von Specht, S. (2020): ICBM—Integrated Combined Baseline Modification: An Algorithm for Segmented Baseline Estimation. - Seismological Research Letters, 91, 1, 475-487.

https://doi.org/10.1785/0220190134

ICBM - Integrated Combined Baseline Modification: An Algorithm for Segmented Baseline Estimation

Sebastian von Specht*[1,2]

Abstract
Accelerograms are the primary source for characterizing strong ground-motion. It is therefore of paramount interest to have high-quality recordings free from any nonphysical contamination. Frequently, accelerograms are affected by baseline jumps and drifts, either related to the instrument and/or a major earthquake. In this work, I propose a correction method for these undesired baseline drifts based on segmented linear least squares. The algorithm operates on the integrated waveforms and combines all three instrument components to estimate a model that modifies the baseline to be at zero continuously. The procedure consists of two steps: first a suite of models with variable numbers of discontinuities is derived for all three instrument components. During this process, the number of discontinuities is reduced in a parsimonious way, for example, two very close discontinuities are merged into a single one. In the second step, the optimal model is selected on the basis of the Bayesian information criterion. I exemplify the application on synthetic waveforms with known discontinuities and on observed waveforms from a unified strong-motion database of the Japan Meteorological Agency (JMA) and the National Research Institute for Earth Science and Disaster Prevention (NIED, Japan) networks for the major events of the 2016 Kumamoto earthquakes. After the baseline jump correction, the waveforms are furthermore corrected for displacement according to Wang et al. (2011). The resulting displacements are comparable to the Interferometric Synthetic Aperture Radar-derived displacement estimates for the Kumamoto earthquake sequence.

Introduction

Accelerometric waveform data are the basis of any ground-motion model (GMM). It is therefore of paramount importance to have clear recordings, free from errors. However, the presence of discontinuities in seismic data is ubiquitous and unavoidable. Discontinuities originate from natural (e.g., ground displacement and tilt [Graizer, 2006; Delorey et al., 2008; Vackar et al., 2015] or insufficient anchoring of the installation) or instrumental sources (e.g., self-noise, digitization noise [Wielandt and Streckeisen, 1982; Wilson et al., 2017]). In case of digitized analog record sections, improper splicing can introduce discontinuities in the time series (Douglas, 2003). Boore (2003) identified analog-to-digital conversion as another source of baseline drift in integrated acceleration recordings.

Irrespective of their origins, in the preparation of seismic data for the development of GMMs these discontinuities must be removed (Boore and Bommer, 2005). The data discontinuities appear as step-like pulses in the acceleration records and result in considerable offsets when integrated to velocity and displacement. The baseline jumps in acceleration result in a segmented and shifted baseline (Fig. 1).

The manual removal of discontinuities (Boore and Bommer, 2005) becomes cumbersome for large data sets, in particular when instrument-related discontinuities appear regularly. In this article, I propose an inversion routine (integrated combined baseline modification [ICBM]), which not only identifies the timing and scale of the individual jump but also determines the number of jumps in the record. ICBM operates on the integrated acceleration traces (velocity) and combines all three instrument components to estimate baseline offsets and their timings. The best baseline correction model is selected with the Bayesian information criterion (BIC, Schwarz, 1978), which not only takes the minimization of the residuals into account but also the number of parameters. Thus,

1. Helmholtz Centre Potsdam, German Research Centre for Geosciences - GFZ, Potsdam, Germany; 2. Institute of Geosciences, University of Potsdam, Potsdam-Golm, Germany
*Corresponding author: specht@gfz-potsdam.de

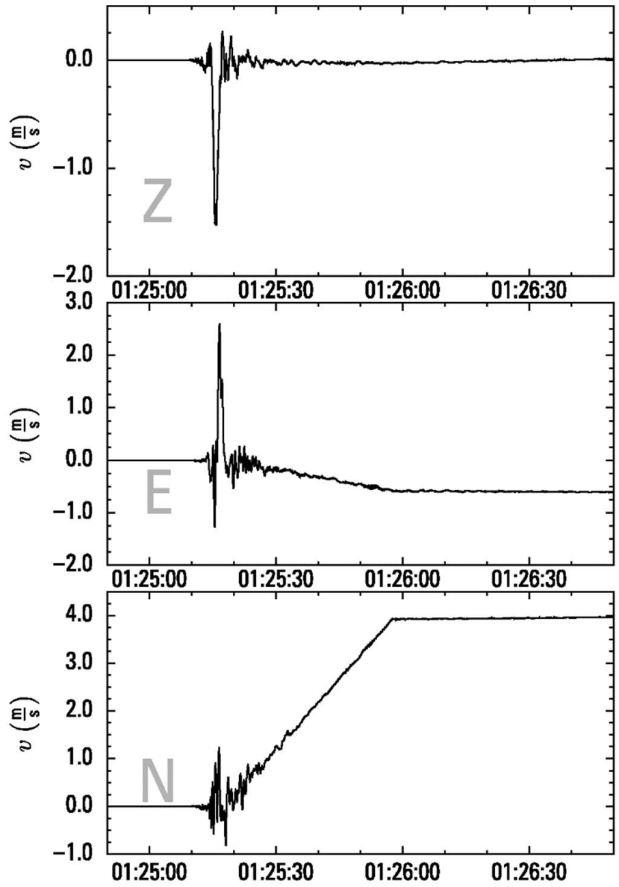

Figure 1. *Result of acceleration baseline jumps in the integrated trace (velocity) at station 93048 (three-component sensor from the Japanese municipal network). The waveform shown is from the Kumamoto 2016 earthquake. All segment shifts occur simultaneously on the traces. While the first jump is probably related to the event itself, the second discontinuity after 60 s is most likely instrument related. Any further processing of this waveform will introduce a bias if left uncorrected.*

the use of BIC safeguards against overfitting and the optimal model is selected in a parsimonious way. The method is applied to a synthetic data set and to waveform data from the 2016 Kumamoto (Japan) earthquake sequence. The baseline corrected waveforms are integrated to displacements after Wang et al. (2011) since displacements are most sensitive to baseline shifts and are compared to Interferometric Synthetic Aperture Radar (InSAR)-derived displacements by Jiang et al. (2017).

Method

A simple baseline jump in acceleration can be expressed as scaled and shifted Heaviside function:

$$a(t) = a_1 h(t - T_1) \tag{1}$$

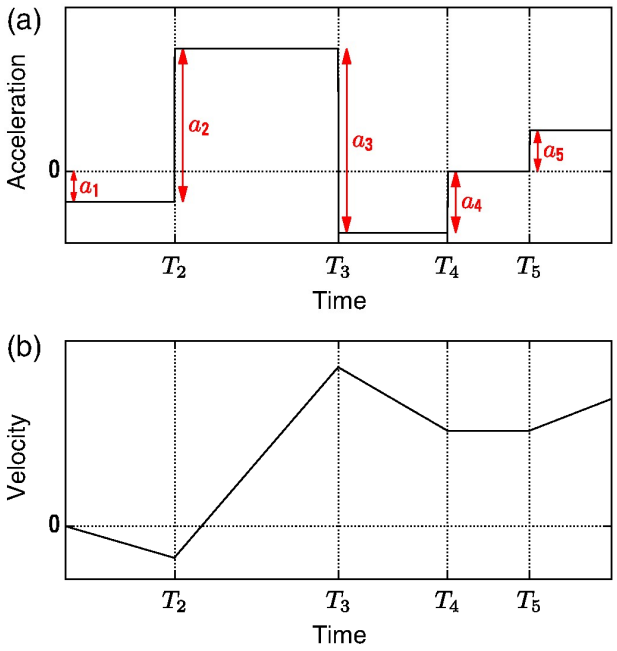

Figure 2. *(a) Concept of several baseline jumps. Note that jump amplitudes are measured with respect to each other and not absolutely. (b) The same as in (a) but for velocity (i.e., the integral of (a)). (a) Concept of several baseline jumps. Note that jump amplitudes are measured with respect to each other and not absolutely. (b) The same as in (a) but for velocity (i.e., the integral of (a)).*

in which a_1 is the jump amplitude and T_1 the time of the jump occurrence. The function $h(t)$ is the Heaviside function

$$h(t - T_1) = \begin{cases} 1 & \text{if } t \geq T \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Figure 2a visualizes the concept.

