Detection of trend changes in time series using Bayesian inference

Nadine Schütz and Matthias Holschneider
Focus Area for Dynamics of Complex Systems, Universität Potsdam, Karl-Liebknecht-Strasse 24, D-14476 Potsdam, Germany
(Received 10 March 2011; revised manuscript received 24 May 2011; published 10 August 2011)

Change points in time series are perceived as isolated singularities where two regular trends of a given signal do not match. The detection of such transitions is of fundamental interest for the understanding of the system's internal dynamics or external forcings. In practice observational noise makes it difficult to detect such change points in time series. In this work we elaborate on a Bayesian algorithm to estimate the location of the singularities and to quantify their credibility. We validate the performance and sensitivity of our inference method by estimating change points of synthetic data sets. As an application we use our algorithm to analyze the annual flow volume of the Nile River at Aswan from 1871 to 1970, where we confirm a well-established significant transition point within the time series.

DOI: 10.1103/PhysRevE.84.021120 PACS number(s): 02.50.Tt, 02.50.Cw, 05.45.Tp, 92.70.Kb

I. INTRODUCTION

The estimation of change points challenges analysis methods and modeling concepts. Commonly change points are considered as isolated singularities in a regular background indicating the transition between two regimes governed by different internal dynamics. In time series scientists focus on change points in observed data to reveal dynamical properties of the system under study and to infer on possible correlations between subsystems. Detecting trend changes within various data sets is under intensive investigation in numerous research disciplines, such as palaeo-climatology [1,2], ecology [3,4], bioinformatics [5,6], and economics [7,8]. In general, the detection of transition points is addressed via (i) regression techniques [9,10], (ii) wavelet based methods [11–14], (iii) recurrence plot based techniques [15–17], or (iv) Bayesian approaches [18–22].

In this work we formulate transition points not only in terms of the underlying regular dynamics, but also as a transition in the heteroscedastic noise level. Our signal model is described by a regular mean undergoing a sudden change and a heteroscedastic fluctuation which undergoes as well a sharp transition at the same time point. We use Bayesian inference to derive estimates for all relevant parameters and their confidence intervals. By applying a kernel based approach, we formally localize the posterior density and the modeling of the subsignals as a linear trend is valid in first order. Consequently, we investigate time series globally and locally for a generalized break point in the signal's statistical properties.

In comparison to established methods our technique is not restricted to uniform time grids, for example, as required for (i) filtering or (ii) generally used wavelet transformations. The majority of existing methods rely on additional approaches to interpret the confidence of the outcome, for example, (i) bootstrapping, (ii) test statistics, or (iii) introducing measures. Whereas our technique provides confidence intervals of the estimates as a byproduct in a natural way. This for us is actually the most convincing argument to approach the detection task via Bayesian inference since besides the parameter estimation on its own, we obtain a degree of belief about our assumed model and about the uncertainties in the parameters [23–25].

Common techniques addressing Bayesian inference (iv) approach on the one hand the plain localization task of the singularity by treating the remaining model's parameters as hidden, for example, in hidden Markov models (HMM) of microarray data [26,27]. On the other hand, hierarchical Bayesian models are used mainly based on Monte Carlo expectation maximization (MCEM) algorithms for the estimation process [6,18,20].

In contrast, we aim for an insight in the parameter structure of the time series. We intend to detect multiple change points without enlarging the model's dimensionality, since this increases substantially the computational time. By addressing the general framework of linear mixed models [28] we are able to factorize the joint posterior density into a family of parametrized Gaussians. This mirrors the separation of the linear from the nonlinear parts and it simplifies considerably the explicit computation of the marginal distributions.

Our technique will be applied to a hydrological time series of the river Nile, which exhibits a well known change point.

