Spatial and temporal pattern of global seismicity extracted by dimensionality reduction

Akihisa Okada, Mitsuhiro Toriumi, and Yoshiyuki Kaneda

***Abstract*—**Constructing a predictive model of earthquake activity is difficult due to the complexity of mechanisms and variety of components. We do not know even what effective features are involved in seismic processes. Thus, we adopted a statistical approach for extracting features of global seismicity. We extracted the features from an open earthquake data catalog by using principal component analysis to reveal the spatial linkages and time dependence of earthquake activity. For principal component analysis, we defined earthquake occurrence rate and regarded its time series as samples and regional labels as the dimensionality. We demonstrate that this method accurately identifies past earthquake activity and reveals correlations among remote locations and time dependence of seismicity features. We also discussed the normality of the sample distribution to confirm the validity of the principal component analysis. These results will help the construction of a predictive earthquake activity model.

***Keywords*—**Global seismicity, feature extraction, dimensionality reduction, principal component analysis

I. Introduction

We can predict and control the behavior of a system by revealing the relations between some features of the system or their time dependence. Examples of such features include macroscopic physical quantities of materials such as their Young’s modulus and magnetic permeability. We can predict the behavior of materials by using a few of these features. On the other hand, from a microscopic viewpoint, several grams of a material consist of Avogadro’s number of atoms. Thus, we need an enormous number of quantities to describe the exact states of the material. In statistical physics, this “gap problem” concerning the number of quantities is resolved by reducing the number of features considered. In concrete terms, macroscopic thermodynamic features are extracted from microscopic kinetic behavior under the assumption that the states of materials follow statistical distributions.

A similar framework will be a promising candidate for solving problems in complex systems wherein the behavior is difficult to predict owing to the variety of the system’s components. An example of such a complex problem is earthquakes [1]. Massive earthquakes kill many people and cause economic losses. If we could predict when earthquakes would occur, the resulting damage would be mitigated. In the case of earthquakes, the nature of the “macroscopic” features is unknown, but we do have observational data that correspond to “microscopic” features.

So far, data-driven approaches have been adopted for revealing some aspects of earthquake activity [2]-[5]. For example, it has been reported that the epidemic-type aftershock sequence (ETAS) model, which is based on a stochastic point process, is effective for predicting the probability of aftershock occurrence [6], [7]. On the other hand, for constructing a physical model and predicting seismicity, linear principal component analysis is employed as a method of dimensionality reduction for extracting features of earthquake activity from observed data [8]-[10].

In this study, we apply this method of dimensionality reduction for global activity of earthquakes and discuss its spatial and temporal features to construct a physical model.

II. Data

We used the open data catalog of earthquakes published on the web by the United States Geological Survey and analyzed earthquakes that met the criteria shown in Table I.

Table I Selection Criteria of Data for Analysis

Period	1990/1/1-2016/12/31
Magnitude	Greater than or equal to 4, less than 5
Depth of epicenter	0-200 km

As indicated, we only considered earthquakes of relatively low magnitude, of which there were many examples, and we excluded deep-focus earthquakes because their mechanism is complex, and different from that of other earthquakes.

A. Okada is with the Toyota Central R&D Labs., Inc., Aichi, Japan (phone: +81-561-71-7293; e-mail: a-okada@mosk.tytlabs.co.jp).

M. Toriumi is with the Japan Agency for Marine-Earth Science and Technology, Tokyo, Japan (e-mail: tori@jamstec.go.jp).

Y. Kaneda is with the Kagawa University, Kagawa, Japan (e-mail: kanedaykg@cc.kagawa-u.ac.jp).

The total number of relevant earthquakes was 208,614. We treated space and time as discrete variables and divided lines of longitude and latitude into 60 sections; in other words, we divided the surface of the earth into 3600 areas. The integrated number of relevant earthquakes over the whole period is shown in Fig. 1 in an equirectangular projection map. It was confirmed that the main contribution originated from plate boundary earthquakes.

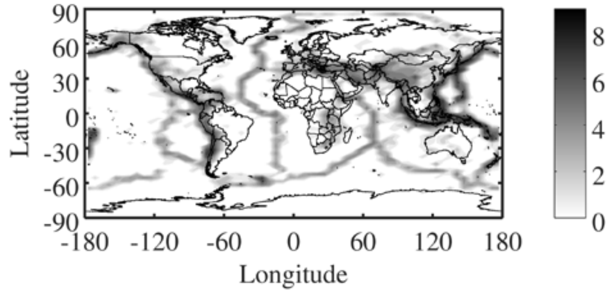


Fig. 1 Integrated earthquake count *N* in divisions of the earth's surface shown as value of ln(1+*N*)

III. METHODS

A. Earthquake Occurrence Rate

First, for extracting features of global earthquakes, we used the earthquake occurrence rate, which represents the average earthquake count per unit area and unit time. Here we defined the variables S_k, T, and $N_k(t)$. S_k and T represent the area of region k and a discrete time interval, respectively. $N_k(t)$ represents the number of earthquakes that occurred in region k during the time interval from t to $t + T$.

Then, earthquake occurrence rate $X_k(t)$ in region k at time t is derived as follows:

$$X_k(t) = \frac{N_k(t)}{S_k T}. \quad (1)$$

It is noted that the area S_k depends on the latitude θ in a strict sense. Assuming that the earth is approximated as a perfect sphere, this dependence is proportional to cos θ in the same way as a small region on a unit sphere in polar coordinates.