The previous concept can be easily extended to an arbitrary number of jumps M:

$$a(t) = \sum_{j=1}^{M} a_j h\big(t - T_j\big). \tag{3}$$

In this definition, the jump amplitude a_{j+1} is the change with respect to the preceding jump a_j, that is subsequent jumps superpose each other. This relative change necessitates a definition that considers a jump that occurred before the record starts and its shift continues in the record. Because the jump time of the first datum is not retrievable, the jump is assumed to be present for the entire record duration:

$$a(t) = a_1 + \sum_{j=2}^{M} a_j h(t - T_j). \quad (4)$$

In other words, a_1 is the baseline offset at the beginning of the record (Fig. 2a).

The discontinuity jumps may be small and well below the signal level. However, because each jump is a single unidirectional change in the record, they become very apparent on integrated traces as time-dependent offsets (Boore and Bommer, 2005). The integral of equation (4) is given by

$$\int_0^t a(\tau)\, d\tau = v(t) = a_0 + a_1 t + \sum_{j=2}^{M} a_j (t - T_j) h(t - T_j), \quad (5)$$

in which a_0 is the constant of integration and has units of velocity, in contrast to a_1, ..., a_M which have units of acceleration. The integral in equation (5) is visualized in Figure 2b.

The model proposed here is a nonlinear inverse problem with an unknown number of parameters, that is not only the parameters must be estimated but also the number of parameters must be determined. The model consists of two stages:

- a nonlinear least-squares (NLSQ) parameter estimation for the baseline discontinuities and their times and
- information-based optimization of number of parameters.

Baseline parameter estimation

The parameter estimation in ICBM for a discrete time series over N time steps is based on least absolute deviation

$$S = \sum_{i=1}^{N} |\Delta v_i| \quad (6)$$

$$= \sum_{i=1}^{N} \left| v_i^{obs} - v_i^{mod} \right|, \quad (7)$$

in which v_i^{obs} and v_i^{mod} are the ith observed and modeled velocities. The minimization of S is achieved by NLSQ, generally stated as

$$\Delta \boldsymbol{v} = \boldsymbol{J} \Delta \boldsymbol{p} \quad (8)$$

in which $\Delta \boldsymbol{v}$ is the vector of velocity residuals (may contain one or several seismic channels), $\boldsymbol{J}$ is the Jacobian matrix with the derivatives of the baseline model $\boldsymbol{v}$ with respect to its parameters, stated by vector $\boldsymbol{p}$. The changes of the model parameters per iteration are given by $\Delta \boldsymbol{p}$. The vector $\Delta \boldsymbol{v}$ represents the residuals between the observed and modeled data, that is

$$\Delta v_i = v_i^{obs} - \left(a_0 + a_1 t_i + \sum_{j=2}^{M} a_j (t_i - T_j) h(t_i - T_j) \right). \quad (9)$$

The algorithm may be defined on single-component records and multi-component (commonly three) records. If baseline jumps appear simultaneously on all three components, then using all three components in the algorithm simultaneously mitigates overfitting and undesired removal (or alteration) of actual signals.

In the following, the more general three-component-based routine is treated but can be easily changed to single component by removing the respective entries (e.g., y and z) in the vectors and the Jacobian. Let $\boldsymbol{v}$ be a vector with N samples from all three components:

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \vdots \\ \boldsymbol{v_i} \\ \vdots \\ \boldsymbol{v_N} \end{pmatrix} \text{ with } \boldsymbol{v_i} = \begin{pmatrix} v_i^x \\ v_i^y \\ v_i^z \end{pmatrix} \quad (10)$$

in which $\boldsymbol{v_i} = \left(v_i^x, v_i^y, v_i^z\right)^T$ are the three velocity components of the ith time sample.

The definition introduced here assumes that jump discontinuities are at same times on all three components, but have different amplitudes on each component. Therefore, baseline jumps are not necessarily correlated in time on all three traces, which is formally realized when one- or two-component amplitudes are zero. Under these assumptions, the parameter vector is defined as

$$\boldsymbol{p} = \begin{pmatrix} \boldsymbol{a_0} \\ \boldsymbol{a_1} \\ \boldsymbol{a_2} \\ T_2 \\ \vdots \\ \boldsymbol{a_j} \\ T_j \\ \vdots \\ \boldsymbol{a_M} \\ T_M \end{pmatrix} \text{with } \boldsymbol{a_j} = \begin{pmatrix} a_j^{(x)} \\ a_j^{(y)} \\ a_j^{(z)} \end{pmatrix}, \tag{11}$$

in which $\boldsymbol{a_j}$ are the amplitudes of the three components of the jth baseline jump and T_j is the time of the jth baseline jump. As a consequence of equation (5) there are no T_0 and T_1. From a model point of view, $\boldsymbol{a_0}$ represents the offset at the time of the first data sample (i.e., the intercept) and $\boldsymbol{a_1}$ is the slope of the baseline in the beginning until time T_2 (Fig. 2). The three-component Jacobian with simultaneously occurring jumps on all three components is

$$\boldsymbol{J} = \frac{\partial v}{\partial \boldsymbol{p}} \begin{pmatrix} \frac{\partial \boldsymbol{v}_1}{\partial \boldsymbol{a_0}} & \frac{\partial \boldsymbol{v}_1}{\partial \boldsymbol{a_1}} & \cdots & \frac{\partial \boldsymbol{v}_1}{\partial \boldsymbol{a_j}} & \frac{\partial \boldsymbol{v}_1}{\partial T_j} & \cdots & \frac{\partial \boldsymbol{v}_1}{\partial \boldsymbol{a_M}} & \frac{\partial \boldsymbol{v}_1}{\partial T_M} \\ \frac{\partial \boldsymbol{v}_2}{\partial \boldsymbol{a_0}} & \frac{\partial \boldsymbol{v}_2}{\partial \boldsymbol{a_1}} & \cdots & \frac{\partial \boldsymbol{v}_2}{\partial \boldsymbol{a_j}} & \frac{\partial \boldsymbol{v}_2}{\partial T_j} & \cdots & \frac{\partial \boldsymbol{v}_2}{\partial \boldsymbol{a_M}} & \frac{\partial \boldsymbol{v}_2}{\partial T_M} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{a_0}} & \frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{a_1}} & \cdots & \frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{a_j}} & \frac{\partial \boldsymbol{v}_i}{\partial T_j} & \cdots & \frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{a_M}} & \frac{\partial \boldsymbol{v}_i}{\partial T_M} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial \boldsymbol{v}_N}{\partial \boldsymbol{a_0}} & \frac{\partial \boldsymbol{v}_N}{\partial \boldsymbol{a_1}} & \cdots & \frac{\partial \boldsymbol{v}_N}{\partial \boldsymbol{a_j}} & \frac{\partial \boldsymbol{v}_N}{\partial T_j} & \cdots & \frac{\partial \boldsymbol{v}_N}{\partial \boldsymbol{a_M}} & \frac{\partial \boldsymbol{v}_N}{\partial T_M} \end{pmatrix}. \tag{12}$$

And the derivatives are given by

$$\frac{\partial v_i^{(k)}}{\partial a_0^{(l)}} = \delta_{kl}, \tag{13}$$

$$\frac{\partial v_i^{(k)}}{\partial a_1^{(l)}} = t\delta_{kl}, \tag{14}$$

$$\frac{\partial v_i^{(k)}}{\partial a_j^{(l)}} = \big(t - T_j\big)h\big(t - T_j\big)\delta_{kl}, \quad \forall j \geq 2, \tag{15}$$

$$\frac{\partial v_i^{(k)}}{\partial T_j} = a_j^{(k)} h\big(t - T_j\big), \quad \forall j \geq 2, \tag{16}$$

in which the three seismic components are given by superscripts k, $l = \{x, y, z\}$, and δ_{kl} is the Kronecker delta.

