Data Assimilation System for Seismoacoustic Waves

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

Abstract—The Great East Japan Earthquake, which took place on March 11, 2011, caused giant tsunamis and seriously damaged the Pacific coast area of the Tohoku district, Japan. Some studies reported that microbarometers in East Asia recorded atmospheric pressure perturbations, which are believed to be evidence of acoustic waves excited by the tsunamis, several tens of minutes after the main shock. This observation suggests a way to establish an early warning system that utilizes the fact that acoustic velocity is much faster than the propagation speed of a tsunami. Such a system would require data assimilation (DA), which is a fundamental technique to combine a numerical simulation and observation data, in order to predict the magnitude and arrival time of a tsunami and to warn people in coastal areas. The present paper introduces a DA system for seismoacoustic waves, which has been under development at the Research and Development Center for Data Assimilation at the Institute of Statistical Mathematics, Japan. We also validate this system by using a twin experiment that assumes a giant earthquake takes place in the Pacific Ocean off the coast of West Japan, which is expected to take place in the near future. The proposed system will eventually contribute to the early warning system for tsunamis, although the time needed for the DA computation and for the handling of observation data still needs to be improved.

I. INTRODUCTION

The Great East Japan Earthquake, which took place on March 11, 2011, caused giant tsunamis and tremendously damaged the Pacific coast area of the Tohoku district, Japan. Although it is likely that such an earthquake occurs periodically at an oceanic plate boundary, even cutting-edge numerical simulation techniques, global-scale observation networks using highly sensitive sensors and satellites, and the Earthquake Early Warning System, which takes full advantage of information technology (IT), underestimated the magnitude of approaching tsunamis and were unable to minimize the damage. Since later research has clarified that observation data contained information on phenomena that indicated the magnitude of the tsunamis, simulation models in which the observation data were properly embedded would have accurately predicted the magnitude of the tsunamis; seismologists greatly regret not having prepared this in advance. It is important to develop a methodology, such as data assimilation (DA), which combines numerical simulations and observation data, in anticipation of the forthcoming Tokai-Tonankai-Nankai Earthquake (Nankai Trough Earthquake), which is predicted to take place in the Pacific Ocean off the coast of West Japan in the near future.

Although seismograms are still the basic form of observation data in earthquake research, recent improvements in Earth observation sensors are capable of detecting seismic phenomena even from non-seismic observations. For instance, the Global Positioning System (GPS) network, which covers all of Japan, has clearly detected long-term spatial and temporal variations of the Earth’s crust. DA research using GPS data has indicated that some regions of oceanic plate boundaries are firmly bonded and others are slippery. In the firmly bonded regions, large magnitude earthquakes occur when there is a sudden release of the accumulated energy due to plate subduction. Another example of useful information from nonseismic data is that acoustic waves that are emitted from a hypocenter and propagated in the upper atmosphere are often detected as “sounds of the earthquake” by distant microbarometers. Such seismoacoustic waves excited by the tsunamis of the Great East Japan Earthquake were reported to travel several thousands of kilometers [1], oscillating the ionosphere at an altitude of several hundreds of kilometers [2]. This phenomenon motivates us to establish a tsunami early warning system that utilizes the fact that acoustic waves propagate much faster than tsunamis.

Section II presents a brief introduction to real observations of seismoacoustic waves and a method for reproducing observed waveforms using numerical simulation, followed in Section III by a presentation of the DA system for seismoacoustic waves that is being developed by the Research and Development Center for Data Assimilation at the Institute of Statistical Mathematics, Japan. In Section IV, we present the results of the twin experiment that we used to validate the DA system. In this experiment, we determined whether the system is capable of accurately estimating the true model parameters from synthetic observation data produced by a simulation that assumes that a giant earthquake takes place in the Pacific Ocean off the coast of West Japan.

II. SEISMOACOUSTIC WAVES

The Great East Japan Earthquake activated the Earth’s crust beneath Japan, and, even now, there are frequent aftershocks, despite two years having passed since the main shock. Fig. 1(a) shows the perturbation in the atmospheric pressure that was due to an aftershock of magnitude of 6.1 that took place on March 14, 2012. This record was obtained by a Nano-Baro Model 6000-16B microbarometer (Fig. 2) installed in Sugadaira, Nagano Prefecture, Japan, 250 km from the epicenter. The microbarometer sensor can observe the absolute atmospheric pressure with an accuracy of 8 pascals (Pa) and a resolution of $10^{-5}$ Pa, with a controlled sampling frequency from 0 to 100 Hz. A Butterworth filter of a periodic band of 1–20 seconds was applied to the observation data in order to remove the background pressure of the Earth’s
Since the present paper assumes a linear system for the Earth, the simulation code is capable of computing a theoretical atmospheric pressure perturbation at an arbitrary place on the Earth in accordance with a given earthquake mechanism, which is called a “centroid moment tensor (CMT)”. Several CMT solutions have been developed for various earthquakes by universities and research institutes, including Harvard University [7] and Japan Meteorological Agency [8]. The present paper models the source of an earthquake, a rupturing fault in the Earth’s crust, by a series of sequentially exploding CMTs that are lined up along the fault.