However, earthquakes happen to a lesser extent in high latitude regions. Thus, the following results, which ignore the dependence on latitude, do not differ greatly from cases in which latitude is considered.

B. Principal Component Analysis

For extracting features of global earthquakes, we adopted linear principal component analysis as a method of dimensionality reduction. We revealed the spatial and temporal features of earthquakes by regarding the time series of $X_k(t)$ values as samples and the regional labels as the dimensionality.

We defined m and n as the total number of regional labels and discrete time intervals. Then, we defined the $m\times n$ matrix X, of which the rows are regional labels and the columns are discrete time intervals, as follows:

$$X = \begin{pmatrix} X_1(t_1) & X_1(t_2) & \cdots & X_1(t_n) \\ X_2(t_1) & X_2(t_2) & \cdots & X_2(t_n) \\ \vdots & \vdots & \ddots & \vdots \\ X_m(t_1) & X_m(t_2) & \cdots & X_m(t_n) \end{pmatrix}. \quad (2)$$

In the usual method of linear principal component analysis, we obtained eigenvectors u from the $m\times m$ covariance matrix $\Sigma = 1/m \cdot XX'$ by eigenvalue decomposition.

For the purpose of dimensionality reduction, when the eigenvalues were arranged in descending order we adopted the minimum number $j_{\min}$ of eigenvalues that meet the following criterion:

$$\frac{\sum_{l=1}^{j_{\min}} \lambda_l}{\sum_{i=1}^{m} \lambda_i} > c. \quad (3)$$

Here, c is a constant used to determine the ratio of retained variance of the data such that $0 < c < 1$.

Furthermore, by ordering $j_{\min}$ eigenvectors in columns, we defined the $m\times j_{\min}$ matrix U. Then, by transforming X using U, we obtained the new matrix Z, which consists of principal components, that is, new features of global earthquakes.

$$Z = U' X \quad (4)$$

The original $m\times n$ matrix X was compressed to the $j_{\min}\times n$ matrix Z. The rows of Z represent new spatial features, whereas the columns still represent time dependence.

The spatial features of Z are visualized by drawing U, because U connects real space with the new feature space.

The parameters used for analysis in this study are summarized in Table II.

Table II Parameters Used in Analysis

Name	Meaning	Value
m	Region of the world	60(lat.)×60(long.)
T	Discrete time interval	1 month
c	Criterion of retained variance	0.8

IV. RESULTS

A. Whole period

First of all, we show the dependence of the retained variance on the number of component in Fig. 2. When we adopted 16 principal components, 80% of the variance was retained. It was confirmed that the dimensionality was greatly reduced from the original 3600 dimensions.

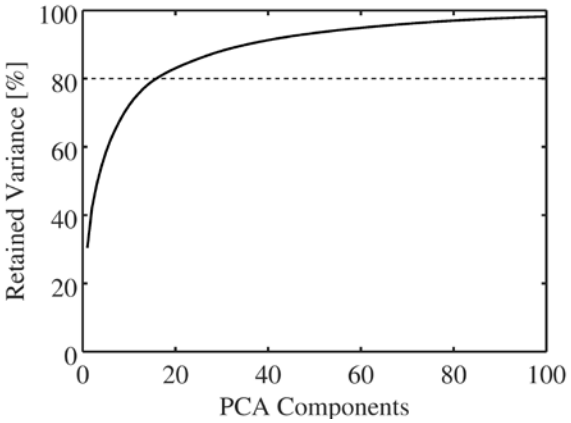


Fig. 2 Dependence of retained variance on number of principal components up to 100. The dashed line represents the criteria that retained 80% variance.

Next, we show the spatial and time dependence of the principal components. By plotting U_{ij} in the real space region i, we can visualize the jth component of new features, as shown in Fig. 3 and 4. The radius of the marker corresponds to the intensity for that location.

We also show the time dependence of principal components in Fig. 5, 6, and 7. For the sake of visibility, we plotted the square root of the intensity.

From the above results, we concluded that the first component corresponds to the 2011 Tohoku Pacific Ocean earthquake (M9.0). We also found that the second component corresponds to the global background activity, the third component corresponds to the 2004 and 2005 Sumatra earthquakes (M9.1, M8.6), and the fourth component corresponds to the 2007 Kuril Islands offshore earthquake (M8.2). These components represent single earthquake events.

On the other hand, some components have broad distributions or multiple locations. For example, the tenth component seems to show a correlation between earthquake activity in the Solomon Sea and around the international boundaries between Peru, Bolivia, and Chile, as shown in Fig. 4. This correlation seems to occur periodically, as shown in Fig. 6 (d).

Here, we show the sample distribution. Since the earthquakes included in our study are almost all plate boundary type earthquakes, it is expected that the distribution of sample data has a peak at the plate boundary in real space. For the validity of principal component analysis, it is desired that the sample data follow a normal distribution. We show the distribution of the background component, namely the second component in Fig. 8. Here, the x-axis shows the intensity normalized by the maximum absolute value and divided into 50 parts. The y-axis shows data counts for the corresponding normalized intensity. This distribution can be approximated by a normal distribution because the data has one peak and seems to have finite variance. Thus, it was reasonable that we adopted the principal component analysis for global seismicity.

B. *Time dependence of principal components*

In the previous section, we analyzed past data and confirmed that this method can identify some massive earthquakes correctly. For the purpose of predicting earthquakes, we must detect some signal in advance. Hence, we investigated the time dependence of the principal components.

