Web Application for Time-Series Analysis based on Particle Filter Available on Cloud Computing System

Hiromichi Nagao
The Institute of Statistical Mathematics
Research Organization of Information and Systems
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
hnagao@ism.ac.jp

Tomoyuki Higuchi
The Institute of Statistical Mathematics
Research Organization of Information and Systems
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
higuchi@ism.ac.jp

Abstract - We develop web application "CloCK-TiME" (Cloud Computing Kernel for Time-series Modeling Engine), which enables users to analyze their time-series data by using a networked PC cluster in a cloud computing system. This software decomposes a given multivariate time-series data into trend, seasonal, autoregressive (AR), and observation noise components, by using the particle filter (PF) algorithm. We also develop a user interface, by which users can set parameters needed in the analysis such as trend order, seasonal period, AR order, and the number of particles. We show an application example in the case of tide gauge data recorded along the coastline of Japan. We are planning to improve our analysis engine in order to obtain not only optimum model parameters but also their posterior distributions eventually by a hybrid method consisting of the PF and the MCMC algorithms.

Keywords: particle filter, multivariate analysis, AR model, MCMC, cloud computing system

1 Introduction

The particle filter (PF) is a powerful tool for model parameter estimation comparing with other Bayesian computational algorithms because the PF is easy to implement on a parallel computer system owing to its scalability. However, requisite number of particles is exponentially increasing as target model is larger because of so-called "curse of dimensionality". For example, Nakamura *et al*. (2009) [1] showed that up to one hundred million particles are required in order to express each posterior distribution sufficiently without a degeneracy in the case of the biological transcription regulatory network model for circadian clock system of mammalian, which has 44 model parameters. Because of a limitation of computer resources in the case of a single PC cluster, an ingenious design would be necessary especially in handling models having much more number of parameters even though the next-generation supercomputer under construction by RIKEN is available.

We introduce, in this paper, web application "CloCK-TiME" (Cloud Computing Kernel for Time-series Modelling Engine), which enables us to analyze a given multivariate time-series data with PC clusters on a cloud computing system. Our method makes it possible to decompose given time-series data into trend, seasonal, autoregressive (AR), and observation noise components. A non-Gaussian distribution function such as the Cauchy distribution is possible to be assumed for the system noise of the trend component for the purpose to extract sudden changes in the trend component. We adopt a multivariate AR model for the purpose to extract correlations among the variates. We apply our method, as an example, to monthly means of tide gauge records observed along the coastline of Japan for several decades. A hybrid method consisting of the PF and the MCMC algorithms would be adopted in our system that enables us to estimate not only optimum parameters but also their posterior distributions with a large-scale parallel computation.

We mention our methodology in Section 2, show an application example in the case of tide gauge records in Section 3, and introduce web application "CloCK-TiME" in Section 4, which provides cloud computing services to those in various fields of science.

2. Methodology

Our analysis engine decomposes given time-series data y_t at time t into four components, i.e., an observation model can be written as

$$y_t = u_t + s_t + p_t + w_t \quad , \tag{1}$$

where each term in the right hand is trend, seasonal, AR, and observation noise component, respectively. Since the engine allows a multivariate analysis, each term in equation (1) has a vector form, and the degree of each vector is assumed here to be L.

The system model that defines the trend component is written as

$$(1-B)^k u_t = v_{u,t} \quad , \tag{2}$$

where k is a trend order, B is the lag operator, and $v_{u,t}$ is a system noise. Either Gaussian or non-Gaussian distribution function is allowed for the system noise in the engine. In the former case, the Gaussian distribution function is assumed to have a mean vector of zero and a

covariance matrix of τ_u, i.e., $v_{u,t} \sim N\left(0, \tau_u\right)$, and we assume here that τ_u is diagonal for the purpose to reduce the number of model parameters. The trend component is allowed to change slightly with time owing to this system noise, but a sudden baseline jump sometimes recorded in geophysical data due to such as earthquakes is never extracted with such a Gaussian system noise. In these cases, a system noise following a non-Gaussian distribution function such as a Cauchy distribution function should be adopted.