The solution of equation (8) is based on least squares. For the least absolute deviation (equation 6) the NLSQ solution is iteratively reweighted:

$$\Delta\boldsymbol{p} = (\boldsymbol{J^T R J})^{-1}\boldsymbol{J^T R}\Delta\boldsymbol{v}. \tag{17}$$

The matrix $\boldsymbol{R}$ is a diagonal matrix, in which ith diagonal element is given by

$$R_{ii} = \frac{1}{max(\delta, |\Delta v_i|)}. \tag{18}$$

The maximum with threshold δ avoids singularities in R. The value for δ is application dependent and should reflect the average noise level of the record (preevent and postevent). A reasonable value is $\delta = 10^{-4}$ ms^{-1}, representing a small velocity for acceleration records.

The incremental change of the parameter vector $\Delta\boldsymbol{p}$ requires initial values for $\boldsymbol{p}_{initial}$ and $\boldsymbol{p}$ is updated after each iteration (i):

$$\boldsymbol{p}^{(i+1)} = \boldsymbol{p}^{(i)} + \Delta\boldsymbol{p}^{(i)} \tag{19}$$

There is no direct *a priori* information about the jump amplitudes available. For the amplitude parameters $\boldsymbol{a_0}$ (offset at the record beginning) and $\boldsymbol{a_1}$ (average slope of the entire record) initial values are estimated from an ordinary least-squares fit to the entire record for all three channels. The remaining amplitude parameters $\boldsymbol{a}_{1+j}$ are set to zero, that is, no baseline jumps are assumed initially. The initial values for the jump times T_j are equally distributed over the record length.

Spectral properties of a baseline jump

In equation (5), the baseline drift on the velocity traces is defined as a superposition of linear function. Although this functional form is practical for the inversion, it is not so for discussing some of its spectral properties. For an arbitrary velocity trace segment between times T_1 and T_2 with constant drift, the baseline model function for that segment is

$$v(t) = \big(p + q(t - T_1)\big)\big(h(t - T_1) - h(t - T_2)\big), \tag{20}$$

in which p is the offset at T_1 due to any previous discontinuities and q is drift due to the acceleration offset associated with time T_1, and at time

T_2 is the subsequent acceleration offset. The Fourier transform of the velocity baseline segment is (for derivation, see Appendix)

$$V(\omega) = \begin{cases} p(T_2 - T_1) + \frac{q}{2}(T_2 - T_1)^2 & \text{if } \omega = 0 \\ \frac{1}{\omega}\left\{\left[\frac{q}{\omega} + (q(T_2 - T_1) + p)i\right] e^{-i\omega T_2} - \left[\frac{q}{\omega} + ip\right] e^{-i\omega T_1}\right\} & \text{if } \omega \neq 0 \end{cases} \quad (21)$$

The spectrum is visualized in Figure 3. An important feature is the main lobe of the spectrum around zero frequency. Its width is related to the length of the segment $\Delta T = T_2 - T_1$ and covers the frequency range below the frequency of the first local minimum of the spectrum at $f_0 = \frac{1}{\Delta T}$.

The spectrum of a series of baseline jumps results from the summation of spectra as defined in equation (21). When correcting baseline jumps it is important to consider the spectra of the baseline drift and of the actual signal. Interference between the signal spectrum and the correction function spectrum should be minimal, that is, it is necessary to define a minimum time difference between baseline jumps for the correction model: ΔT_{min}.

The time threshold ΔT_{min} is related to the width of the main lobe. And as the main lobe contains most the baseline drift energy (Fig. 3), an estimate of ΔT_{mi} can be derived from the spectral properties of the seismic signal by different means to reduce interference between the baseline spectrum main lobe and the seismic signal spectrum:

- Investigation of the usable bandwidth of the signal spectrum by computing the spectral ratio between signal and noise spectra. The lower frequency at which the spectral ratio exceeds a given amplitude threshold (e.g., 3) is used as the inverse of ΔT_{min}. This definition is most useful for records of seismic signals where background noise is of considerable influence.
- Inversion of parameters for a Brune spectrum model (Brune, 1970; magnitude M, corner frequency f_c, and along-path attenuation κ). Then one can set $\Delta T_{min} = f_c^{-1}$.
- Estimation of the signal duration, T_{dur} based on the running Arias intensity (e.g., Bommer et al., 2009) as a proxy for threshold time difference. The signal duration is the time difference between, for example, the times where the running Arias intensity reaches 5% and 95% of the total Arias intensity of the seismic signal.

When setting $\Delta T_{min} \geq \Delta T_{dur}$, only a single baseline jump can occur within the baseline model. This safeguards against baseline overfitting in the actual seismic signal and restricts potential contamination of the signal spectrum to the frequency band below the corner frequency. The corner frequency f_c is related to the source duration, resulting in a lower (and thus less conservative) ΔT_{min} compared to usage of T_{dur}, which also includes path-related effects of wave guide propagation (Herrmann, 1985).

Redundancy of modeled jumps

During the iteration process, some jump time T_m may become similar - or even identical for larger M - to another jump time T_n. Such overfit is undesired as it renders the model overly complicated and with increasing M, the baseline correction will approximate the actual waveform data. As described in the Spectral Properties of a Baseline Jump section, the baseline jump spectrum can contaminate the seismic spectrum, if the modeled baseline jumps are too frequent in time, in particular when subsequent jumps are less than the time threshold ΔT_{min}.

After each iteration of equation (17), all time intervals $|T_m - T_n|$ are examined to avoid unnecessary or redundant jumps in the baseline model. If for any two times it holds $|T_m - T_n| < \Delta T_{min}$, then the parameter sets are merged and the nth parameter set is updated:

$$\boldsymbol{a}_n^{updated} = \boldsymbol{a}_m + \boldsymbol{a}_n, \quad (22)$$

$$T_n^{updated} = \frac{1}{2}(T_m - T_n). \quad (23)$$

The mth parameter set is then removed. The four updated parameters are a combination of

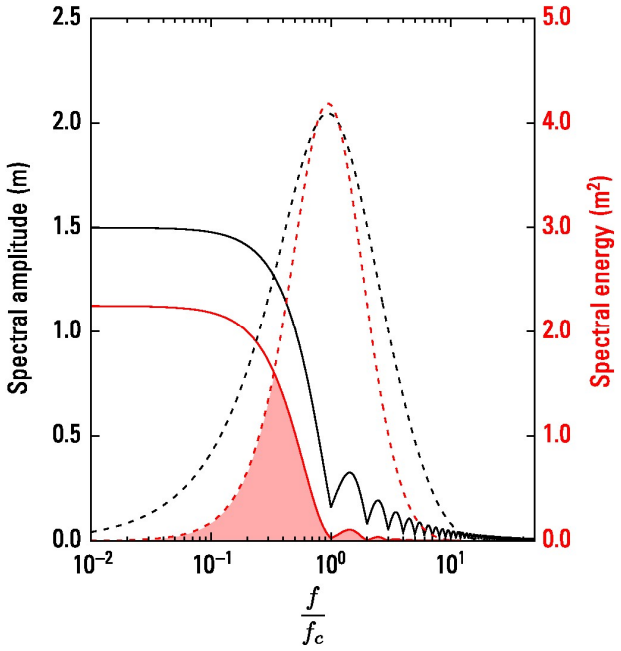

Figure 3. *Theoretical spectra of a single baseline drift on a velocity trace (solid lines) and Brune source spectrum (dashed lines). The frequency axis is normalized to the main lobe width of the baseline spectrum and respectively to the corner frequency of the Brune source spectrum. Spectra in black are amplitude spectra, in red are energy spectra (squared amplitude spectra). The baseline segment spectrum is based on a 10 s segment with 0.1 ms^{-1} initial offset and a velocity drift of 0.01 ms^{-2}. The Brune source spectrum has corner frequency f_c 0.1 Hz, M_w 7.0, along-path attenuation $\kappa = 680f^{0.38}$, and a distance of 30 km; other parameters are taken from Atkinson (2000, his table 4) and Boore (2003). The light red area indicates the overlap of the baseline segment and Brune source energy spectra. When defining the minimum segment length in the baseline model, the earthquake spectra, on which the model is applied, should be taken into account. The source corner frequency is an adequate proxy for the definition of the minimum segment length (as the inverse of the corner frequency) at shorter distances. At longer distances - that is when the travel path duration adds substantially to the source duration - the total signal duration should be considered as a more conservative proxy.*

the original eight parameters.