In particular, we focused on the first principal component. As mentioned before, this component corresponds to the 2011 Tohoku Pacific Ocean earthquake. Thus, we divided the entire period into three parts; the first, second, and third parts consisted of the data from 1991 to 2000, from 2001 to 2010, and from 2012 to 2016, respectively. Furthermore, we added the data from 1981 to 1990. We excluded the 2011 data in order to make a before and after comparison for the 2011 Tohoku Pacific Ocean earthquake.

We defined the index of seismicity with the following equation:

$$Y_i = \sum_{k=1}^{j_{\min}} U_{ik} \ . \qquad (5)$$

Y_i represents the locations where we determined seismicity is active; these can be seen, in Fig. 9. From 1981 to 1990, earthquake activity was widespread, but it was concentrated on the western side of the Pacific Plate from 1991 to 2000, and activity around Japan was calm from 2001 to 2010. After that, the 2011 Tohoku Pacific Ocean earthquake occurred. From 2012 to 2016, there was a resurgence in widespread activity.

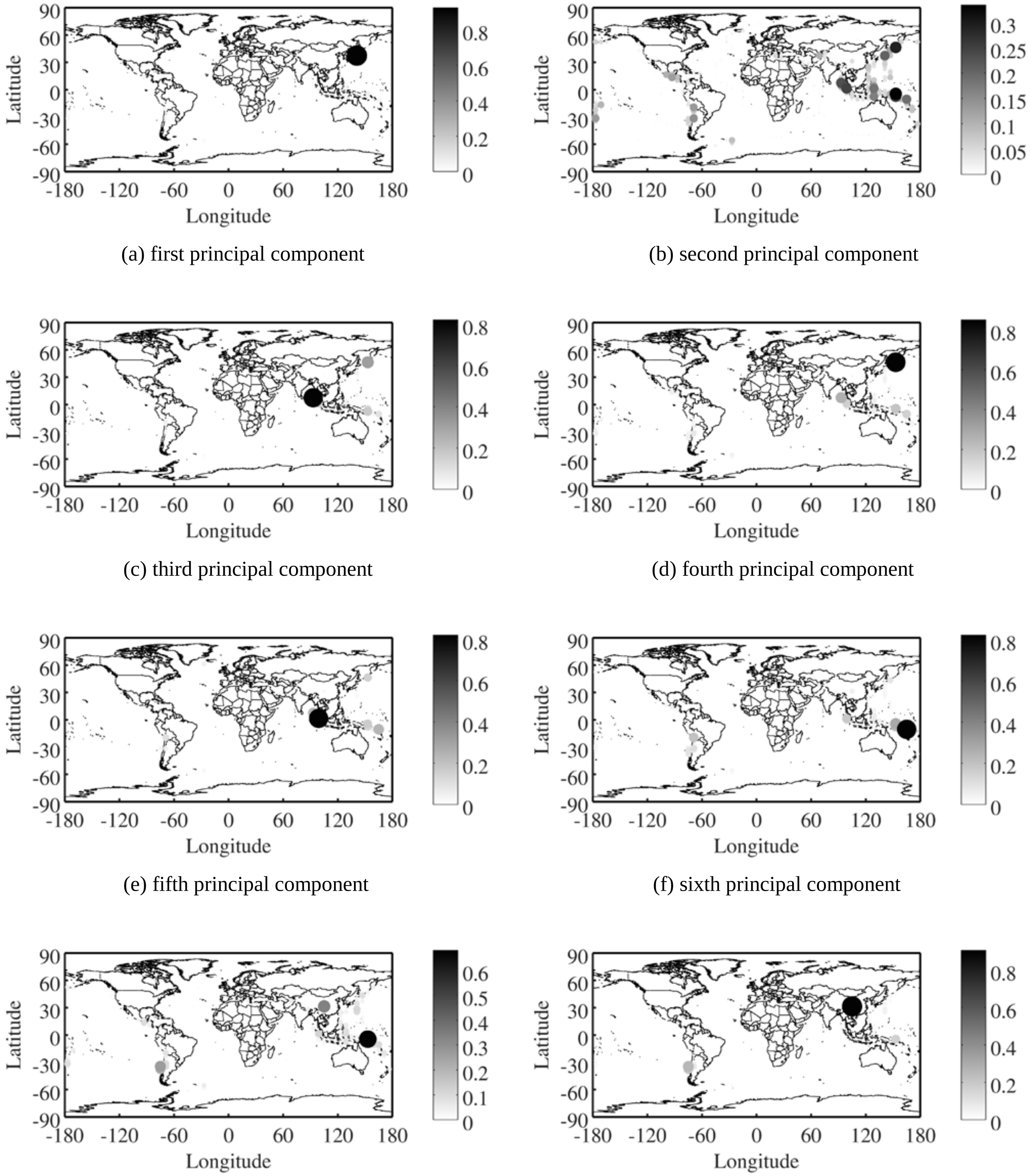
V. Discussion

The new features extracted effectively identified past major earthquakes, even though we only considered relatively low magnitude earthquakes. This is because the Gutenberg-Richter law holds locally. Furthermore, the method seems to show a correlation among earthquakes that occurred at remote locations as the tenth principal component. It is expected that we will be able to extract features with more local detail by applying this method not to the whole world but to a local area. However, we must point out that the correlation between earthquakes revealed in this study does not definitely originate from a physical cause. There is a possibility that this feature was extracted because of coincidences in the occurrence of earthquakes. Thus, for understanding and predicting the activity of earthquakes, we must investigate this feature by collecting data and providing a physical rationale.