The seasonal component s_t extracts a periodic variation clearly seen in the time series. The system model for the seasonal component can be derived from an assumption that the sum of the seasonal components through a period is almost zero, i.e.,

$$s_t = -\sum_{i=1}^{l} s_{t-i} + v_{s,t} \quad , \tag{3}$$

where l is a seasonal period, and $v_{s,t}$ is a system noise that follows a Gaussian distribution function having mean vector zero and diagonal covariance matrix τ_s, i.e., $v_{s,t} \sim N\left(0, \tau_s\right)$. Owing to this system noise, the seasonal component can vary its amplitude and phase gradually with time.

The AR component p_t extracts periodic variations (e.g., Higuchi (1991) [2]) having different periods with the seasonal component. We adopt a multivariate AR model, i.e.,

$$p_t = \sum_{m=1}^{M} A_m p_{t-m} + v_{p,t} \quad , \tag{4}$$

where M is an AR order, A_m $(m = 1, \cdots, M)$ are AR coefficient square matrices of degree L, and $v_{p,t}$ is a system noise that follows a Gaussian distribution function of mean vector zero and diagonal covariance matrix τ_p, i.e., $v_{p,t} \sim N\left(0, \tau_p\right)$.

The observation noise component w_t is a residual that never be explained by other three components, which follows a Gaussian distribution function of mean vector zero and diagonal covariance matrix σ, i.e., $w_t \sim N\left(0, \sigma\right)$.

A set of the system models (i.e., equations (2)-(4)) and observation model (i.e., equation (1)) is summarized as a state space model

$$x_t = F x_{t-1} + G v_t \tag{5}$$

$$y_t = H x_t + w_t \quad , \tag{6}$$

where a state vector x_t, a system noise vector v_t, and three matrices F, G, and H have the forms

$$x_t = \left(u_t' \quad u_{t-1}' \quad s_t' \quad \cdots \quad s_{t-10}' \quad p_t' \quad \cdots \quad p_{t-m+1}'\right)' \tag{7}$$

$$v_t = \left(v_{u,t} \quad v_{s,t} \quad v_{p,t}\right)' \tag{8}$$

$$F = \begin{bmatrix} F_u & & \\ & F_s & \\ & & F_p \end{bmatrix} \tag{9}$$

$$G = \begin{bmatrix} G_u & & \\ & G_s & \\ & & G_p \end{bmatrix} \tag{10}$$

$$H = \begin{bmatrix} H_u & H_s & H_p \end{bmatrix} \quad , \tag{11}$$

where the prime denotes a transposed matrix. The submatrices in each matrix in equations (9), (10), and (11) have the forms

$$F_u = \begin{bmatrix} {}_kC_2 I_k & -{}_kC_3 I_k & \cdots & (-1)^k {}_kC_k I_k \\ I_k & & & \\ & \ddots & & \\ & & I_k & O \end{bmatrix} \tag{12}$$

$$F_s = \begin{bmatrix} -I_l & -I_l & \cdots & -I_l \\ I_l & & & \\ & \ddots & & \\ & & I_l & O \end{bmatrix} \tag{13}$$

$$F_p = \begin{bmatrix} A_1 & A_2 & \cdots & A_M \\ I_L & & & \\ & \ddots & & \\ & & I_L & O \end{bmatrix} \tag{14}$$

