriate separation

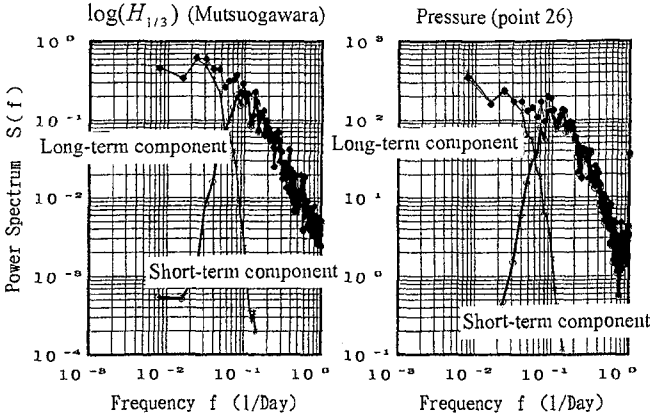

Figure 3 Spectra of the logarithm of the significant wave height at Mutsuogawara port and atmospheric pressure with the separated components

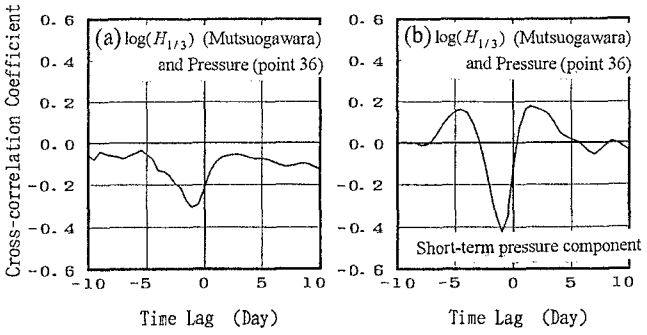

Figure 4 Cross-correlation coefficients between the logarithm of the significant wave height at Mutsuogawara port and the atmospheric pressure data

can be done by the Kalman filter by choosing a proper value of the trade-off parameter σ_v^2/σ_w^2, where σ_v^2 and σ_w^2 are the variances of v_n and w_n in equations (9) and (8), respectively. The trade-off parameter of the Kalman filter is used to control the magnitude of the change of the model in each time step of the computations. That is, by changing the value of σ_v^2/σ_w^2, the ratio of the energy of the separated long-term component and the short-term component can be controlled.

Figure 4 shows the cross-correlation coefficients between the logarithm of the significant wave height measured at Mutsuogawara port and the atmospheric pressure data at a point. The cross-correlation coefficient in figure (a) was estimated from the original atmospheric pressure data. The cross-correlation coefficient in figure (b) was estimated from the short-term atmospheric pressure component separated from the original data. The cross-correlation coefficient in figure (a) shows distinct positive and negative peaks with clear time lag between the two time series, though figure (b) shows vague negative peaks and time lag.

The purpose of separating the atmospheric pressure data into two components is to reduce negative effects from the long-term component to the short-term forecasting. This is necessary since we focus on the short-term wave forecasting and the long-term component may cause a bias in the relation between the criterion variable and the predictor variables.

2) Principal Component Analysis of atmospheric pressure data

The second step is the PCA by which the separated time series data on the grid points are projected onto the empirical eigen-vectors obtained from each component of the atmospheric pressure data on the grid points. Here, the empirical eigen-vectors are preliminarily computed by using the past three-years' atmospheric pressure data. The outline of the PCA is as follows.

Atmospheric pressure field $P_{z,t}$ can be approximately expressed by the

superposition of the eigen-vectors $e_{n,z}$ with weighting coefficients $c_{n,t}$:

$$P_{z,t} = \sum c_{n,t} e_{n,z} \tag{10}$$

where atmospheric pressure field $P_{z,t}$ is normalized by the mean value and standard deviation of $P(x,y,t)$, and each eigen-vector $e_{n,z}$ is assumed to be orthogonal to each other as

$$\sum e_{n,z} e_{m,z} = \delta_{n,m} \tag{11}$$

Then the eigen-vector $e_{n,z}$ can be estimated by solving the following equation.

$$\mathbf{A} = \lambda_{\mathbf{n}} \mathbf{e}_{\mathbf{n}} \tag{12}$$

where λ_n: eigen-value, $a_{i,j}$ is the (i,j) component of matrix $\mathbf{A}$ and is expressed by

$$a_{i,j} = \frac{1}{n_z n_t} \sum_{t=1}^{n_t} P_{i,t} P_{j,t} \tag{13}$$

Using the orthogonal condition of the eigen-vector, the weighting coefficient $c_{n,t}$ can be obtained by

$$c_{n,t} = \sum P_{z,t} e_{n,z} \tag{14}$$

Figure 5 shows examples of the eigen-vectors of the long-term component of the atmospheric pressure system. From the left to the right in **Figure 5**, each figure shows the 1-st, the 2-nd, the 3-rd, and the 4-th principal component, respectively.