The time dependence of the index of seismicity shows periods of concentrated activity, followed by quiescent periods. It is considered that energy for generating massive earthquakes is conserved by this quiescence. More detailed analysis in which a limited region is surveyed must be conducted to reveal the signals for individual earthquakes.

VI. CONCLUSION

In this study, we extracted new features of global earthquakes using linear principal component analysis from observational data published on the web. Some past massive earthquakes appeared as principal components, even though we only considered relatively low-magnitude earthquakes. Furthermore, it is implied that there has been a correlation between South American and Solomon Sea earthquakes, which seems to be periodic. Before massive earthquakes, periods of concentrated activity followed by quiescence are confirmed by the time dependence in the index of seismicity Y_i. In future work to predict earthquake activity, we must investigate whether or not these new features originate from physical causes.



(a) first principal component

(b) second principal component

(c) third principal component

(d) fourth principal component

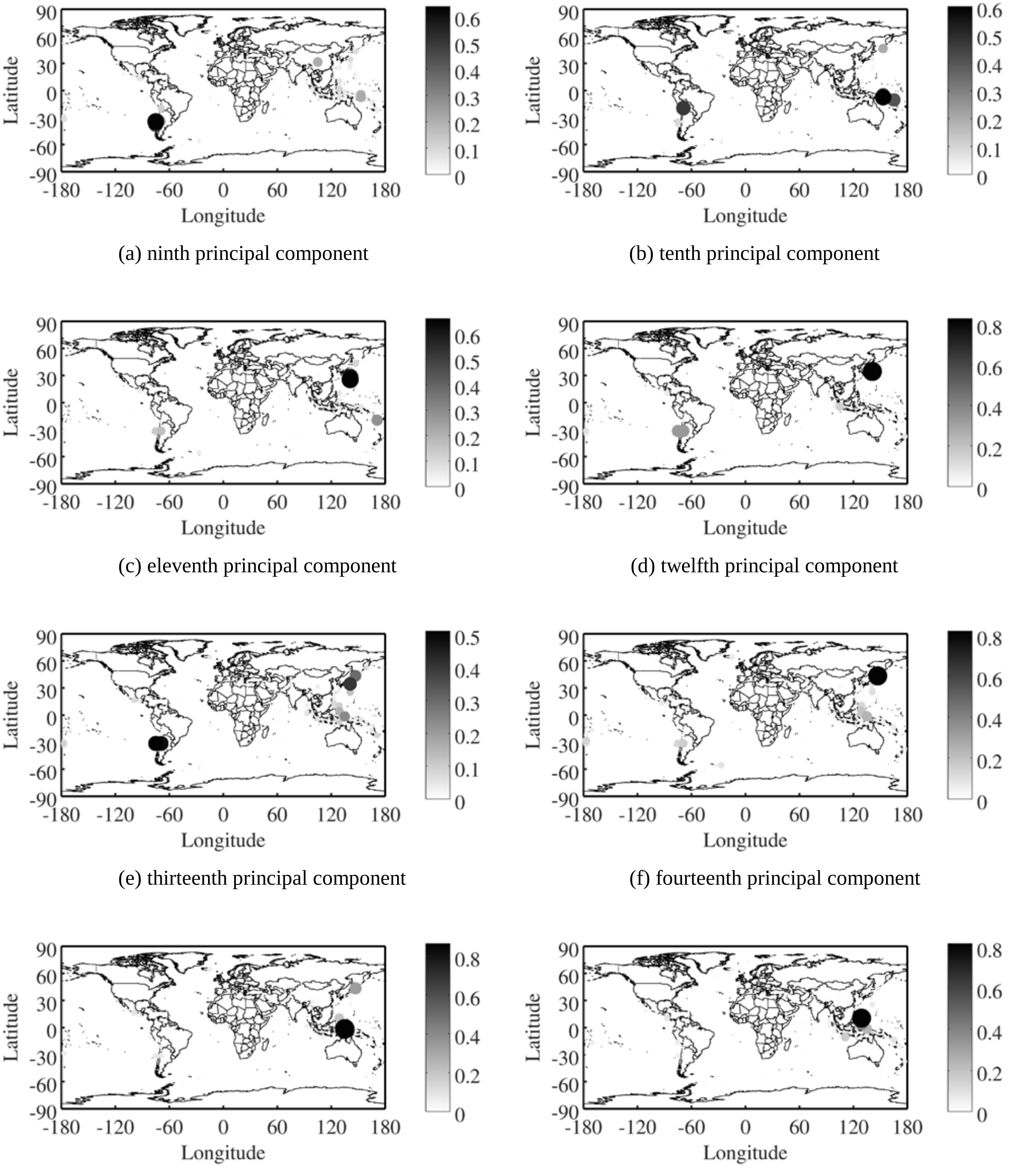
(e) fifth principal component

(f) sixth principal component

(g) seventh principal component

(h) eighth principal component

Fig. 3 Spatial distribution of principal components from first to eighth



(a) ninth principal component

(b) tenth principal component