$$G_u = \begin{bmatrix} I_k \\ \vdots \\ 0 \end{bmatrix} \tag{15}$$

$$G_s = \begin{bmatrix} I_l \\ \vdots \\ 0 \end{bmatrix} \tag{16}$$

$$G_p = \begin{bmatrix} I_M \\ \vdots \\ 0 \end{bmatrix} \tag{17}$$

$$H_u = \begin{bmatrix} I_k & \cdots & 0 \end{bmatrix} \tag{18}$$

$$H_s = \begin{bmatrix} I_l & \cdots & 0 \end{bmatrix} \tag{19}$$

$$H_p = \begin{bmatrix} I_M & \cdots & 0 \end{bmatrix} \quad , \tag{20}$$

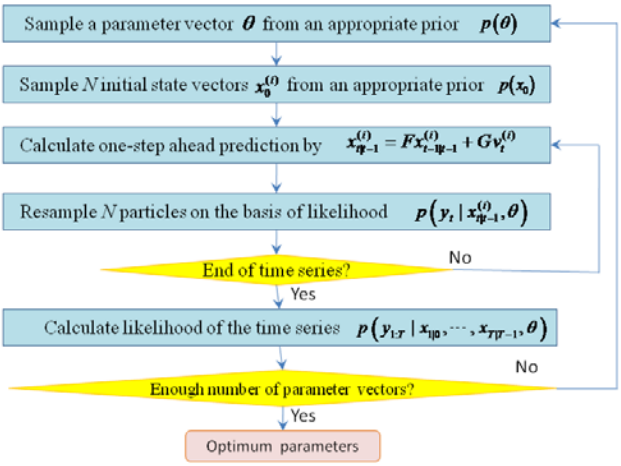

Figure 1: Flow chart to optimize model parameters in the framework of the particle filter.

where I_n is a unit matrix of degree n. Unknown parameters in our time series model are summarized in a hyperparameter

$$\theta = \left(\{A_1\}_{ij} \quad \cdots \quad \{A_M\}_{ij} \quad \{\tau_u\}_i \quad \{\tau_s\}_i \quad \{\tau_p\}_i \quad \{\sigma\}_i \quad x_0' \right)' \quad (21)$$

where the suffixes i and j denote components of the matrices, x_0 is an initial state vector, i.e., the number of parameters to be estimated is $N_p = ML^2 + (k + l + M + 3)L$. All of the parameters are to be estimated by the maximum likelihood principle. Nagao *et al.* (2002, 2003) [3][4] adopted the Kalman filter algorithm to determine the model parameters, however the Kalman filter can be applied only when all of the probability distribution functions in the model are the Gaussian distribution functions. A non-Gaussian distribution function can be adopted for the system noise of the trend component in our model, so that the Kalman filter is no longer available in this case. We apply, in this paper, the PF algorithm, which is a kind of the Monte Carlo method, in order to estimate the model parameters. Figure 1 shows a schematic of the procedure to estimate the model parameters. First we set a prior distribution for the hyperparameter as a multi-dimensional Gaussian distribution, i.e.,

$$p(\theta) = \frac{1}{(2\pi)^{N_p/2}|S|} \exp\left[-\frac{(\theta - \theta_0)' S^{-1} (\theta - \theta_0)}{2} \right], \quad (22)$$

where θ_0 is a mean vector, S is a covariance matrix, and the vertical bars denotes a determinant. Our code is assumed to work without a given prior distribution, so that it is necessary to generate the prior distribution automatically from input time series data. The Gaussian assumption is sufficient in this situation, although the prior distribution is, of course, allowed generally to be a non-Gaussian. How to give the mean vector θ_0 is always an important problem especially in the case of the initial vector x_0, and we set it as following algorithm in this paper. First we fit a regression line having an order of k, i.e., the trend order specified in advance, to each time-series, and the value of the obtained regression line at time t = 0 is regarded as means of prior distributions of the trend components in the initial state vector. Then the regression line is subtracted from the original time-series data, and the residuals are stacked into a time interval of l, i.e., the seasonal period specified in advance. This stacked data are assumed to be means of seasonal components in the initial state vector. We assume that means of all AR components in the initial state vector are zeroes.