Another issue that may arise during the iteration is that some jump time T_m is out of data range, that is $T_m < 0$ or $T_m \geq T_N$. The jump parameters associated to the former case are equivalent to those in $\boldsymbol{a}_1$, that is spanning the entire record. Therefore, the parameters $\boldsymbol{a}_m$ in the former case $T_m < 0$ are added to $\boldsymbol{a}_1$:

$$\boldsymbol{a}_1^{updated} = \boldsymbol{a}_1 + \boldsymbol{a}_m, \tag{24}$$

and the jump amplitudes $\boldsymbol{a}_m$ and time T_m are removed from the parameters. In the latter case $T_m \geq T_N$ all parameters moved out of scope, such that neither the jump time T_m nor the amplitudes am have any impact on the model. Hence, these parameters are removed from the parameter vector.

Redundancy of parameter estimates may arise if the jump amplitudes am become very small. If it holds that

$$\bar{a} < \sum_{k=1}^{3} \left| a_m^{(k)} \right|, \tag{25}$$

then the parameters $\boldsymbol{a}_m$ and time T_m are removed. The value of $\bar{a}$ is application dependent, but when considering strong-motion records, a value of $\bar{a} = 10^{-6}$ ms^{-2} is sufficient to remove small jump amplitudes that have a negligible contribution on the entire inversion and are below the level of seismic background noise.

Optimization of number of baseline jumps

To determine the optimal number of jumps, the NLSQ inversion is performed P times with changing M. For each inversion the BIC (Schwarz, 1978; Burnham and Anderson, 2002) is then calculated. The BIC is given by

$$\mathrm{BIC} = N_p \ln N_d - 2 \ln \mathcal{L} \tag{26}$$

in which N_p is the number of free parameters, N_d the number of data, and $\mathcal{L}$ the likelihood of the model. For (non-) linear least squares the assumption is that residuals are normally distributed, for the likelihood it then holds (Burnham and Anderson, 2002)

$$\mathcal{L} \sim \hat{\sigma}^2 = \frac{(\Delta \boldsymbol{v})^T (\Delta \boldsymbol{v})}{3N}. \tag{27}$$

The $3N$ arises from the number of data in all three components and $\Delta \boldsymbol{v}$ is the vector of residuals from equation (9). The number of free parameters N_p for the three-component baseline model in terms of the number of segments M is given by

$$N_p = 7 + 4(M - 1). \tag{28}$$

Note that the sum of the residuals is also a free parameter. Thus, the BIC for the baseline model is

TABLE 1
Events from the 2016 Kumamoto Earthquake Sequence Used for Displacement Estimation

Date (yyyy/mm/dd)	Time (hh:mm:ss)	M_{JMA}	Longitude (°)	Latitude (°)	Depth (km)
2016/04/14	21:26:00	6.5	130.808	32.742	11
2016/04/14	22:07:00	5.8	130.848	32.775	8
2016/04/15	00:03:00	6.4	130.777	32.700	7
2016/04/16	01:25:00	7.3	130.762	32.753	12
2016/04/16	01:44:00	5.4	130.760	32.752	15
2016/04/16	03:55:00	5.8	131.190	33.025	11
2016/04/16	09:48:00	5.4	130.835	32.847	16

Date and time are in local time (Japan Standard Time [JST]).

$$\mathrm{BIC} = \big(7 + 4(M-1)\big)\ln(3N) + 3N\ln(\hat{\sigma}^2). \quad (29)$$

The optimal model of the P models is the one that minimizes BIC.

With the optimal model selected, the acceleration traces are corrected by

$$a^{corr}(t) = a^{obs}(t) - a(t). \quad (30)$$

Examples

Synthetic data

This section illustrates the baseline correction with synthetic data. The purpose of the synthetic data is to demonstrate the reliability of the inversion as the number of jumps is set a priori and the times are known. Synthetic waveform data are generated from stochastic simulations for a magnitude 7 earthquake at 30 km distance using standard parameters from Boore (2003). Figure 4 shows a synthetic three-component acceleration waveform with noise. Three baseline jumps with different amplitudes on each component are added at the same times. While the acceleration traces are negligibly affected, the integrated traces show strong baseline drifts if left uncorrected. Similarly, the effect of the baseline jumps in the amplitude spectra is more apparent in the velocity spectra where the baseline jump spectra dominate the low frequency range below the source corner frequency (Fig. 5a). Compared to the velocity spectra, the acceleration spectra are less affected by the baseline jumps, although at very low frequencies the baseline jump spectra show a strong effect (Fig. 5b). ICBM corrects the baseline shifts by identifying the jump amplitudes and times and the velocity baseline is around zero. Because of the usage of BIC, the correction model is selected in a parsimonious way.

The 2016 Kumamoto earthquake

On 16 April 2016 central Kyushu (Japan) was hit by an M_w 7.1 strike-slip earthquake east of Kumamoto city. The rupture reached the surface and cause substantial surface displacements (Shirahama et al., 2016). Because of the rupture location within Kyushu, the event is well recorded on the dense seismological networks of Japan. Strong-motion data are available from the National Research Institute for Earth Science and Disaster Prevention (NIED, Japan) (K-NET, KiK-net) and the Japan Meteorological Agency (JMA) (seismic stations used for the determination of the JMA seismic intensity [shindo]). Given the dense seismological network, the available waveform database is extensive with several hundreds of records for the mainshock, and large foreshock and aftershocks. The database covers all events where data from NIED and JMA are available (Table 1).

All available JMA stations show baseline jumps on all three components (Fig. 6). These jumps would have implications on estimations of response spectra at low oscillatory periods, and a considerable impact on radiated seismic energy and peak ground velocity, and in particular static displacement and peak ground displacement. The static displacement estimation is most sensitive to baseline jumps due to the double integration of the accelerograms. The automated displacement estimation method of