(c) eleventh principal component

(d) twelfth principal component

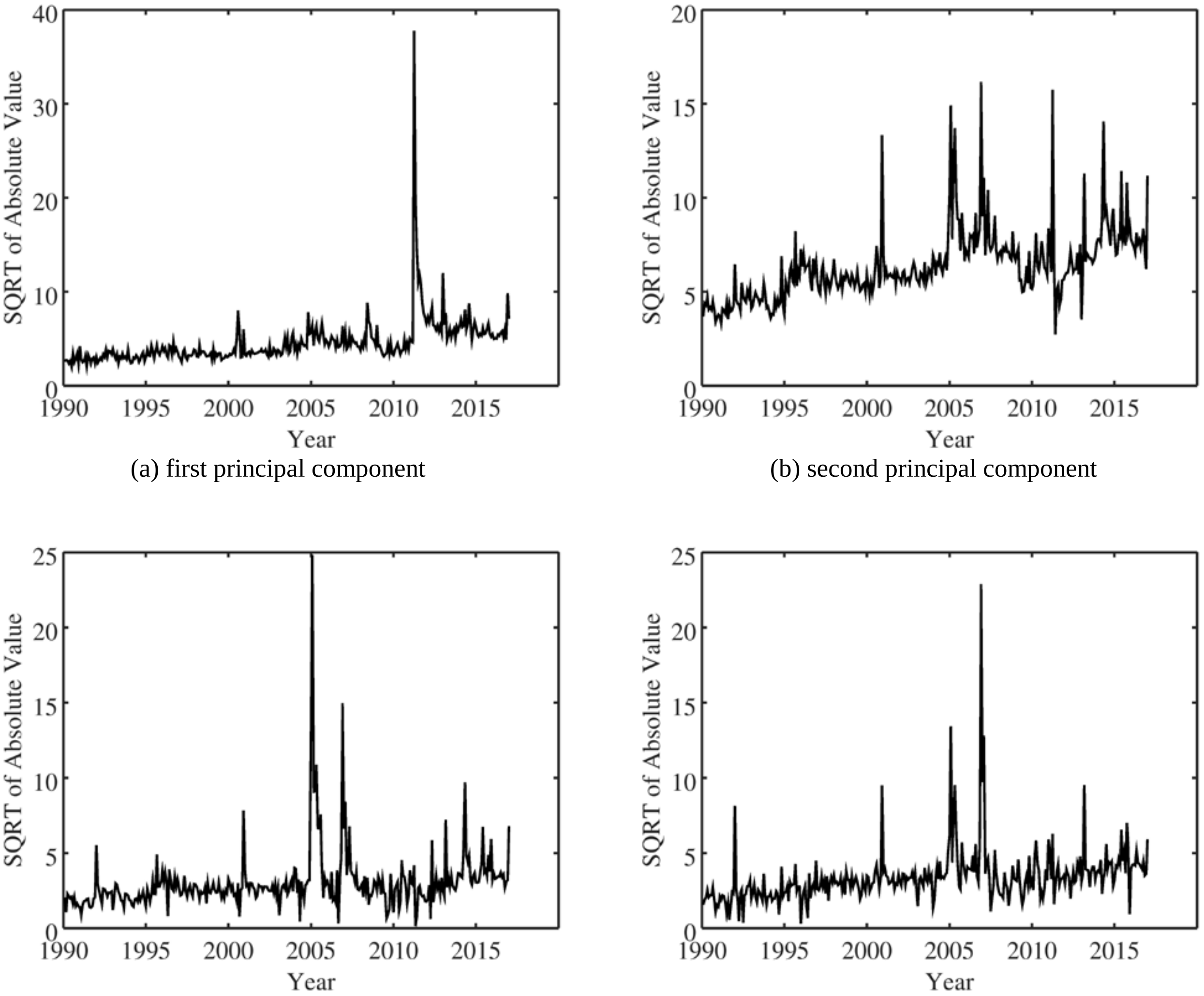
(e) thirteenth principal component

(f) fourteenth principal component

(g) fifteenth principal component

(h) sixteenth principal component

Fig. 4 Spatial distribution of principal components from ninth to sixteen

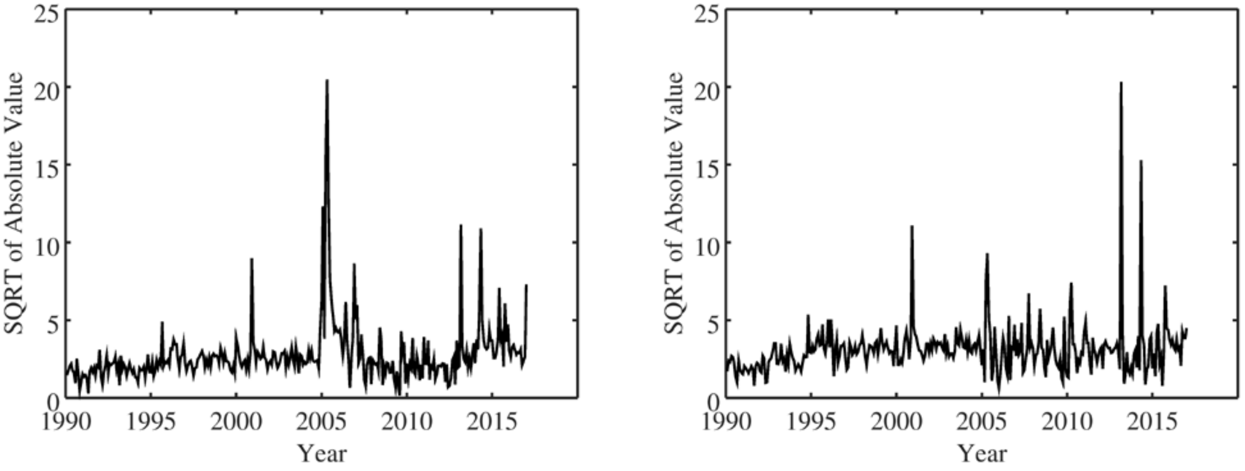


(a) first principal component

(b) second principal component

(c) third principal component

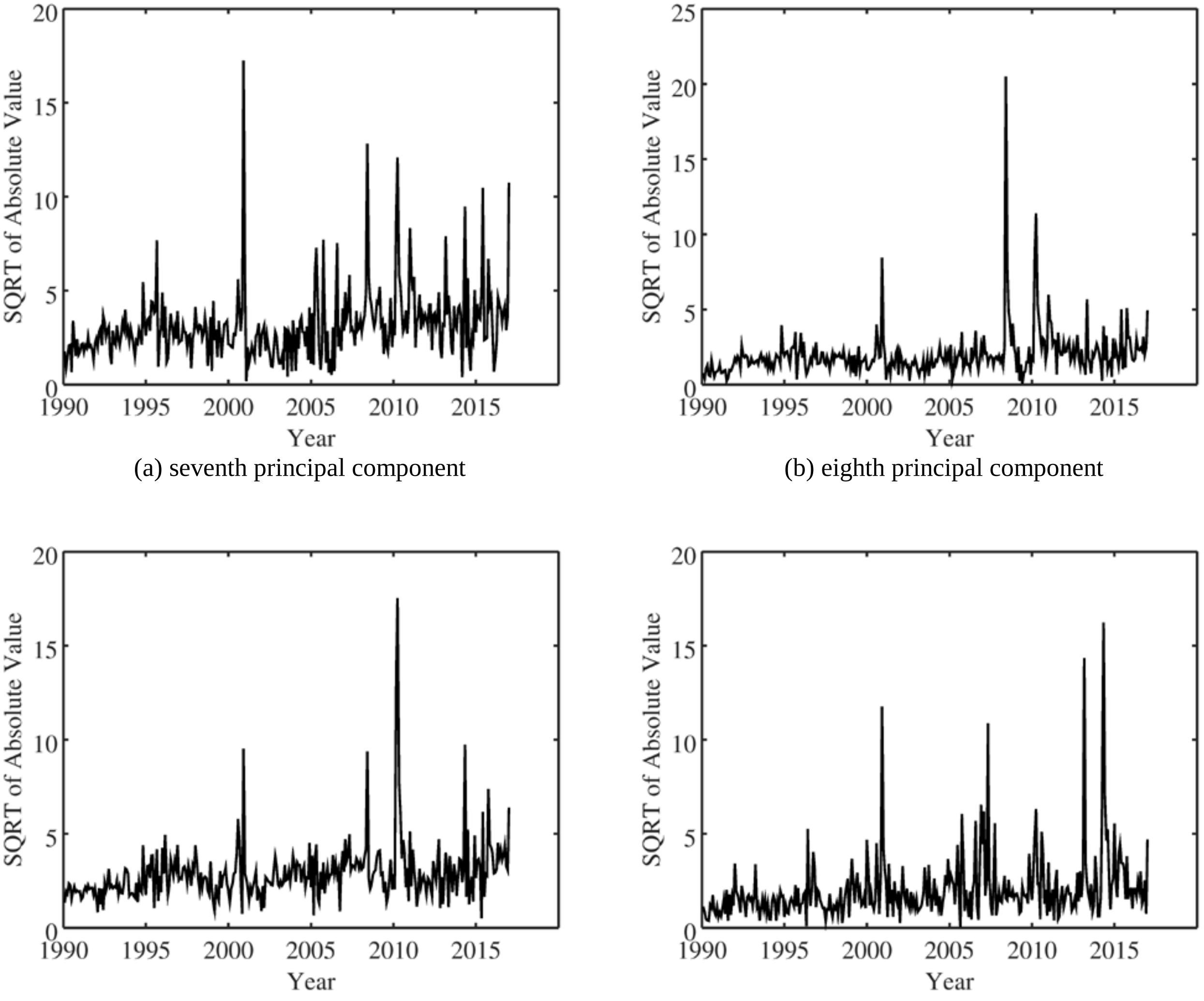
(d) fourth principal component



(e) fifth principal component

(f) sixth principal component

Fig. 5 Time dependence of principal components from first to sixth

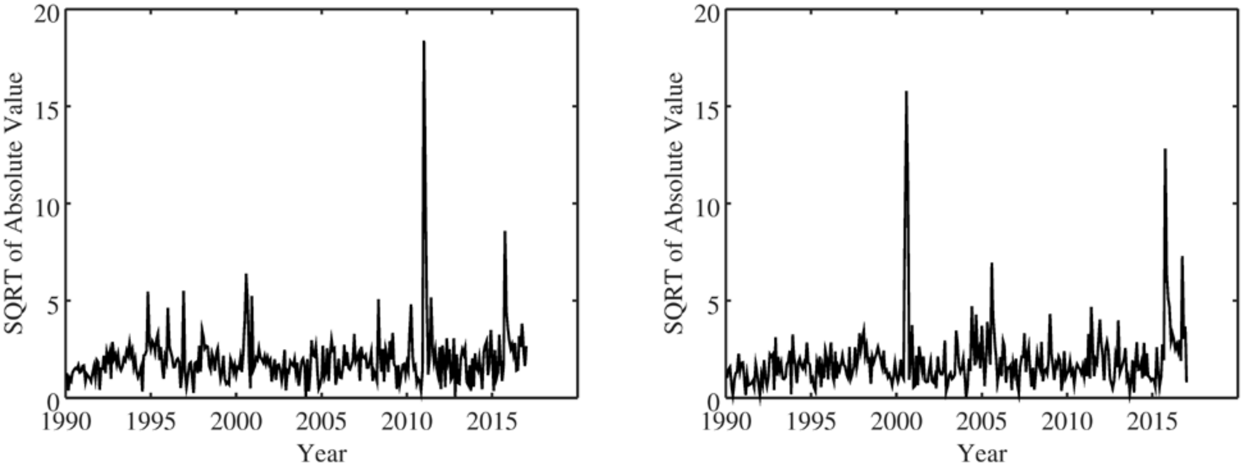


(a) seventh principal component