We determine prior distributions for multivariate AR coefficient matrices by solving the Yule-Walker equation

$$C_0 = \sum_{m=1}^{M} A_m C_{-m} + W \quad (23)$$

$$C_k = \sum_{m=1}^{M} A_m C_{k-m} \quad (k = 1, 2, \cdots) \quad , \quad (24)$$

where W is a covariance matrix of the system noise $v_{p,t}$. C_k $(k = 1, 2, \cdots)$ are symmetric cross-covariance matrices with lag k, i.e.,

$$\{C_k\}_{ij} = E\left[\left(y_t^{res,i} - \mu^i \right) \left(y_t^{res,j} - \mu^j \right) \right] \quad , \quad (25)$$

where $y_t^{res,i}$ is the residual obtained by subtracting the above regression line and stacked data from the original data, and μ^i denotes the mean of the i-th residual data. We apply the Levinson's algorithm (e.g., Wiggins and Robinson (1965) [5]) in the process to obtain the AR coefficient matrices and covariance matrix, and let the solution be the mean of prior distribution. When we sample initial particles from the prior distribution with appropriate variances, all of the samples should satisfy a condition of AR stationary. This condition is equivalent to that all of roots of an equation

$$\left| I_L - \sum_{m=1}^{M} \omega^m A_m \right| = 0 \quad (26)$$

are outside of the unit circle in the complex plane. We use the Lehman-Schur method to check whether this condition is satisfied or not for each sampled particle.

We carry out the PF for each sampled particle to calculate likelihood function $p(y_{1:T} | \theta)$, where $y_{1:T}$ denotes observation data at all time points. We adopt the residual systematic resampling method in the filtering step of PF, and apply a particle smoother algorithm with a lag of double of seasonal period (e.g., Kitagawa (1996) [6]).

3. Application to Tide Gauge Data

In order to show the performance of our methodology, we apply the analysis engine to tide gauge data obtained in Japan. Sea level along the coastline of Japan has been observed since 1872 at ~150 tide gauges operated by

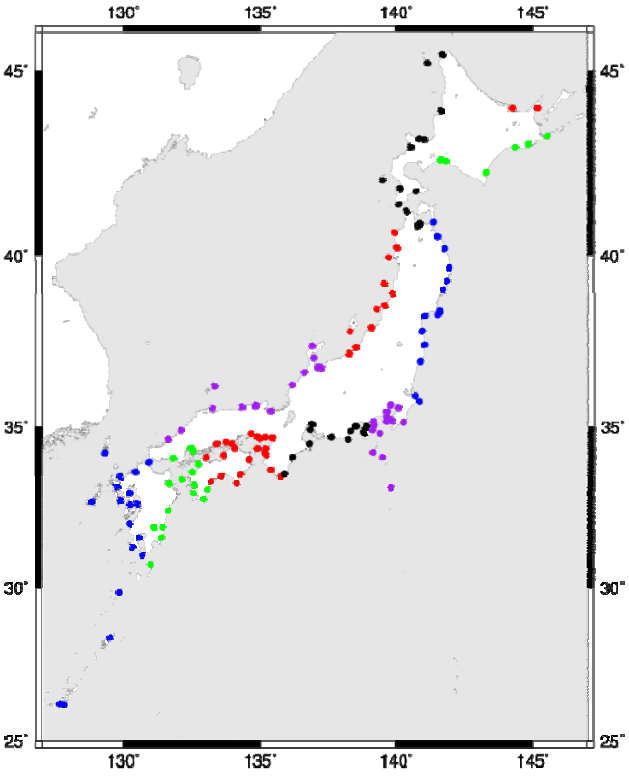

Figure 2: Distribution of tidal observatories along the coastline of Japan. ~150 observatories have been operating for >100 years. The color coding denotes the sea area divisions defined by Tsumura (1963) [8].

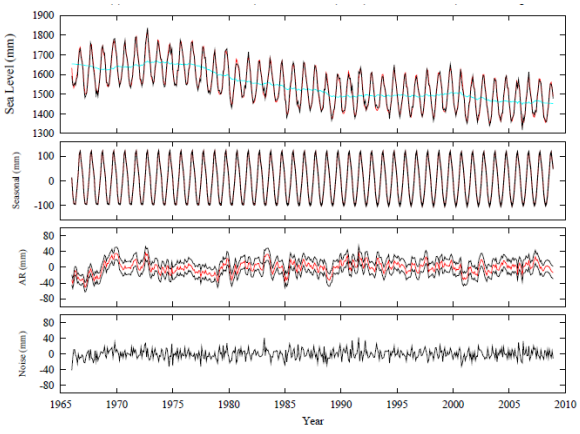

Figure 3: Decomposition of tide gauge record at an observatory. (top panel) Original observation data traced by the trend component shown in blue, (second panel) seasonal component, (third panel) AR component shown in red with standard deviation shown in black, and (bottom panel) observation noise component.