II. SPECIFICATION OF THE ALGORITHM

In our modeling approach we consider two aspects of change points in a time series. On the one hand, a change point is commonly associated with a sudden change of local trend in the data. This indicates a transition point between two regimes governed by two different internal dynamics or external influences. On the other hand, we assume that the systematic evolution of the local variability of the data around its average value undergoes a sudden transition at the change point. As we will show, both aspects can be combined into a linear mixed model. Moreover, our formulation allows the separation of the Gaussian from the intrinsic nonlinear parts of the estimation problem, which besides clarifying the structure of the model, speeds up computations considerably. Based on our analytical calculations we accomplish the inference on the model's parameters by Bayesian inversion.

A. Formulation of the linear mixed model

The simplest type of signal undergoing a change point at time θ can be expressed as

$$y(t) = \beta_0 + \beta_1|\theta - t|_- + \beta_2|\theta - t|_+ + \xi(t). \quad (1)$$

Here we use piecewise linear basis functions [29,30], also called ramp functions, defined through

$$|\theta - t|_- = (\zeta_-^\theta) = \begin{cases} \theta - t & \text{if } t \leqslant \theta \\ 0 & \text{else} \end{cases} \tag{2}$$

and

$$|\theta - t|_+ = (\zeta_+^\theta) = \begin{cases} t - \theta & \text{if } t \geqslant \theta \\ 0 & \text{else} \end{cases}. \tag{3}$$

Natural data series can in general not be modeled by such a simple behavior as given by these functions. Therefore we add some random fluctuations ξ around the mean behavior. These random fluctuations can be due to measurement noise as well as to some intrinsic variability, which is not captured by the low dimensional mean dynamics on both sides of the change point θ. For this fluctuating part of the signal we suppose that its amplitude is essentially constant around the change point. However, the intrinsic variability may, like the mean behavior of the system itself, undergo a sudden change in its evolution of amplitude. Hence we consider stochastic fluctuations $\xi(t)$ whose amplitudes undergo a transition themselves according to

$$\mathrm{STD}(\xi(t)) = \sigma(1 + s_1|t-\theta|_- + s_2|t-\theta|_+). \tag{4}$$

The scale factor σ could be the level of the measurement noise or some background level of the intrinsic fluctuations, whereas the constants $s_{1,2}$ describe the systematic evolution of the model's intrinsic variability prior and after the change point measured in units of σ. Although the fluctuating component may naturally contain coherent parts, we assume throughout this work that the fluctuations are Gaussian random variables, which at different time points are uncorrelated,

$$\mathbb{E}(\xi(t)\xi(t')) = 0, \quad t \neq t'. \tag{5}$$

This clearly is an approximation and its validity can be questioned in concrete applications. However, this assumption allows us to implement highly efficient algorithms for the estimation of the involved parameters. From now on, we call this fluctuating part simply "noise." A realization of such a time series is presented in Fig. 1. Given a data set of n time points t_i, $i = 1,\ldots,n$, the observation vector $\boldsymbol{y} = [y(t_i)]^T \in \mathbb{R}^n$ can be written as follows:

$$\boldsymbol{y} = F\boldsymbol{\beta} + \boldsymbol{\xi}. \tag{6}$$

Here the fixed effect vector $\boldsymbol{\beta} = (\beta_0,\beta_1,\beta_2)^T \in \mathbb{R}^3$ corresponds to the coefficients of the linear combination of the ramp functions modeling the mean behavior. The system matrix of the fixed effects $F \in \mathbb{R}^{n\times 3}$ is then given by the sampling of the ramp functions $\boldsymbol{\zeta}_\pm^\theta$ defined in Eqs. (2) and (3) at the observation points t_i

$$F_\theta = \begin{pmatrix} 1 & (\zeta_-^\theta)_1 & (\zeta_+^\theta)_1 \\ \vdots & \vdots & \vdots \\ 1 & (\zeta_-^\theta)_n & (\zeta_+^\theta)_n \end{pmatrix}, \quad (\zeta_\pm^\theta)_i = (\zeta_\pm^\theta)(t_i). \tag{7}$$

