Appendix

To estimate the temporal variation in μ, we used the same statistical approach as in IW2008. In this method, the temporal variation was represented by a piecewise linear function or linear spline[19], and the breaking points of the piecewise linear function were taken at all occurrence times t_i ($i = 1, 2, \ldots, N$) of the N earthquakes in an examined sequence. Hence, the temporal variation in μ was represented by

$$\mu(t) = \frac{\mu_{i+1} - \mu_i}{t_{i+1} - t_i}(t - t_i) + \mu_i \quad \text{for } t_i \leq t < t_{i+1}. \tag{A1}$$

We then incorporated a smoothness constraint or roughness penalty on $\mu(t)$. We intended to optimize $\boldsymbol{\theta} = (\mu_1, \mu_2, \ldots, \mu_N)$ (and β and σ), but practically optimization of such a huge number of parameters is an unstable process; the incorporation of the constraint enhances the stability of the optimization.

The likelihood function of the parameters was obtained using Equation 5 and is as follows:

$$L(\beta, \sigma, \boldsymbol{\theta}) = \prod_{i=1}^{N} f(M_i|\beta, \mu_i, \sigma). \tag{A2}$$

This formula agrees with Equation 6 if all values of μ_i's are the same. The smoothness constraint was quantified by the following function:

$$\Phi(\boldsymbol{\theta}|v) = v \int_0^T \left[\frac{\partial}{\partial t}\mu(t)\right]^2 dt, \tag{A3}$$

which can be rewritten as

$$\Phi(\boldsymbol{\theta}|v) \;\; = \;\; v \sum_{i=1}^{N-1} \frac{(\mu_{i+1} - \mu_i)^2}{t_{i+1} - t_i} \tag{A4}$$

by substituting Equation A1 into Equation A3.

Let us now consider the penalized log-likelihood function[22,23] $Q(\boldsymbol{\theta}|v, \beta, \sigma)$ as follows:

$$Q(\boldsymbol{\theta}|v, \beta, \sigma) = \ln L(\beta, \sigma, \boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}|v). \tag{A5}$$

The maximization of $Q(\boldsymbol{\theta}|v, \beta, \sigma))$ provides the best estimate of $\boldsymbol{\theta}$; however, the result depends on the value of v, which controls the trade-off between the goodness-of-fit of the model to the data and the smoothness, and the values of β and σ.

A Bayesian approach enables us to objectively determine the values of v, β, and σ by means of the type II maximum likelihood approach[24] or the maximization of the marginal likelihood[20]. We supposed that the assumed probability density function (i.e. prior distribution) of $\boldsymbol{\theta}$ is proportional to $\exp[-\Phi(\boldsymbol{\theta}|v)]$, and hence, the prior distribution $\pi(\boldsymbol{\theta}|v)$ was given by

$$\pi(\boldsymbol{\theta}|v) = \prod_{i=1}^{N-1} \sqrt{\frac{v}{\pi(t_{i+1} - t_i)}} \exp\left[-\frac{v(\mu_{i+1} - \mu_i)^2}{(t_{i+1} - t_i)}\right]. \tag{A6}$$

Then, if we integrate out the product of the likelihood function $L(\beta, \sigma, \boldsymbol{\theta})$ that appears in Equation A2 and the prior distribution $\pi(\boldsymbol{\theta}|v)$ over $\boldsymbol{\theta}$, we can obtain the marginal likelihood[25] with respect to v, β, and σ.

The integration described above, however, is impractical because the integration of $\pi(\boldsymbol{\theta}|v)$ over $\boldsymbol{\theta}$ is not finite; $\pi(\boldsymbol{\theta}|v)$ with respect to $\boldsymbol{\theta}$ is the so-called improper prior. Instead, we isolated μ_N from $\boldsymbol{\theta}$ because the integration of the prior over $\boldsymbol{\theta}_{-N} = (\mu_1,\ \mu_2,\ \ldots, \mu_{N-1})$ is finite. Thus, we integrated out the product over $\boldsymbol{\theta}_{-N}$, and the marginal likelihood $\mathcal{L}$ with respect to v, β, σ, and μ_N was obtained as follows:

$$\mathcal{L}(v, \beta, \sigma, \mu_N) = \int_{\Theta} L(\beta, \sigma, \boldsymbol{\theta}) \pi_{-N}(\boldsymbol{\theta}_{-N}|v, \mu_N) d\boldsymbol{\theta}_{-N}, \tag{A7}$$

where Θ denotes the parameter space of $\boldsymbol{\theta}_{-N}$, and the prior distribution in Equation A6 is rewritten here as $\pi_{-N}(\boldsymbol{\theta}_{-N}|v, \mu_N)$. The set of values of v, β, σ, and μ_N that maximizes the marginal likelihood are the best estimates[20,24]. In this Bayesian framework, the four parameters are often referred to as hyperparameters.

The optimization was carried out through the repetition of the following two steps. In the first step, for a particular set of values of the four hyperparameters, we searched the value of $\boldsymbol{\theta}_{-N}$ that maximizes the penalized log-likelihood function $Q(\boldsymbol{\theta}|v, \beta, \sigma)$ in Equation A5. In the second step, we computed the value of the logarithm of the marginal likelihood $\ln \mathcal{L}(v, \beta, \sigma, \mu_N)$. For this computation, the logarithm of the integrand in Equation A7 $\ln L(\beta, \sigma, \boldsymbol{\theta}) \pi_{-N}(\boldsymbol{\theta}_{-N}|v, \mu_N)$ was approximated by a quadratic form at the optimum of $\boldsymbol{\theta}_{-N}$ found in the first step. Then, the Laplace approximation[26] was used for the integration in

Equation A7. We changed the values of the four hyperparameters, and repeated these two steps until the value of $\ln \mathcal{L}(v, \beta, \sigma, \mu_N)$ was maximized.

For the model comparison, we introduced the Akaike Bayesian Information Criterion (ABIC)[20]:

$$\begin{aligned} \mathrm{ABIC} &= -2(\text{maximum } \ln \mathcal{L}) \\ &\quad +2(\text{the number of optimized hyperparameters}). \end{aligned} \tag{A8}$$

In the case where the temporal variation in μ is allowed, the number of optimized hyperparameters is four.

The value of v is assumed to approach infinity when we do not consider temporal variation. In this case, the prior distribution approaches the Dirac delta function $\delta(\boldsymbol{\theta}_{-N})$, and the limit of the marginal likelihood is

$$\mathcal{L}(v, \beta, \sigma, \mu_N) \to \int_{\Theta} L(\beta, \sigma, \boldsymbol{\theta})\delta(\boldsymbol{\theta}_{-N})d\boldsymbol{\theta}_{-N} = L(\beta, \sigma, \mu_N) \quad \text{as} \quad v \to \infty. \tag{A9}$$

Consequently, maximization of the marginal likelihood agrees with the ordinary maximum likelihood method, and ABIC is equivalent to the Akaike Information Criterion[21]. A more precise and theoretical justification of the equivalence is provided by Akaike[27]. In this case, the number of the optimized (hyper)parameters is three (β, σ, and μ_N).