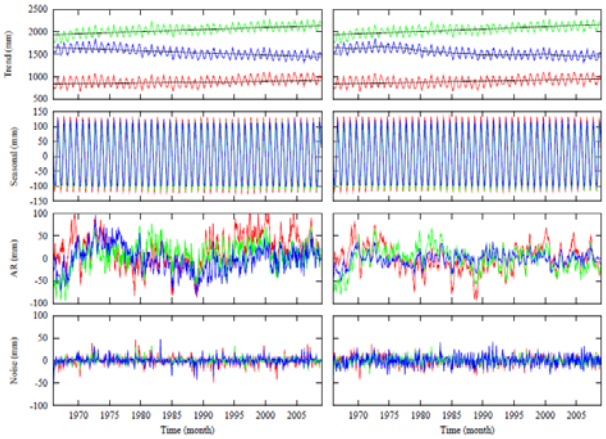

Figure 4: Comparison between univariate (left) and multivariate (right) analyses of tide gauge records at three nearby observatories.

Geographical Survey Institute (GSI), Japan Meteorological Agency (JMA), Japan Coast Guard (JCG), and other institutes. The day-to-day tidal records clearly show the stationary oceanic tides sometimes including transient events such as tsunamis. On the other hand, long-term sea level variations for years are considered to demonstrate oceanic variations and/or tectonic deformations. In particular, a secular trend up to several millimeters per year is caused by tectonic process such as subduction of oceanic plates, and sometimes shows a sudden baseline jump due to interplate earthquakes. Kato and Tsumura (1979) [7] improved the analysis method of Tsumura (1963) [8] that enables us to extract evidences of tectonic deformation from the long-period tidal records through determination of a trend variation from monthly means with modeling both seasonal variation and spatial correlation with nearby observatories. One of the important results they derived is that the coasts of Japan can be classified into ten sea areas according to features of the spatial correlations (e.g., Kobayashi (2008) [9]). GSI has adopted this powerful tool as an official analysis method for more than ten years, but it should be improved at this moment especially in parts relying on subjective experiences. Our method can exactly cover the shortcomings of the previous method especially in points that objective estimations of long-term variations in the trend and seasonal components, and automatic interpolation of missing observation data.

Figure 2 shows the distribution of tide gauge stations located along the coastline of Japan with demonstrating the sea area divisions proposed by Tsumura (1963) [8]. We use, in this paper, monthly means of the tide gauge data obtained from 1966 to 2008, i.e., for 43 years. The sea level is affected by the atmospheric pressure at an observatory; the sea level decreases ~10mm when the atmospheric pressure increases 1hPa. We correct the observed tide gauge data to those under a condition of 1000hPa by the following relation

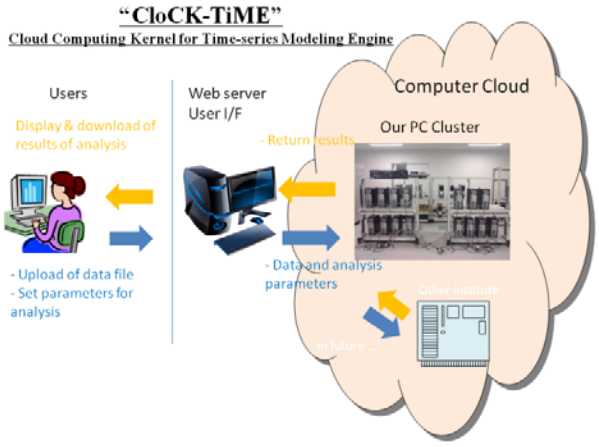

Figure 5: Overview of "CloCK-TiME" (Cloud Computing Kernel for Time-series Modelling Engine). A user can obtain results of time-series analysis of a hybrid method consisting of particle filter and MCMC performed by PC cluster in a cloud computing system.