A CMT is a tensor of order two and degree three, having the form

$$M = \begin{pmatrix}
M_{rr} & M_{r\theta} & M_{r\phi} \\
M_{\theta r} & M_{\theta\theta} & M_{\theta\phi} \\
M_{\phi r} & M_{\phi\theta} & M_{\phi\phi}
\end{pmatrix}$$

(1)

where \((r, \theta, \phi)\) is expressed in spherical coordinates with the origin at the hypocenter. \(M_{\alpha\beta}\) (where \(\alpha\) and \(\beta\) indicate either \(r\), \(\theta\), or \(\phi\)) is the moment that acts in the \(\alpha\)-direction and affects the \(\beta\)-direction. Equilibrium of the moment requires symmetry in the CMT, i.e., \(M_{\alpha\beta} = M_{\beta\alpha}\), so the CMT has six free parameters. As mentioned above, a convolution of the Green’s function and a CMT provides a response function, which is a temporal variation in the atmospheric pressure at an observatory. Fig. 1(b) shows a theoretical waveform computed by the normal mode summation technique for the aftershock on March 14, 2012; it assumes that the fault consists of a single source, which is described by a CMT that JMA developed. The simulated traveltimes, i.e., the times from the main shock to the arrival of the Rayleigh and acoustic waves, explain much of the observation shown in Fig. 1(a). However, inconsistencies in both waveforms imply that a fault should be modeled by multiple sources rather than a single source, and/or that a more accurate estimation of the hypocentral parameters is needed. The DA system for seismoacoustic waves, which will be introduced in Section III, is able to deal with a more complex fault model and to accurately estimate the model parameters by using the Markov chain Monte Carlo (MCMC) method.

III. DATA ASSIMILATION SYSTEM FOR SEISMOACOUSTIC WAVES

A. State Space Model

DA is a fundamental technique for combining a numerical simulation model and observational/experimental data. Although DA was originally developed for meteorology and oceanology [9], it is now applied to various fields such as solid Earth, life, and industrial sciences [10] [11] [12]. The ultimate goal of DA is to provide numerical simulation models that are capable of predicting future states by embedding “real” observation data into “virtual” simulation models [13]. Using a sequential Bayesian filter, DA estimates a probability density function of the parameters contained in a simulation model as well as a probability density function of the state variable at each time step. In the present paper, we focus on estimating the model parameters, since our aim is to establish a DA procedure that can rapidly analyze streaming seismogram and microbarometer data.
DA begins with a generalized state space model (SSM):

\[
x_t = f_t(x_{t-1}, v_t), \quad v_t \sim q(\cdot) \tag{2}
\]

The system model (Eq. (2)) predicts the state \( x_t \) at a time \( t \), given that simulation \( f_t \) was applied to state \( x_{t-1} \) at the previous time \( t-1 \). The system noise \( v_t \), which follows the distribution \( q(\cdot) \), absorbs the “modeling error”, which measures the degree to which the simulation model cannot explain the observed phenomena. The observation model (Eq. (3)) compares observable quantities, which are extracted and/or computed from the state \( x_t \) by applying an observation operator \( h_t \), with the observation data point \( y_t \) taken into consideration the observation noise \( w_t \), which follows the distribution \( r(\cdot) \). Since the present DA system adopts the normal mode summation technique, a linear computation, the SSM can be rewritten in a linear form:

\[
x_t = F_t x_{t-1} + G_t v_t \tag{4}
\]

\[
y_t = H_t x_t + w_t \tag{5}
\]