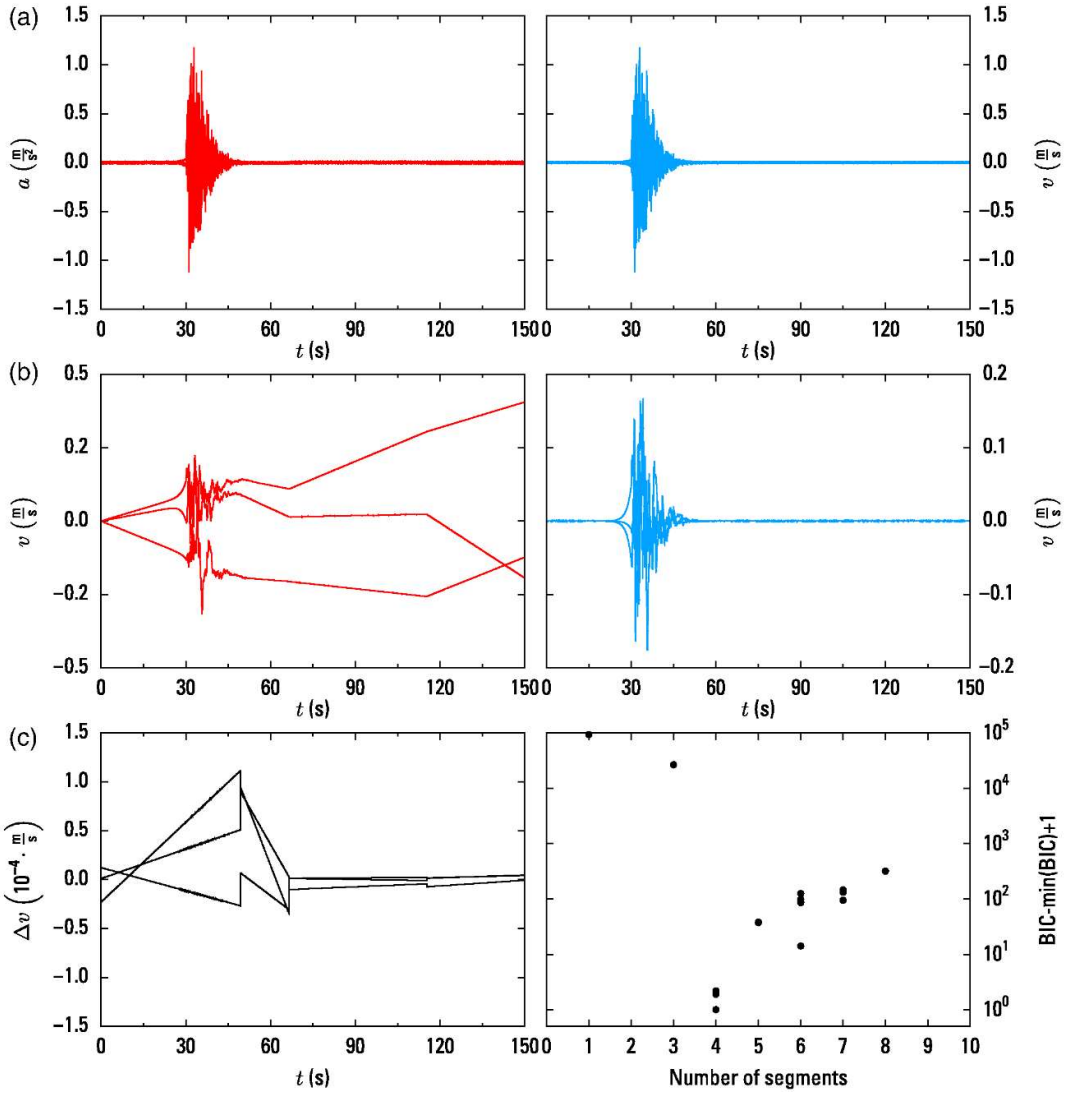

Figure 4. *Comparison of three-component acceleration waveform (a) without baseline correction (left panel, red waveforms) and the same waveforms with baseline correction (right panel, blue waveforms). The synthetic waveforms are generated by the stochastic method after Boore (2003). Three random baseline jumps have been added at random times (resulting in four linear segments), which are covered by noise due to the jumps small amplitudes. (b) The acceleration traces integrated to velocity and the uncorrected traces show now strong baseline drifts without correction (left), while the velocity baselines after correction are around zero (right). (c) The left panel: the residual of the baseline estimate of integrated combined baseline modification to the baseline shift function used to simulate the baseline jumps (i.e., ground truth). The deviation is minimal (note the y-axis scale), demonstrating that both amplitudes and times of the baseline jumps are sufficiently approximated. The right panel shows the Bayesian information criterion (BIC) against the number of baseline segments (equals baseline jumps plus one). The baseline correction used here was obtained after 20 runs with increasing number of segments (1–20). Because the algorithm can reduce the number of segments during the iteration process, some segment combinations have been eliminated before the BIC evaluation. The best models are correctly found with four segments.*

Wang et al. (2011) is applied to the waveform database. Their algorithm finds parameters to correct the waveform such that it best fits a step function. Because this algorithm expects one discontinuity only, it can fail to estimate static displacements in the presence of additional discontinuities (Fig. 7).

I compare the resulting cumulative static displacement of the major Kumamoto events with InSAR-based displacement estimates by Jiang et al. (2017). InSAR-derived displacements perform similarly to the seismically inferred displacements and source parameters (Weston et al., 2012; Kobayashi, 2017). The presequence InSAR imagery was acquired on 8 April and the post-sequence data on 20 April, thus covering in addition to the mainshock all major foreshock and aftershocks. The InSAR and accelerometer-derived displacements correlate highly and displacements scale equally with both methods

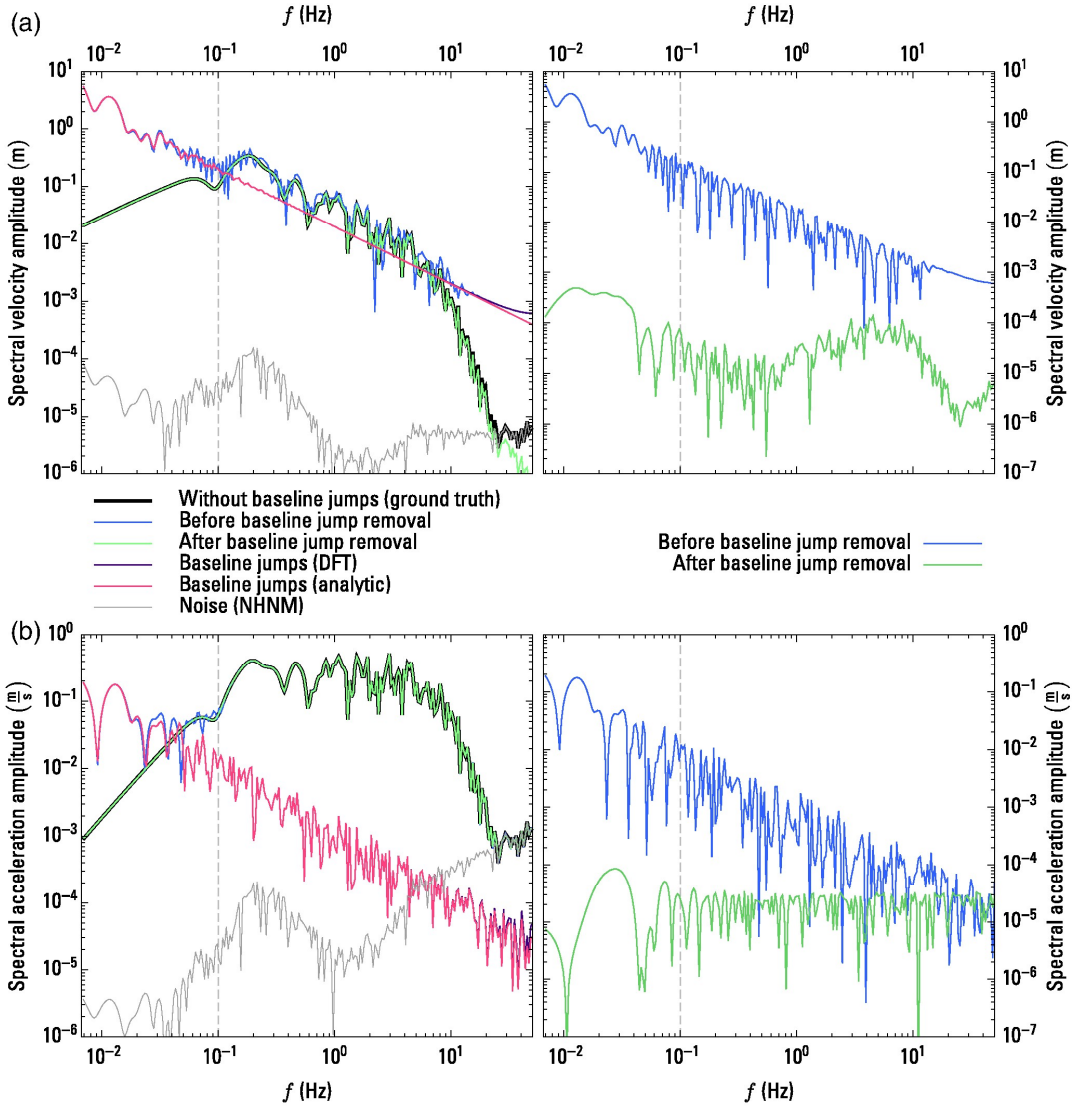

Figure 5. *Waveform spectra of a component (lowest of the three) in Figure 4b, left. The source spectrum is designed after Boore (2003) with parameters from Atkinson (2000) for the Kumamoto earthquake with corner frequency f_c 0.1 Hz, M_w 7.0, along-path attenuation $\kappa = 680f^{0.38}$ (see also Fig. 3). To the earthquake signal is a noise signal added based on the new high-noise model (NHNM Peterson, 1993). (a) Velocity and (b) Acceleration spectra. (left) The spectrum of the waveforms with baseline drift (blue) is a superposition of the baseline drift spectrum (magenta) and the source spectrum (black). The baseline drift spectrum dominates at low frequencies, in particular below the source corner frequency (dashed line). The difference between the analytic expression of the baseline segment spectrum (magenta) and the discrete Fourier transform (DFT) of the baseline segment time series (dark purple) is negligible and increases relatively only slightly close to the Nyquist frequency at 50 Hz. After the removal of the baseline drift, the corrected spectrum (green) recovers completely the uncontaminated spectrum (ground truth, black). The major difference between the velocity and acceleration trace is relatively much smaller effect of the baseline shift at higher frequencies. (Right) Difference between spectra before (blue) and after (green) baseline correction with respect to the ground truth for velocity (a) and acceleration (b). The colors of the graphs correspond to the ones from the left column. The baseline correction removes the spectral contamination nearly completely with very little deviation from the ground truth for both velocity and acceleration.*

(Fig. 8). The functional relation between the two is

$$d_{\mathrm{InSAR}} = 0.89 d_{acc} + 0.01, \qquad (31)$$

in which parameters were estimated by Deming regression (orthogonal least squares Deming, 1943), with a very high Pearson correlation $r =$ 0.932.