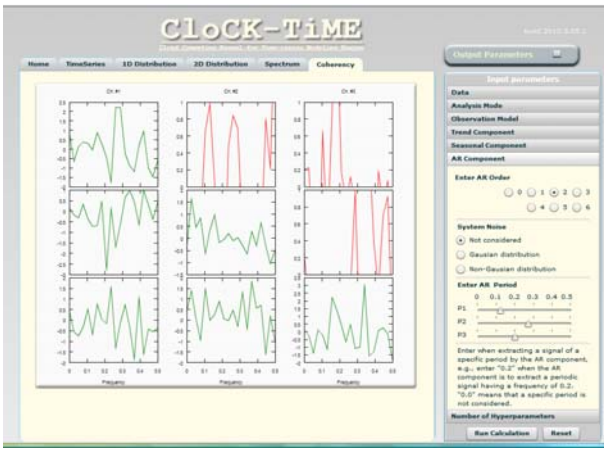

Figure 6: User interface of "CloCK-TiME". It is possible to set parameters needed in an analysis, and obtain resulting figures and a set of optimized parameters through this interface.

$$y_t^l = y_t^{l,\mathrm{obs}} + 10\left(P_t^l - 1000\right) \qquad (27)$$

where $y_t^{l,\mathrm{obs}}$ and P_t^l are monthly means of raw tide gauge data in millimeter and atmospheric pressure in hPa at the l-th observatory, respectively. When a tide gauge observatory does not measure the atmospheric pressure, we use the most nearby barometer record. We hereafter use the corrected observation data y_t^l in later analysis.

Figure 3 shows, for example, a result of decomposition obtained from tide gauge records at an observatory. The original time series has a long-term trend variation and clear annual variation, and we successfully extract these variations in trend and seasonal components, respectively. The AR component has a variation of several years, and this is considered to relate to oceanic variations related to the oceanic current "Kuroshio" and the oceanic general circulation in the North Pacific Ocean (e.g., Yasuda and Sakurai (2006) [10]).

Figure 4 shows a comparison between univariate analysis, i.e., individual application to each time-series, and multivariate analysis, i.e., simultaneous application to whole time-series, of nearby three observatories, which locate in the same sea area defined by Tsumura (1963) [8]. It is well-known that the oceanic variation in the same sea area shows a similar behaviour, so that it is a key issue to extract such spatial correlation towards a better time-series modelling. The observation noise component obtained by multivariate analysis is successfully reduced comparing with the case of the univariate analysis. This owes to the consideration of a spatial correlation between tide gauge observatories by the multivariate AR model.

4. Web Application "CloCK-TiME"

We develop web application "CloCK-TiME", which makes it possible to utilize our time series analysis method for users in various fields of science with a PC cluster in a cloud computing system. Figure 5 shows an overview of CloCK-TiME, and Figure 6 shows its user interface (I/F) coded by Flash. A user uploads his/her time-series data to our web server through the I/F, and set parameters such as trend order, seasonal period, AR order, and the number of particles (i.e., hyperparameters) needed in a time-series analysis. Then the I/F calls the analysis engine installed in a PC cluster with the input data file and parameters, and receives the calculation results. Finally the I/F shows resulting figures such as time-series of each component as shown in Figure 1, prior and posterior distributions for each model parameter, correlation between parameters, spectrum and coherency for each time-series. The user can also obtain a set of optimized numerical values of model parameters. All of these results can be downloaded via internet.