As seen in the figure, the 1-st component is invariable with respect to time t. This component seems to be the average of the atmospheric pressure system. The 2-nd component seems to show the phenomenon in which the atmospheric pressure

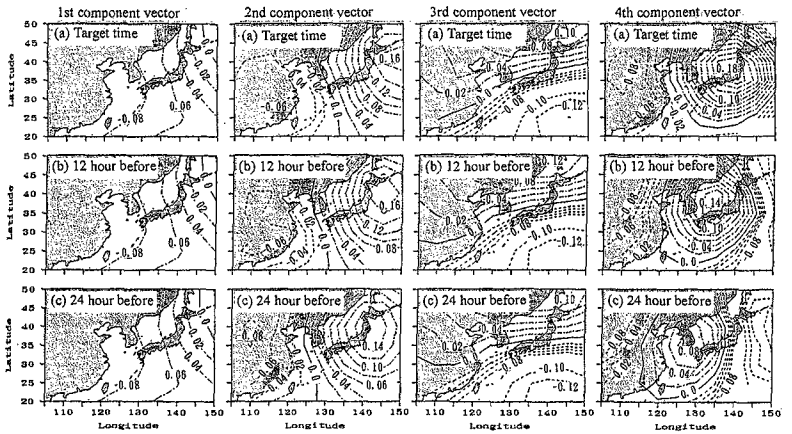

Figure 5 Components of PCA for target time and 12, 24 hours before target time

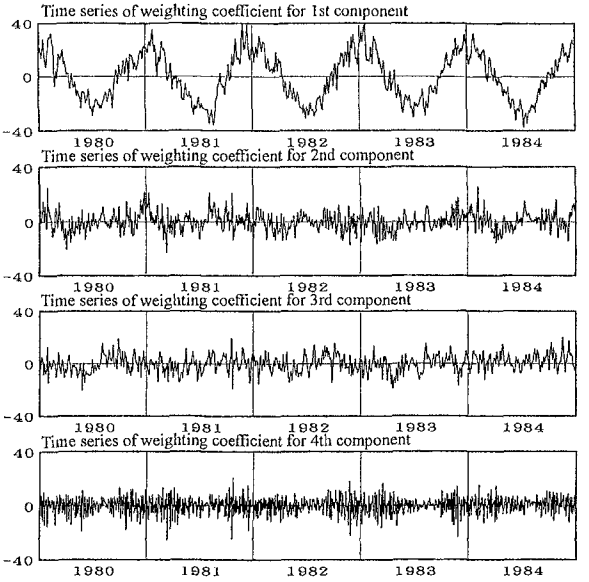

Figure 6 Time series of the weighting coefficients of the PCA

system moves from the west to the east. The 3-rd component shows from the south to the north. The 4-th component from the south-west to the north-east with the developing pressure system. The behavior of the atmospheric pressure system is assumed to be approximated by the superposition of these orthogonal empirical eigen patterns in this study.

The weighting coefficient of each eigen-vector is stored to be used in the following 3-rd step computation. Through this procedure, the atmospheric pressure data on the grid points are transformed and condensed into fewer, yet more efficient and independent predictor variables. The purpose of introducing the PCA is to avoid the unfavorable effect of multicollinearity of the atmospheric pressure data on the space-time grid points, by which new predictor variables are generated through the PCA. **Figure 6** is an example of the time series of the weighting coefficient $c_{n,t}$, which is used as the predictor variable in the next step.

3) Time dependent regression model for wave forecasting by Kalman Filter

The final step is the wave forecasting where the weighting coefficients previously obtained are used as the input data for the time-dependent regression model utilizing the Kalman filter.

The equation for the time-dependent regression model is assumed by

$$\log_{10} H_{1/3} = a_0 + \sum_{i=1}^{N} a_i z_i + \varepsilon \tag{15}$$

This equation can be reduced to

$$y_n = H_n x_n + w_n \quad \text{(Observation equation)} \tag{16}$$

where $y_n = \log_{10} H_{1/3}$, $x_n = (a_0, a_1, \cdots, a_N)^t$, $H_n = (1, z_1, z_2, \cdots, z_N)$, $w_n = \varepsilon$ and the weighting coefficient $c_{n,t}$ is expressed as z_i for convenience.