Displacements are also in agreement with in-field measurements by Shirahama et al. (2016) in the vicinity of the surface rupture and displacements estimated from light detection and ranging measurements by Scott et al. (2018). Major discrepancies appear only for few stations in the northern part of the caldera of Mt.

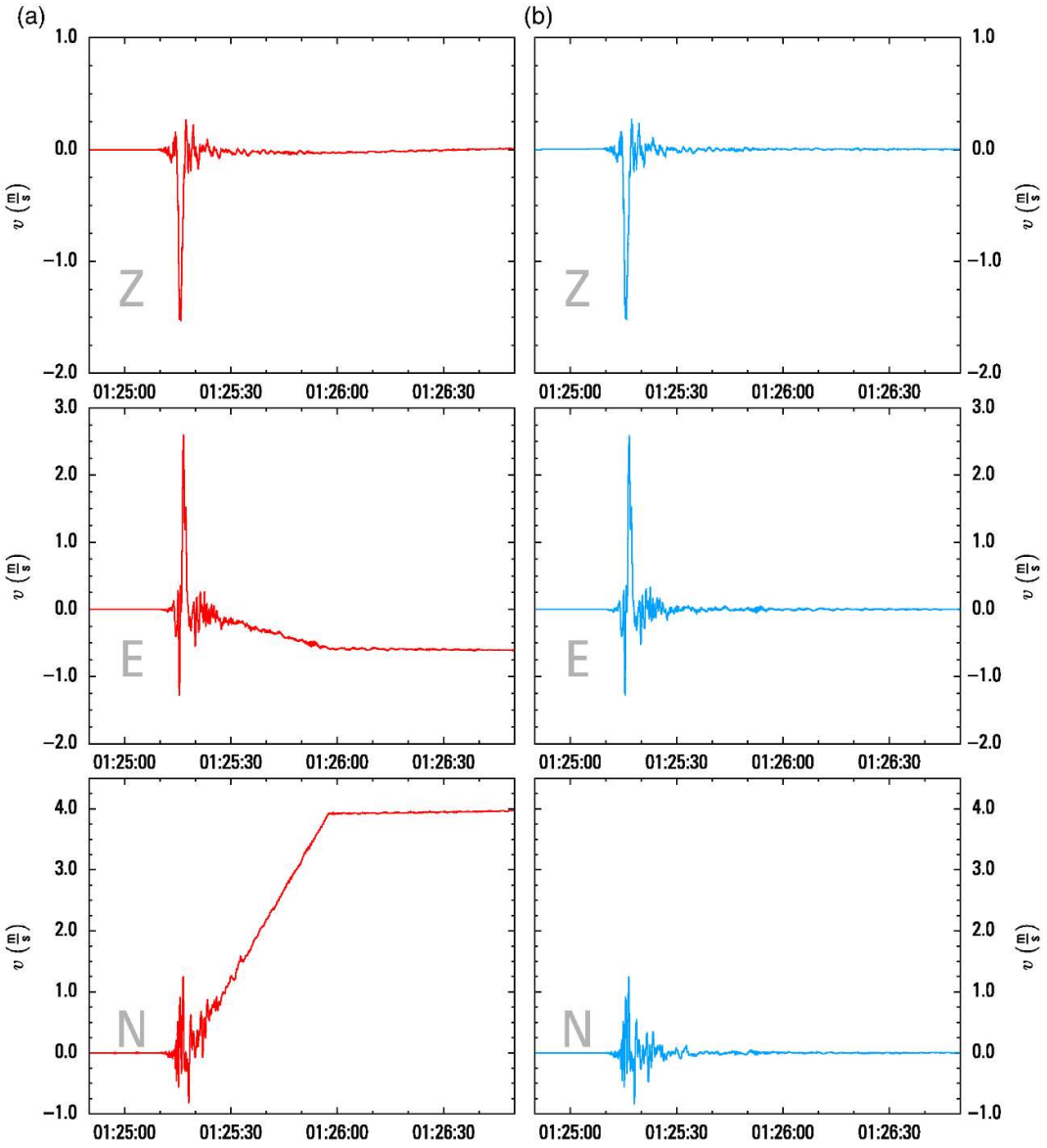

Figure 6. *The example waveform of Figure 1. Accelerations are integrated to velocities. (a) Raw velocity traces without baseline corrections. There are two notable discontinuities, one in the strong-motion portion (most likely event related) and another 40 s later (most likely instrument related). (b) Velocities after baseline correction. All three velocity traces are now flat with both discontinuities removed.*

Aso, northeast of the area with large static displacement (Fig. 8). The differences in displacements can be attributed to uncertainties of either displacement estimation method and might also be related to, for example, highly localized displacements at smaller surface ruptures in the rupture vicinity (Fujiwara et al., 2016).

Discussion and Conclusion

In this article, I introduced an automated baseline correction for three-component signals based on nonlinear least absolute deviation with model optimization related to the BIC. The method corrects for baseline jumps in accelerograms by identifying signal drifts and discontinuities in the velocity traces (time-integrated accelerograms). The correction function is a segmented linear function with an arbitrary number of segments. Because baseline jumps occur spuriously, several restrictions are imposed on the model to avoid overfitting (no clustering of jumps) and to stabilize the inversion process (jumps may occur simultaneously on all three components). Proxies for the minimum time length of the baseline segments are proposed that are related to the source corner frequency, usable spectral bandwidth, and the

signal duration. These definitions ensure that the earthquake signal is not contaminated by many baseline corrections that might arise due to overfitting. The BIC also selects the baseline model that best fits the baseline jumps in a parsimonious way.

Not only can the robust automated process handle large amounts of waveform data but the determination of the jump times is performed on up to three components simultaneously, improving the robustness of estimates even at noisier times, for example, during an earthquake or when the jumps are in the coda. During the iteration of ICBM, all baseline jump estimates are improved concurrently, which contrasts with the approach by Boore and Bommer (2005) with a sequential fit of jumps. The concurrent estimation is an advantage, as estimation errors in ICBM are more evenly distributed over all jumps, while in the sequential fit errors from earlier jumps are propagated to later jumps, where estimates of later jumps can be overall more erroneous than earlier jumps.

The applicability of the method has been demonstrated on synthetic signals and real signals from the 2016 Kumamoto earthquake sequence. Comparison of the static displacements estimated from accelerograms with InSAR displacements by Jiang et al. (2017) highlights that the automated baseline correction can be used as a preprocessing routine for strong-motion records as described by Boore and Bommer (2005) and to improve static displacement estimation routines (Wang et al., 2011).

Data and Resources

The waveform data of the Kumamoto earthquake are obtained from K-NET/KiK-net (http://www.kyoshin.bosai.go.jp/kyoshin/, last accessed May 2019) and from the Japanese Meteorological Agency (https://www.data.jma.go.jp/svd/eqev/data/kyoshin/jishin/index.html, last accessed May 2019). The Interferometric Synthetic Aperture Radar (InSAR) data of Jiang et al. (2017) are available as electronic supplement at (https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2016GL072253, last accessed May 2019). The Light Detection and Ranging (LiDAR) data of Scott et al. (2018) are available as an electronic supplement at (https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JB015581, last accessed May 2019). I provide the code for integrated combined baseline modification (ICBM) as C++ header based on the linear algebra library Armadillo by Sanderson and Curtin (2016). The code is available at https://github.com/EmperorOfTheMoon/ICBM (last accessed November 2017).