(b) eighth principal component

(c) ninth principal component

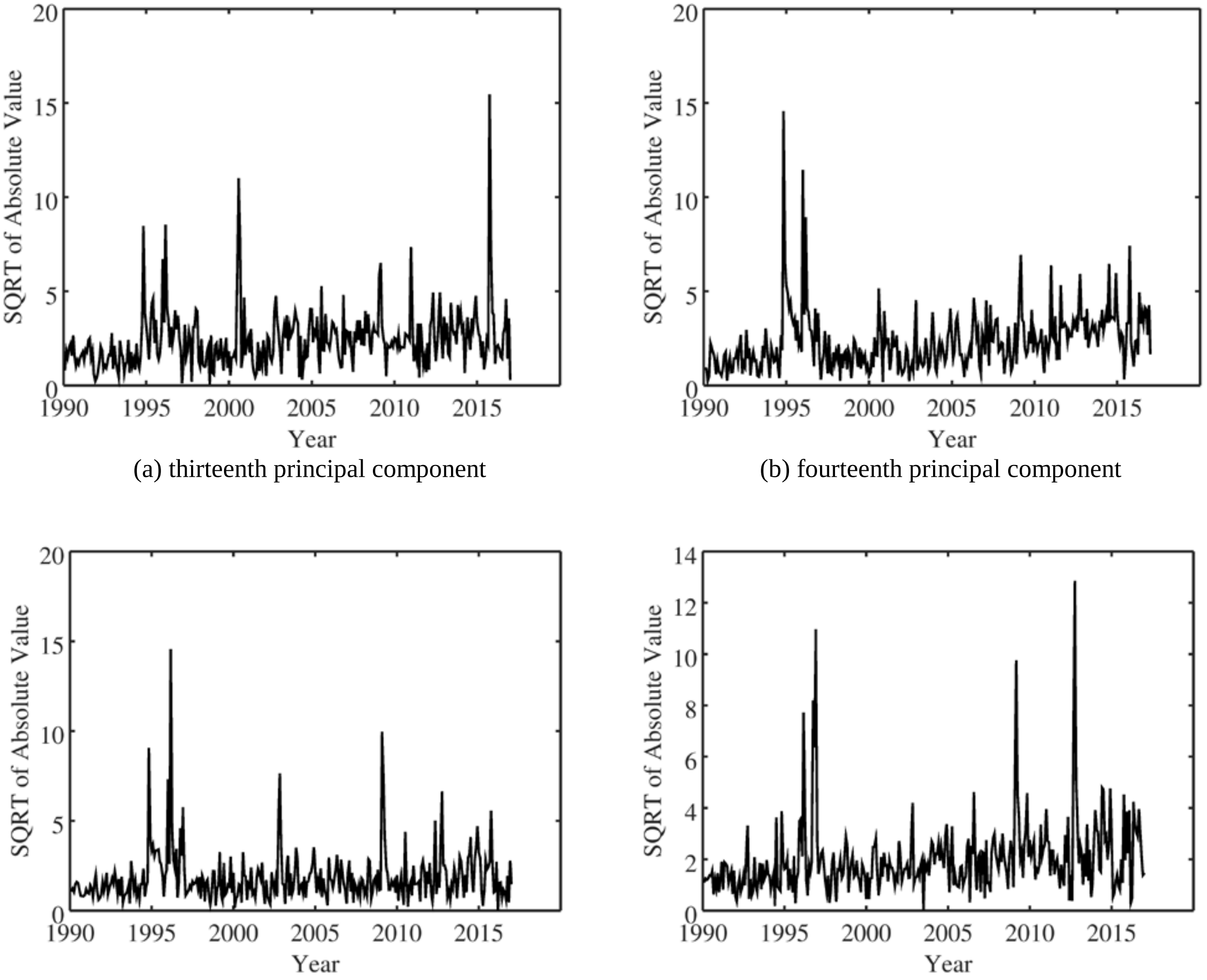
(d) tenth principal component



(e) eleventh principal component

(f) twelfth principal component

Fig. 6 Time dependence of principal components from seventh to twelfth



(a) thirteenth principal component

(b) fourteenth principal component

(g) fifteenth principal component

(h) sixteenth principal component

Fig. 7 Time dependence of principal components from thirteenth to sixteenth

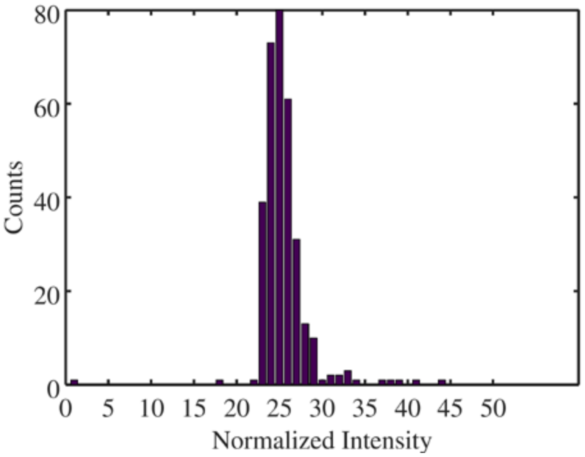


Fig. 8 data distribution of the background component

References

[1] B. Gutenberg and C. F. Richter, "Seismicity of the earth," Geological Society of America Special Papers, vol. 34, pp. 1–126, 1941.