The observation noise is assumed to follow a normal distribution with mean zero and covariance \( \text{diag}(\sigma_w^2) \), i.e., \( w_t \sim N(0, \text{diag}(\sigma_w^2)) \), where \( \sigma_w^2 \) is the variance of the observation noise at the \( k \)-th observatory. The state and observation vectors are

\[
x_t = (\delta p_{1,1,t}^{\text{sim}}, \cdots, \delta p_{1,N,t}^{\text{sim}}, \delta p_{2,1,t}^{\text{sim}}, \cdots, \delta p_{N,N,t}^{\text{sim}})' \tag{6}
\]

\[
y_t = (\delta p_{1,1,t}^{\text{obs}}, \cdots, \delta p_{N,N,t}^{\text{obs}})' \tag{7}
\]

where the apostrophe denotes transposing a matrix, \( \delta p_{n,k,t}^{\text{sim}} \) is the waveform responding to the \( n \)-th hypocenter (referred to as “the \( n \)-th subevent" hereinafter) to be observed at the \( k \)-th observatory, and \( \delta p_{n,k,t}^{\text{obs}} \) is the waveform that is observed at the \( k \)-th observatory. The system matrix \( F_t \), which represents the simulation model, contains information about the normal modes, the associated eigenfunctions, and the CMTs. The present paper does not consider the system noise \( v_t \) since our current purpose is to evaluate the estimation capability of the proposed DA system. Since a sequence of multiple subevents along a fault is just a model for the sake of convenience, the components of the state variable \( x_t \), which are the responses to each subevent, are unobservable. What is observable is a linear sum of the responses weighted by the moment magnitude of each subevent, thus the observation matrix \( H_t \) is designed as

\[
H_t = (m_1 B_1(t_w) \cdots m_N B_N(t_w)) \tag{8}
\]

where \( m_n \) (\( n = 1, \cdots, N \)) is the moment magnitude of each subevent. \( B_n(t_w) \) is a time-delay operator, which acts on an arbitrary time series \( z(t) \) as

\[
B_n(t_w) z(t) = z\left(t - t_w - \frac{nL}{(N-1)V}\right) \tag{9}
\]

where \( L \) is the length of the fault and \( V \) is the rupture velocity, i.e., \( B_n(t_w) \) expresses the sequence of exploding subevents. The parameter \( t_w \) is a time shift to account for...
background wind, which sometimes affects the propagation of seismoacoustic waves.

B. Parameter Estimation

Fig. 3 illustrates the workflow of the proposed DA system. The model parameters to be estimated in the procedure are summarized in the following parameter vector;

$$\theta = (C', m', t, \sigma^2)'$$ (10)

where $C = (C_0, C_0, C_0)'$ is the location of a hypocenter with latitude $C_0$, longitude $C_0$, and depth $C_0$, $m = (m_1, \cdots, m_N)'$ contains the moment magnitudes of the subevents, and $\sigma^2 = (\sigma_1^2, \cdots, \sigma_K^2)'$ contains the variance of the observation noise at each observatory. Hence, the number of model parameters is $\dim \theta = N + K + 5$. From a list of past earthquakes, the DA system obtains the source parameters related to the earthquake event selected by the user; these parameters include the location of the hypocenter $C_{\text{model}}$, CMT $M_{\text{model}}$, the length of the fault $L_{\text{model}}$, and the rupture velocity $V_{\text{model}}$. The subscript “model” means the proposed value of the parameter in a model developed by seismic inversion. The system sets the following initial values for the MCMC procedure to estimate the parameters;

$$\theta_{\text{model}} = (C_{\text{model}}', m_{\text{model}}', V_{\text{model}}', 0, 1^2)'$$ (11)

where $1^2$ denotes a vector, all of the components of which are $1^2$. Each of the moment magnitudes of the $N$ CMTs is $m_{n,\text{model}} = M_{\text{model}}/N$ ($n = 1, \cdots, N$), i.e., $M_{\text{model}}$ is divided equally and the sections are located along the fault at equally-spaced intervals. Summarizing Eqs. (4)–(9) and setting an appropriate prior distribution $p(\theta)$, the posterior distribution can be defined as

$$p(\theta | \delta P_{\text{obs},K,t:T}) \propto p(\theta) \prod_{t=1}^{T} \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi \sigma_k^2}} \exp \left[- \frac{\left( \delta p_{\text{obs},t} - \sum_{n} m_{n} \delta p_{n,k,t-1,t}^{\text{sim}} \right)^2}{2\sigma_k^2} \right]$$ (12)

where $\delta p_{n,k,t}$ on the right-hand side is the mean of one-step-ahead prediction conditioned on the observation data in the period from 1 to $t - 1$, that is, assuming we are given $\delta p_{n,k,t-1}$. Since $\delta p_{n,k,t}$ in Eq. (12) is a distribution, we would normally assume that the notation should distinguish it from the one that appeared in Eq. (6). However, this distinction is not needed when the system noise is neglected, such as in the present case, because instead of the filtering step, it is sufficient to compare the simulated state variable $\delta p_{n,k,t}^{\text{sim}}$ with the observed value $\delta p_{n,k,t}^{\text{obs}}$. Eq. (12) compares the observed waveform $\delta p_{n,k,t}^{\text{obs}}$ with the simulated waveform, which is a linear sum of $\delta p_{n,k,t}^{\text{sim}}$, weighted by the moment magnitude $m_n$, and stacked by considering the rupture velocity $V$.