If the coefficient a_n is slowly varying value, then

$$\Delta^k x_n = v_n \quad \text{(System equation)} \tag{17}$$

where Δ^k is the k-th order difference operator, and the first order difference operator is applied in this study as $a_n - a_{n-1} = v_n$.

The trade-off parameter defined by σ_v^2/σ_w^2 is used to control the magnitude of the change of the model in each time step of the computations, where σ_v^2 and σ_w^2 are the variances of v_n and w_n in equations (17) and (16), respectively. **Figure 7** shows examples of the variations of the time series of the state variables a_i $(i = 0, \cdots, N)$. Figure (a) shows the results calculated with $\sigma_v^2/\sigma_w^2 = 10^{-3}$, while figure (b) shows the results calculated with $\sigma_v^2/\sigma_w^2 = 10^{-10}$. When choosing a small trade-off parameter, the time variations of the state variables a_i $(i = 0, \cdots, N)$ can be suppressed so as to fluctuate around the mean values as seen in figure (b). The rapid change of the state variables a_i $(i = 0, \cdots, N)$ in figure (a) seems to reflect the over-fitting of the model to the data. In other words, the problem of the over-fitting can be reduced by choosing a proper value of the trade-off parameter σ_v^2/σ_w^2 in the model.

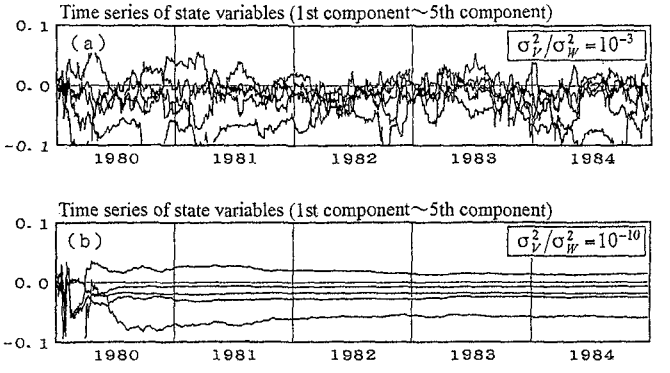

Figure 7 Time series of the state variables a_i $(i = 0, \cdots, N)$ of the time-dependent regression model

The purpose of adopting a time-dependent regression model utilizing Kalman filtering is to detect a gradual change such as a seasonal variation in the atmospheric and wave systems, and to reflect it in the forecasting model to improve the accuracy of the wave forecasting.

These three steps are repeated for each time step. That is, by obtaining new wave data or atmospheric pressure data, the wave forecasting equation is updated to improve the accuracy of the wave forecasting.

Numerical simulations of wave forecasting based on 5-years of data

The applicability and reliability of the proposed method is examined for six wave observation stations around Japan, shown in **Figure 1,** through simulations based on 5-years of wave data and weather charts.

Figure 8 shows the accuracy of the proposed method applied for Mutsuogawara port. The horizontal axis is the lead time for wave forecasting. The vertical axis is the standard deviation of the prediction errors. It is seen that the predicted wave height by the proposed method shows different characteristics depending on the magnitude of the trade-off parameter, σ_v^2/σ_w^2, of the Kalman filter used in the 3-rd step. The trade-off parameter controls the magnitude of the change of the coefficients in each time step of the computations. If an appropriate trade-off parameter is chosen, the wave height errors, predicted several-steps-ahead, can be controlled within an allowable range, although the prediction error of one-step-ahead may not be the minimum. In other words, the problem of the over-fitting of the criterion variable to the data can be resolved by choosing an appropriate value of the trade-off parameter in the model.