[2] G. Ekström, "Global seismicity: results from systematic waveform analyses, 1976–2005," in Treatise on geophysics, vol. 4, no. 4.16, H. Kanamori, Ed., Amsterdam: Elsevier, pp. 473–481, 2007.

[3] T. Hori, S. Miyazaki, M. Hyodo, R. Nakata, and Y. Kaneda, "Earthquake forecasting system based on sequential data assimilation of slip on the plate boundary," Theoretical and Applied Mechanics Japan, vol. 62, pp. 179–189, 2014.

[4] R. Nakata, T. Kuwatani, M. Okada, and T. Hori, "Geodetic inversion for spatial distribution of slip under smoothness, discontinuity, and sparsity constraints," Earth, Planets and Space, 68:20, 2016.

[5] T. Kuwatani, K. Nagata, M. Okada, and M. Toriumi, "Data-driven approaches to petrology." in *Annual Meeting of the Geological Society of Japan The 120th Annual Meeting*, 2013, p.111.

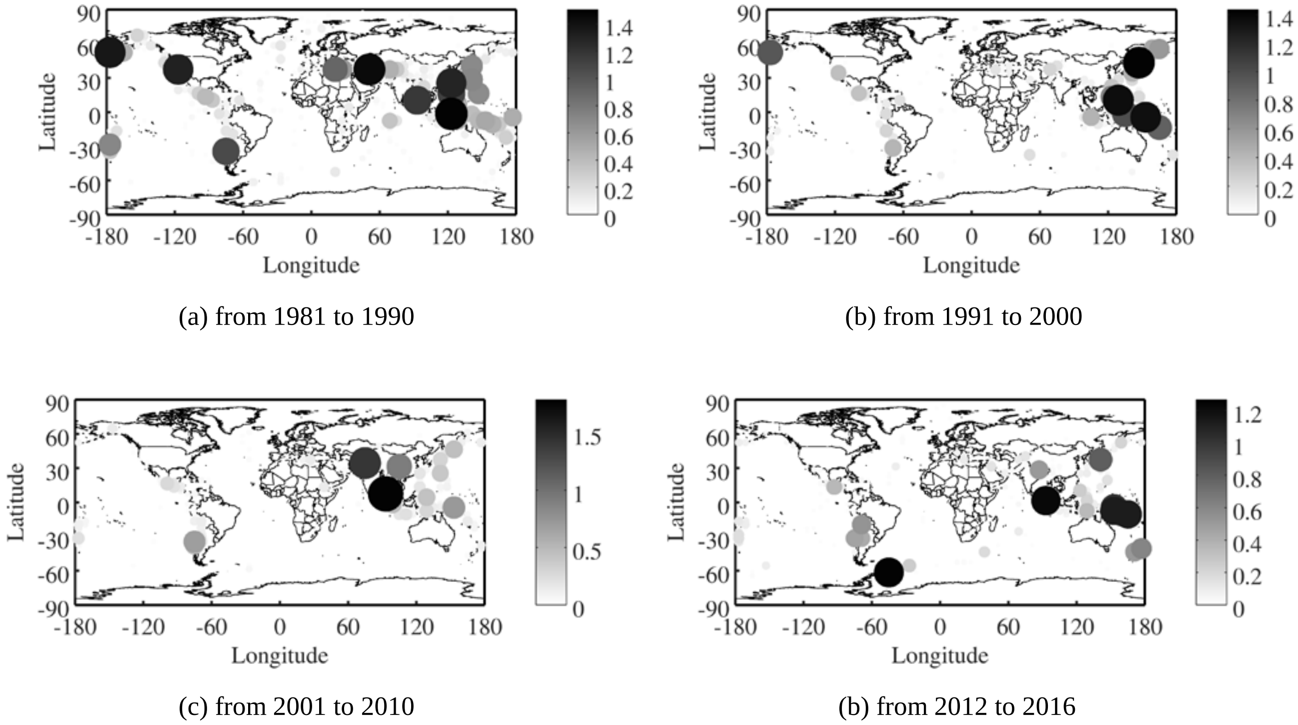
[6] Y. Ogata, "Statistical models for eearthquake occurrences and residual analysis for point processes," Journal of the American Statistical association, vol. 83, no. 401, pp. 9–27, 1988.

[7] Y. Ogata, "Seismicity analysis through point-process modeling: A review," Pure and applied geophysics, vol. 155, no. 2-4, pp. 471–507, 1999.

[8] M. Toriumi, "Principal component analyses of seismic activity in the plate boundary zone of northeastern japan arc," in Japan Geoscience Union Meeting Abstracts, 2009.

[9] M. Toriumi, "Dimension reduction analysis of microseismicity rate in sw japan," in Abstracts, Japan Geoscience Union Meeting, 2011.

[10] M. Toriumi, "Data-driven geochemistry in solid earth processes," Geochemistry, vol. 47, pp. 159–162, 2013, in Japanese.



(a) from 1981 to 1990

(b) from 1991 to 2000

(c) from 2001 to 2010

(b) from 2012 to 2016

Fig. 9 time dependence of locations of activity Y_i