IV. VALIDATION OF THE DATA ASSIMILATION SYSTEM THROUGH A TWIN EXPERIMENT

A newly developed DA procedure is often tested by a twin experiment, which determines if the DA procedure can reproduce the true solution from observation data that were synthesized by a numerical simulation of an assumed true solution. We therefore performed a twin experiment in order to validate the data assimilation system for seismoacoustic waves. A sequence of five subevents is located along a 40 km fault, and each subevent is assumed to generate an earthquake. The perturbations in atmospheric pressure due to seismoacoustic waves excited by the earthquakes are observed at Sugadaira and Shionomisaki observatories.

Fig. 4. Synthetic earthquake that was assumed in the twin experiment in order to validate the data assimilation system for seismoacoustic waves. A sequence of five subevents is located along a 40 km fault, and each subevent is assumed to generate an earthquake.
where the observation noise that follows the normal distribution; the moment magnitudes of the subevents, with additional as a linear sum of these simulated waveforms, weighted by real observations (Fig. 1(a)) are clearly reproduced. Synthetic true value for each of the parameters. Fig. 7(a) shows the of a synthetic earthquake. Despite the intentionally bad initial mechanisms developed by universities and research institutes, prior distribution stationary state after approximately 25,000 steps. Although a method in the MCMC, indicating that the sampling attains a experiment are the depth of the hypocenter \( \hat{\theta} \), the rupture velocity \( V \), the moment magnitudes of the subevents \( \mathbf{m} \), and the variances of the observation noise \( \sigma^2 \).

Fig. 5 shows the theoretical atmospheric perturbations \( \delta p_{\text{obs}} \) that correspond to each subevent, computed by the normal mode summation technique. The two large wave packets, the Rayleigh waves and acoustic waves, that were found in the real observations (Fig. 1(a)) are clearly reproduced. Synthetic observation data used in the twin experiment were obtained as a linear sum of these simulated waveforms, weighted by the moment magnitudes of the subevents, with additional observation noise that follows the normal distribution;

\[
\delta p_{\text{obs}} = \sum_{n=1}^{N} \sigma_k \tilde{p}_{n,k,t} + w_{k,t} \quad w_{k,t} \sim N(0, \sigma_{k,0}^2) \quad (13)
\]

where \( \sigma_{k,0}^2 \) is the variance of the observation noise. The present paper assumes \( \sigma_{1,0}^2 = 20^2 \text{ Pa}^2 \) and \( \sigma_{2,0}^2 = 50^2 \text{ Pa}^2 \), which means that the noise level at the coastal Shionomisaki observatory is larger than that at the Sugadaira observatory in the mountains. The synthetic observation data are shown in the bottom subplot of Fig. 5.

Fig. 6 shows the sampling process using the Metropolis method in the MCMC, indicating that the sampling attains a stationary state after approximately 25,000 steps. Although a prior distribution \( p(\theta) \) could be constructed based on source mechanisms developed by universities and research institutes, we have assumed a uniform distribution because a prior distribution cannot, of course, be obtained in such the case of a synthetic earthquake. Despite the intentionally bad initial values, the marginalized posterior distributions were near the true value for each of the parameters. Fig. 7(a) shows the marginalized posterior distribution for each model parameter \( \hat{\xi}(\delta p_{1:K,1:T}) \), where \( \xi \) is one of the model parameters contained in \( \theta \) and \( \hat{\xi} \) indicates the unnormalized probability density function. The posterior distribution was estimated from 10,000 samples chosen from the MCMC samples obtained from 30,005–80,000 steps at intervals of 5 steps (Fig. 6). Table 1 summarizes the true value \( \xi_0 \) and the initial value \( \xi_{\text{model}} \) for each model parameter, the variance of the proposal density used in the MCMC, the covariance of which is assumed to be diagonal. The mean, standard deviation, and mode of the estimated marginalized posterior distribution are also shown. Since the present simulation code is only capable of computing the cases of the depth of the hypocenter at intervals of 1.25 km and it converges much more rapidly than the other parameters, the obtained marginalized posterior distribution has the shape of the Dirac delta function. This rapid convergence in the depth of the hypocenter is due to the extreme sensitivity of the variation in the atmospheric pressure to the depth of the hypocenter [3]. Finally, Fig. 7(b) compares the synthetic observational waveform \( \delta p_{1:K,1:T} \) and the theoretical waveform \( \delta p_{\text{K,MAP}} \), which was reproduced from the maximum-a-posteriori (MAP) model parameter set that attained the posterior distribution maximum. The MAP solution clearly explains the observations well, and thus proves that the proposed DA system for seismoacoustic waves performs properly.