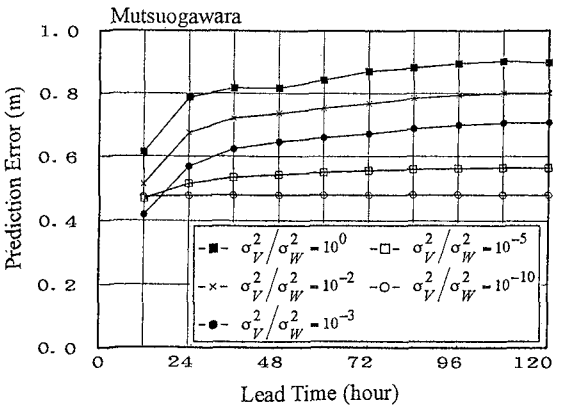

Figure 8 Accuracy of the proposed method (Error vs. lead time)

	Consideration of the separation of long-term component and short-term component	Trade-off parameter σ_v^2 / σ_w^2
(a)	×	10^{-3}
(b)	×	10^{-10}
(c)	○	10^{-3}
(d)	○	10^{-10}

Table 1 Simulation conditions for wave forecasting

To examine the validity of the proposed method, we applied the proposed method for 4 different computation conditions. **Table 1** shows the 4 cases. **Figure 9** shows the scatter diagram between the observed wave height and the forecasted wave height for a lead time of 120 hours under the 4 different conditions for Mutsuogawara port.

The low correlation coefficient between the forecasted value and the observed value can be seen in the case (a) where the separation of the long-term component is not considered, and the trade-off parameter is also inappropriate. In the case (b), the trade-off parameter is properly chosen though the separation of the long-term component is not considered. In this case, the correlation coefficient is improved compared to the case (a). However, prominent bias between the forecasted value and observed value can be seen. In the case (c) where the separation of the long-term

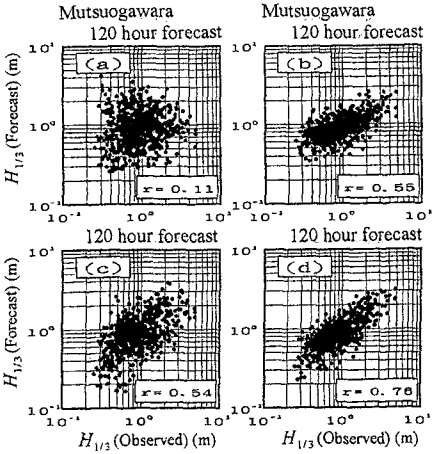

Figure 9 Scatter diagram between the observed wave height and the forecasted wave height for a lead time of 120 hours under the 4 different conditions

component is considered while the trade-off parameter is not appropriate, the correlation coefficient is not improved when compared to the case (b) although it does not show the prominent bias. In the case (d) where the separation of the long-term component is considered and the trade-off parameter is appropriate, this case shows the highest correlation coefficient in the 4 cases and does not show the prominent bias between forecasted and observed values.

From these comparative results obtained under the different conditions using the same data, it is demonstrated that each technique introduced in each step of the proposed method is effective. It is also demonstrated that the accuracy of the proposed method for short-term wave forecasting is better than the other methods (Kobune, et.al., 1988, 1990 and Suda and Yuzawa, 1983) when an appropriate trade-off parameter is properly chosen.