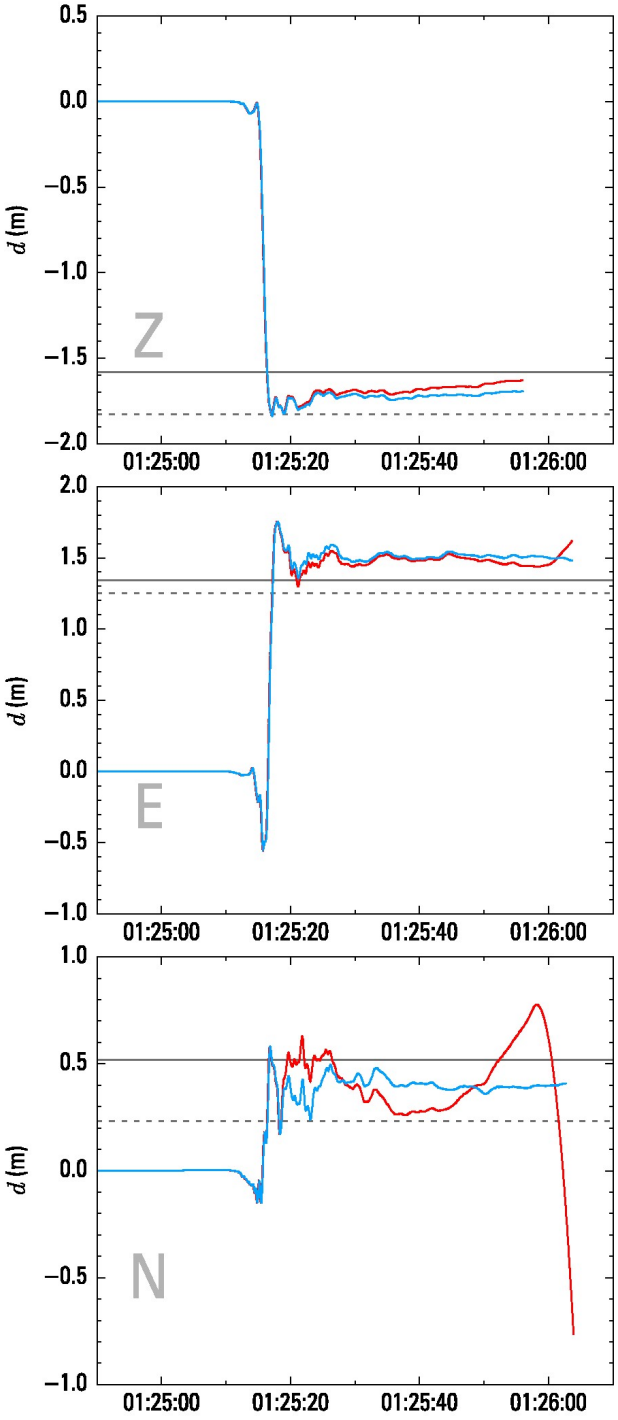

Figure 7. *Static displacement estimation after Wang et al. (2011) for the waveforms in Figure 6. Displacements without baseline correction are in red, with baseline correction in blue. The waveform length is cut according to Wang et al. (2011) to estimate the static displacements. While the Z and N components are less affected by the baseline jumps, the N component is strongly affected. The solid gray lines are displacement estimates based on the high-resolution Light Detection and Ranging (LiDAR) model of Scott et al. (2018) at the accelerometer location. The dashed gray lines are displacements inferred by Interferometric Synthetic Aperture Radar (InSAR) after Jiang et al. (2017) approximately 500 m away in north-northwest direction.*

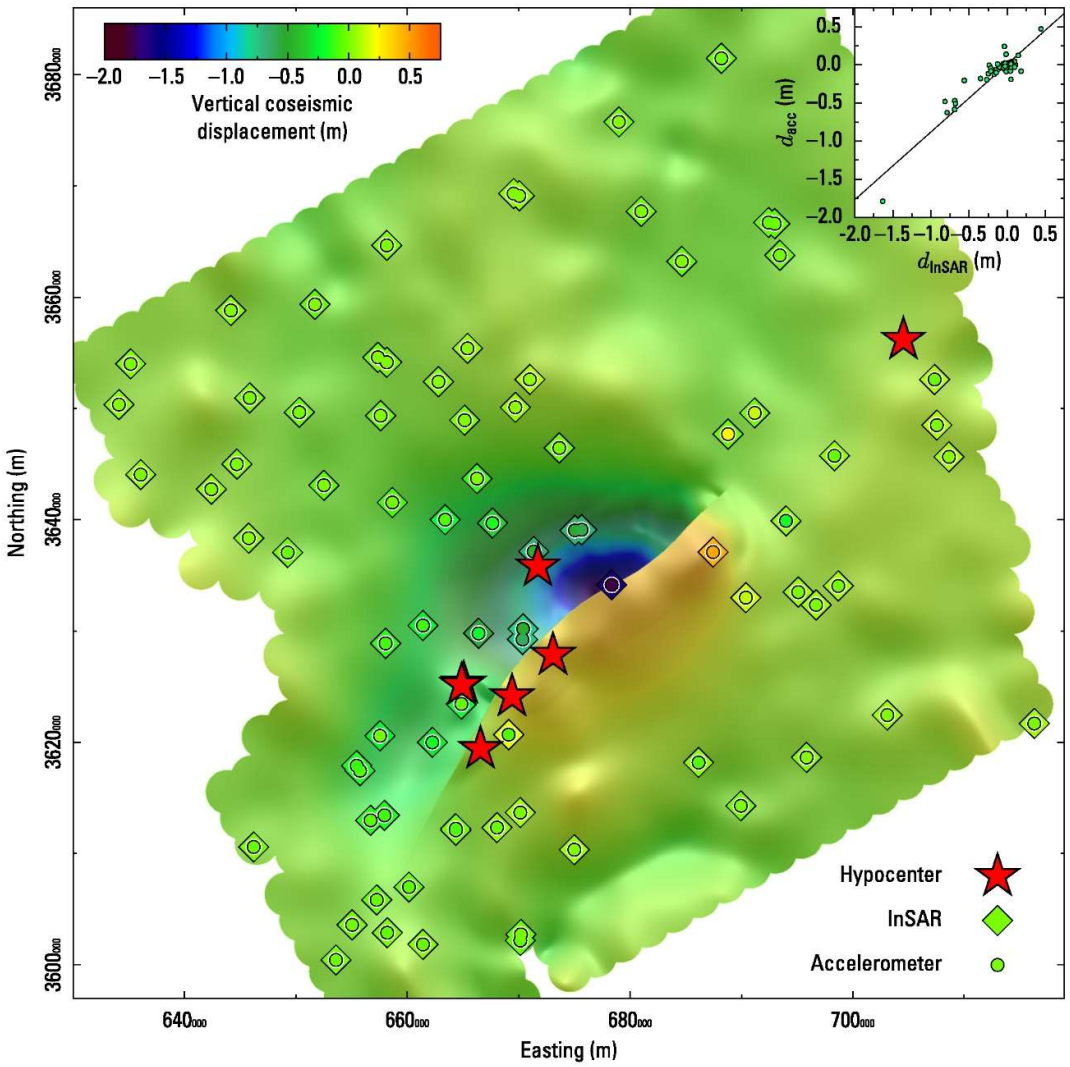

Figure 8. *Map of the vertical coseismic displacement of the 2016 Kumamoto earthquake sequence derived from InSAR (Jiang et al., 2017). The surface rupture of the mainshock is indicated by the sudden displacement offset. Locations of major events (Table 1) are shown as stars. InSAR displacements at the accelerometer sites (diamonds) fit spatially to the accelerometer-based displacements (circles in diamonds). The inset shows the comparison between accelerometer-derived displacement (d_{acc}) and InSAR-derived displacement (d_{InSAR}).*

Acknowledgments

Many thanks to Julian Bommer and Dave Boore for reviewing this article, which resulted in a very appreciable improvement of the article. Also, thanks to Fabrice Cotton for helping to improve the initial version of this article.

References

Atkinson, G. M. (2000). Stochastic modeling of California ground motions, Bull. Seismol. Soc. Am. 90, no. 2, 255–274.