V. C ONCLUSION

The present paper introduced the DA system for seismoacoustic waves. This system estimates the model parameters related to the source mechanism, and integrates a numerical simulation based on the normal mode summation technique.
accuracy and speed of the DA computations. Nonetheless, we need to be solved, such as improving of the microbarometer observation networks, making the system sufficiently robust to sustain large ground shakes or electrical outages, and improving the observed microbarometer data. A validation by a twin experiment, which assumed a very large earthquake in the Pacific Ocean off the coast of West Japan being predicted to occur in the near future, has ensured that the DA system can successfully reproduce the true model parameters related to a sequence of subevents by using the MCMC method.

The computation time needed for this DA technique, which was approximately one hour for the twin experiment, should be reduced to 10% or less, considering that there will need to be retrieval and initial handling of the observation data was approximately one hour for the twin experiment, should be reduced to 10% or less, considering that there will need to be retrieval and initial handling of the observation data. A valida

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Observed Value</th>
<th>True Value</th>
<th>Standard Deviation of Proposal Density for MCMC</th>
<th>Mean of Estimated Marginalized Posterior Distribution</th>
<th>Standard Deviation of Estimated Marginalized Posterior Distribution</th>
<th>Mode of Estimated Marginalized Posterior Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{1}$</td>
<td>10.00</td>
<td>10.00</td>
<td>0.00</td>
<td>0.00</td>
<td>10.00</td>
<td>10.00</td>
</tr>
<tr>
<td>$m_1$</td>
<td>0.150</td>
<td>0.150</td>
<td>0.00</td>
<td>0.00</td>
<td>0.150</td>
<td>0.150</td>
</tr>
<tr>
<td>$m_2$</td>
<td>0.250</td>
<td>0.250</td>
<td>0.00</td>
<td>0.00</td>
<td>0.250</td>
<td>0.250</td>
</tr>
<tr>
<td>$m_3$</td>
<td>0.350</td>
<td>0.350</td>
<td>0.00</td>
<td>0.00</td>
<td>0.350</td>
<td>0.350</td>
</tr>
<tr>
<td>$m_4$</td>
<td>0.200</td>
<td>0.200</td>
<td>0.00</td>
<td>0.00</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>$m_5$</td>
<td>0.050</td>
<td>0.050</td>
<td>0.00</td>
<td>0.00</td>
<td>0.050</td>
<td>0.050</td>
</tr>
<tr>
<td>$V$</td>
<td>2.00</td>
<td>2.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>0.151</td>
<td>0.151</td>
<td>0.00</td>
<td>0.00</td>
<td>0.151</td>
<td>0.151</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>0.254</td>
<td>0.254</td>
<td>0.00</td>
<td>0.00</td>
<td>0.254</td>
<td>0.254</td>
</tr>
<tr>
<td>$\sigma_3$</td>
<td>0.352</td>
<td>0.352</td>
<td>0.00</td>
<td>0.00</td>
<td>0.352</td>
<td>0.352</td>
</tr>
<tr>
<td>$\sigma_4$</td>
<td>0.200</td>
<td>0.200</td>
<td>0.00</td>
<td>0.00</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>$\sigma_5$</td>
<td>0.050</td>
<td>0.050</td>
<td>0.00</td>
<td>0.00</td>
<td>0.050</td>
<td>0.050</td>
</tr>
</tbody>
</table>

The authors appreciate Profs. Ichiro Tomizawa, Taichi Hayashi, Toshikiko Iyemori, and Masaki Kanao for their cooperation in the establishment of the infrasound observations at Sugadaira and Shionomisaki. Theoretical simulations of seismoacoustic wave propagations were carried out by using a code provided by Dr. Naoki Kobayashi on the Supercomputer System for Statistical Science of the Institute of Statistical Mathematics, Japan. This work is supported by the Grant-in-Aid “Growing Fusion Research Project” of the Transdisciplinary Research Integration Center, Research Organization of Information and Systems, Japan. The map used in this paper was generated by Generic Mapping Tools (GMT).

**REFERENCES**