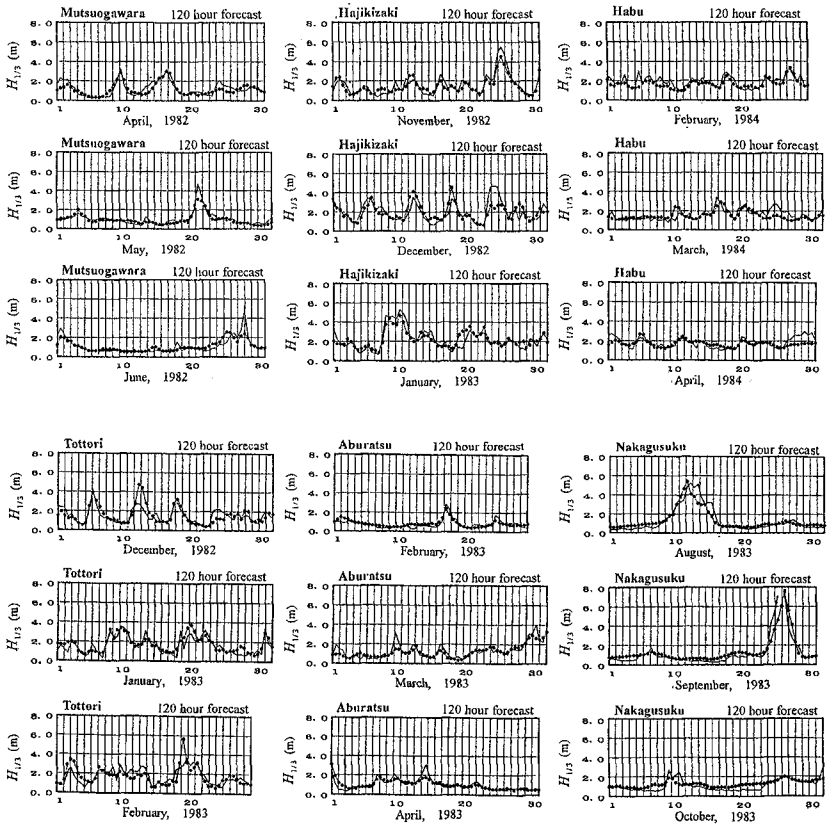

Figure 10 Comparison of the time series of the forecasted wave heights (●) for a lead time of 120 hours and the observed wave height (solid line)

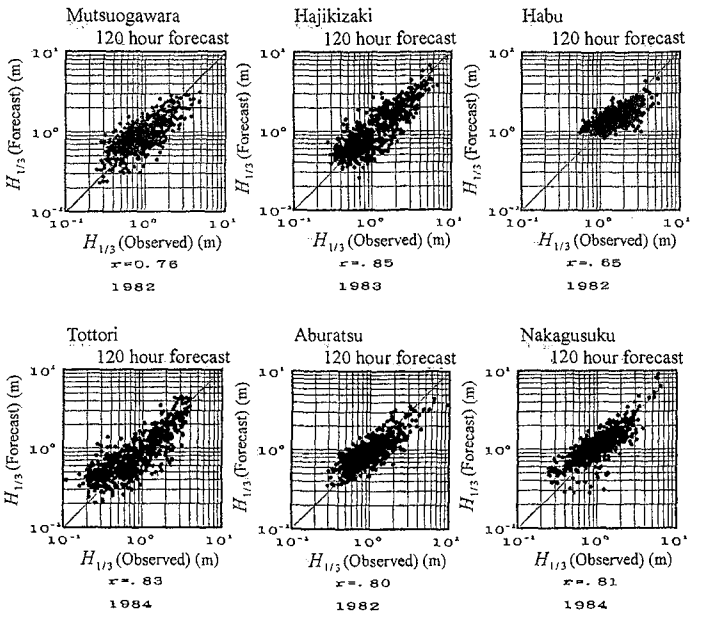

Figure 11 Scatter diagram between the observed wave heights and the forecasted wave heights for a lead time of 120 hours.

The examinations of the wave forecasting for other observation stations were carried out using the atmospheric pressure data at 500km × 500km grid points around the wave observation stations shown in **Figure 1**. **Figure 10** shows an example of the comparison of the time series of the forecasted wave heights for a lead time of 120 hours and the observed wave height, where ● is the forecasted wave height and solid line is the observed wave height. As seen in the figure, the tendency of the time delay of the forecasted wave height to the real wave height is not recognized although most of the conventional statistical wave forecasting methods show such a drawback (Kobune, et.al., 1988, 1990 and Suda and Yuzawa, 1983).

Figure 11 shows the scatter diagram between the observed wave heights and the forecasted wave heights for a lead time of 120 hours. The examples for 6 wave observation stations around Japan are shown in the figure. All the data in one year are plotted in the figure. These examples demonstrate that the reliability of the proposed method for short-term wave forecasting can be acceptable for practical use if errors are allowed to a certain extent.

Conclusions

The overall conclusions of this study are summarized below.

1) The proposed method utilizing the Kalman filter combined with Principal Component Analysis can be a useful tool for a short-term wave forecast if errors are allowed to a certain extent.
2) If an appropriate trade-off parameter is chosen in the model, the wave height errors predicted several-steps-ahead can be controlled within an allowable range, although the prediction error one-step-ahead may not be the minimum.
3) The proposed method is easy to handle, which enables us to use the proposed method with a small personal computer.

Acknowledgement

We wish to express our sincere gratitude to Mr. Sidney Walter thurston III for critical review and valuable comments on this paper.

References

Kobune, K. and N. Hashimoto (1988): On the reliability of the wave forecasting by the multiple regression model, Coastal Engineering in Japan, Vol. 31, No.2, pp.199-205.

Kobune, K., N. Hashimoto and Y. Kameyama (1990): On the reliability of wave forecasting by Emprical wave forecasting models, Technical Note of P.H.R.I., No.673, 42p (in Japanese).

Suda, H. and A. Yuzawa (1983): A fundamental study on the queuing theory of the open sea berth based on the wave forecasting, Proc. Of JSCE, Vol.339, pp.177-185.