The CloCK-TiME is provided online, so that the computation time is one of the primary key issues. The computation time would be longer when the number of the particles is necessarily larger in case that a non-Gaussian distribution function is selected for a system noise or that the order of observation becomes larger. For the purpose to reduce the number of particles, the Rao-Blackwellized particle filter algorithm has a potential, which applies both Kalman filter and PF to the linear and the non-linear parts of the state space model (e.g., Doucet *et al.* (2001) [11]). We will implement eventually this algorithm to the CloCK-TiME, and set out convenient online software without unnecessarily long communication and/or calculation latency.

5. Conclusion

We develop a PF algorithm for a decomposition of multivariate time-series data, and make it available online named “CloCK-TiME”. It enables us to decompose a given multivariate time-series data into trend, seasonal, AR, and observation noise components. In this paper, we show only results of parameter optimization in the case of tide gauge records obtained in Japan. However, such an only parameter optimization is thought to be insufficient from a point of view of computation efficiency and/or “remodelling”, since a posterior distribution for each model parameter could give information of which parameter is dominant/indominant. We will eventually improve our method by plugging-in an MCMC algorithm (e.g., Gilks *et al*. (1994) [12]) to the PF, and make it possible to obtain the posterior distributions. This improvement must require more computer resources because a number of particles are necessary to express the distributions sufficiently with little degeneracy. Therefore an installation of our web application on a cloud computing system is an important issue for our project.

Acknowledgments

We appreciate Prof. Satoshi Miura, Drs. Daisuke Inazu, Ryo Yoshida, Keisuke Hayashi, Masaya M. Saito for their invaluable discussions on this research. We used tide gauge data supplied from Geographical Survey Institute (GSI), Japan Meteorological Agency (JMA), Japan Coast Guard (JCG), and other institutes. The map used in this paper is created by the Generic Mapping Tools.

References

[1] Nakamura, K., R. Yoshida, M. Nagasaki, S. Miyano, and T. Higuchi, Parameter estimation of *in silico* biological pathways with particle filtering towards a petascale computing, *Pacific Symposium on Biocomputing*, 14, 227-238, 2009.

[2] Higuchi, T., Frequency domain characteristics of linear operator to decompose a time series into the multi-components, *Ann. Inst. Statist. Math.*, 43, 469-492, 1991.

[3] Nagao, H., T. Iyemori, T. Higuchi, S. Nakano, and T. Araki, Local time features of geomagnetic jerks, *Earth Planets Space*, 54, 117-131, 2002.

[4] Nagao, H., T. Higuchi, T. Iyemori, and T. Araki, Lower mantle conductivity anomalies estimated from geomagnetic jerks, *J. Geophys. Res.*, doi:10.1029/2002JB001786, 2003.

[5] Wiggins, R. A. and E. A. Robinson, Recursive solution to the multichannel filtering problem, *J. Geophys. Res.*, 70, 1885-1891, 1965.

[6] Kitagawa, G., On Monte Carlo filter and smoother, *Proceedings of the Institute of Statistical Mathematics*, 44, 31-48, 1996 (in Japanese).

[7] Kato, T. and K. Tsumura, Vertical crustal deformation in Japan deduced from tide gauge records (1951-1978), *Bull. Earthq. Res. Inst.*, 54, 559-628, 1979 (in Japanese).

[8] Tsumura, K., Investigation of the mean sea level and its variation along the coast of Japan (Part I), *J. Geod. Soc. Japan*, 9, 49-90, 1963 (in Japanese).

[9] Kobayashi, A., Reexamination of sea area divisions defined by Tsumura for vertical crustal movement estimation using tidal records, *Shinken Jiho*, 71, 1-17, 2008 (in Japanese).

[10] Yasuda T. and K. Sakurai, Interdecadal variability of the sea surface height around Japan, *Geophys. Res. Lett.*, 33, doi:10.1029/2005GL024920, 2006.

[11] Doucet, A., J. F. G. Freitas, and N. Gordon, Sequential Monte Carlo Methods In Practice, *Springer-Verlag*, New York, 2001.

[12] Gilks, W. R., G. O. Roberts, and E. I. George, Adaptive direction sampling, *The Statistician*, 43, 179-189, 1994.