Bommer, J. J., P. J. Stafford, and J. E. Alarcon (2009). Empirical equations for the prediction of the significant, bracketed, and uniform duration of earthquake ground motion, Bull. Seismol. Soc. Am. 99, no. 6, 3217–3233.

Boore, D. M. (2003). Simulation of ground motion using the stochastic method, Pure Appl. Geophys. 160, no. 3, 635–676.

Boore, D. M., and J. J. Bommer (2005). Processing of strong-motion accelerograms: Needs, options and consequences, Soil Dynam. Earthq. Eng. 25, no. 2, 93–115.

Brune, J. N. (1970). Tectonic stress and the spectra of seismic shear waves from earthquakes, J. Geophys. Res. 75, no. 26, 4997–5009.

Burnham, K. P., and D. R. Anderson (2002). Model Selection and Multi-model Inference, Second Ed., Springer-Verlag New York, Inc., New York, New York.

Delorey, A. A., J. Vidale, J. Steim, and P. Bodin (2008). Broadband sensor nonlinearity during moderate shaking, Bull. Seismol. Soc. Am. 98, no. 3, 1595–1601.

Deming, W. E. (1943). Statistical Adjustment of Data, John Wiley & Sons, New York, New York.

Douglas, J. (2003). What is a poor quality strong-motion record?, Bull. Earthq. Eng. 1, no. 1, 141–156.

Fujiwara, S., H. Yarai, T. Kobayashi, Y. Morishita, T. Nakano, B. Miyahara, H. Nakai, Y. Miura, H. Ueshiba, Y. Kakiage, et al. (2016). Small-displacement linear surface ruptures of the 2016 Kumamoto earthquake sequence detected by ALOS-2 SAR interferometry, Earth Planets

Space 68, no. 1, 160.
Graizer, V. (2006). Tilts in strong ground motion, Bull. Seismol. Soc. Am. 96, no. 6, 2090–2102.
Herrmann, R. B. (1985). An extension of random vibration theory estimates of strong ground motion to large distances, Bull. Seismol. Soc. Am. 75, no. 5, 1447–1453.
Jiang, H., G. Feng, T. Wang, and R. Bürgmann (2017). Toward full exploitation of coherent and incoherent information in Sentinel-1 TOPS data for retrieving surface displacement: Application to the 2016 Kumamoto (Japan) earthquake, Geophys. Res. Lett. 44, no. 4, 1758–1767.
Kobayashi, T. (2017). Earthquake rupture properties of the 2016 Kumamoto earthquake foreshocks (Mj 6.5 and Mj 6.4) revealed by conventional and multiple-aperture InSAR, Earth Planets Space 69, no. 1, 7.
Peterson, J. R. (1993). Observations and modeling of seismic background noise, Technical Report USGS Grant No. 93-322.
Sanderson, C., and R. Curtin (2016). Armadillo: a template-based C++ library for linear algebra, J. Open Source Softw., 1, 26.
Schwarz, G. (1978). Estimating the dimension of a model, Ann. Stat. 6, no. 2, 461–464.
Scott, C. P., J. R. Arrowsmith, E. Nissen, L. Lajoie, T. Maruyama, and T. Chiba (2018). The M 7 2016 Kumamoto, Japan, Earthquake: 3-D deformation along the fault and within the damage zone constrained from differential Lidar topography, J. Geophys. Res. 123, no. 7, 6138–6155.
Shirahama, Y., M. Yoshimi, Y. Awata, T. Maruyama, T. Azuma, Y. Miyashita, H. Mori, K. Imanishi, N. Takeda, T. Ochi, et al. (2016). Characteristics of the surface ruptures associated with the 2016 Kumamoto earthquake sequence, central Kyushu, Japan, Earth Planet. Space 68, no. 1, 191.
Vackar, J., J. Burjanek, and J. Zahradnik (2015). Automated detection of long-period disturbances in seismic records; MouseTrap Code, Seismol. Res. Lett. 86, no. 2A, 442–450.
Wang, R., B. Schurr, C. Milkereit, Z. Shao, and M. Jin (2011). An improved automatic scheme for empirical baseline correction of digital strong-motion records, Bull. Seismol. Soc. Am. 101, no. 5, 2029–2044.
Weston, J., A. M. Ferreira, and G. J. Funning (2012). Systematic comparisons of earthquake source models determined using InSAR and seismic data, Tectonophysics 532/535, 61–81.
Wielandt, E., and G. Streckeisen (1982). The leaf-spring seismometer: Design and performance, Bull. Seismol. Soc. Am. 72, no. 6A, 2349–2367.
Wilson, D., A. T. Ringler, and C. R. Hutt (2017). Detection and characterization of pulses in broadband seismometers, Bull. Seismol. Soc. Am. 107, no. 4, 1773–1780.

Appendix

The Fourier transform of

$$v(t) = \big(p + q(t - T_1)\big)\big(h(t - T_1)\big) - \big(h(t - T_2)\big) \tag{A1}$$

is

$$V(\omega) = \int_{-\infty}^{\infty} v(t) e^{-i\omega t} dt \tag{A2}$$

and can be stated as

$$V(\omega) = \int_{T_1}^{T_2} \big(p + q(t - T_0)\big)\, e^{-i\omega t} dt \tag{A3}$$

$$= p \int_{T_1}^{T_2} e^{-i\omega t} dt + q \int_{T_1}^{T_2} (t - T_1) e^{-i\omega t} dt. \tag{A4}$$

The left integral of equation (A4) is readily given by

$$\int_{T_1}^{T_2} e^{-i\omega t} dt = \begin{cases} T_2 - T_1 & \text{if } \omega = 0 \\ \dfrac{i}{\omega}\left(e^{-i\omega t T_2} - e^{-i\omega t T_1}\right) & \text{if } \omega \neq 0 \end{cases} \tag{A5}$$

Integration by parts gives the right integral of equation (A4)

$$\int_{T_1}^{T_2} (t - T_1) \frac{du}{dt} dt \tag{A6}$$

with

$$\frac{du}{dt} = e^{-i\omega t} \text{ and } u = \frac{i}{\omega} e^{-i\omega t} \tag{A7}$$

and thus

$$\int_{T_1}^{T_2} (t - T_1) e^{-i\omega t} dt = \left[\frac{1}{2}(t^2 - T_1 t) u\right]_{T_1}^{T_2} - \int_{T_1}^{T_2} \frac{d(t - T_1)}{dt} u dt \tag{A8}$$

$$= \begin{cases} \frac{1}{2}(T_2 - T_1)^2 & \text{if } \omega = 0 \\ \omega^{-1}(i(T_2 - T_1) + \omega^{-1})e^{-i\omega t T_2} - \omega^{-2}e^{-i\omega t T_1} & \text{if } \omega \neq 0 \end{cases}. \quad \text{(A9)}$$

Plugging equations (A5) and (A8) into equation (A4) gives the Fourier transform of the baseline drift segment

$$V(\omega) = \begin{cases} p(T_2 - T_1) + \frac{q}{2}(T_2 - T_1)^2 & \text{if } \omega = 0 \\ \frac{1}{\omega}\left\{\left[\frac{q}{\omega} + (q(T_2 - T_1) + p)i\right]e^{-i\omega t T_2} - \left[\frac{q}{\omega} + ip\right]e^{-i\omega t T_1}\right\} & \text{if } \omega \neq 0 \end{cases} \quad \text{(A10)}$$

Because of the finiteness of the signal, the spectrum of the derivative of $v(t)$ is not simply $i\omega V(\omega)$. In addition, the combined segments form a continuous function, while each segment alone is discontinuous. Therefore, the Fourier transform of a baseline jump segment comes from

$$a(t) = q\big(h(t - T_1) - h(t - T_2)\big). \quad \text{(A11)}$$

The Fourier transform is as in equation (A5)

$$A(\omega) = q\int_{T_1}^{T_2} e^{-i\omega t}dt \quad \text{(A12)}$$

$$= \begin{cases} q(T_2 - T_1) & \text{if } \omega = 0 \\ \frac{iq}{\omega}\left(e^{-i\omega t T_2} - e^{-i\omega t T_1}\right) & \text{if } \omega \neq 0 \end{cases}. \quad \text{(A13)}$$