The noise $\boldsymbol{\xi} \in \mathbb{R}^n$ is a Gaussian random vector with zero mean and covariance matrix $\Omega \in \mathbb{R}^{n\times n}$,

$$\boldsymbol{\xi} \sim \mathcal{N}(0,\sigma^2\Omega). \tag{8}$$

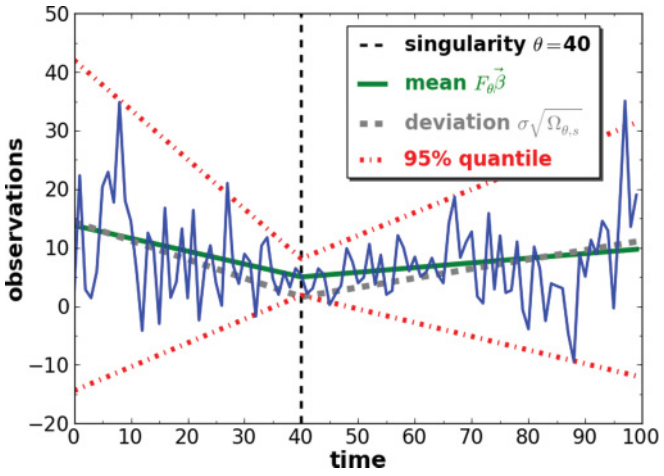

FIG. 1. (Color online) Realization of a time series of $n_{\mathrm{obs}} = 100$ data points generated by Eq. (6), whereas the mean is parametrized by $F_\theta\boldsymbol{\beta} = 5 + 0.22\,\boldsymbol{\zeta}_-^\theta + 0.08\,\boldsymbol{\zeta}_+^\theta$ and the variance is modeled as $\sigma^2\Omega_{\theta,s} = [1.6(1 + 0.2\,\boldsymbol{\zeta}_-^\theta + 0.1\,\boldsymbol{\zeta}_+^\theta)]^2$.

The covariance itself is structured noise, which is parametrized by the two slope parameters $\boldsymbol{s} = (s_1,s_2)$ and the change point θ itself as

$$\left(\Omega_{\theta,s_1,s_2}\right)_{ij} = ([1 + s_1(\zeta_-^\theta)_j + s_2(\zeta_+^\theta)_j]^2)\cdot\delta_{ij}. \tag{9}$$

In conclusion, the probability density of the observations for fixed parameters (i.e., fixed effects $\boldsymbol{\beta}$, change point θ, slope parameters $\boldsymbol{s}$) can be written as

$$\boldsymbol{y} \sim \mathcal{N}(F_\theta\hat{\boldsymbol{\beta}},\sigma^2\Omega_{\theta,s}). \tag{10}$$

The Likelihood function of the parameters given the data can then be formulated as

$$\mathcal{L}(\boldsymbol{\beta},\sigma,\boldsymbol{s},\theta|\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}\sqrt{|\Omega|}}\, e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-F\boldsymbol{\beta})^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta})}. \tag{11}$$

Note that the functional dependency of $\boldsymbol{\beta}$ is a Gaussian density. Clearly in the exponent $\boldsymbol{\beta}$ is of a quadratic form and since $\Xi = F^T\Omega^{-1}F$ is positive definite we may write

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}\sqrt{|\Omega|}}\, e^{-\frac{\mathcal{R}^2}{2\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\Xi(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}, \tag{12}$$

where the mode of the Gaussian in $\boldsymbol{\beta}$ is the best linear unbiased predictor of the fixed effects (BLUP) [31]

$$\begin{aligned}\boldsymbol{\beta}^* &= \underset{\boldsymbol{\beta}\in\mathbb{R}^3}{\operatorname{argmin}}(\boldsymbol{y}-F\boldsymbol{\beta})^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta}) \\ &= (F^T\Omega^{-1}F)^{-1}F^T\Omega^{-1}\boldsymbol{y},\end{aligned} \tag{13}$$

and the residuum $\mathcal{R}$ measured in the Mahalanobis distance [32], induced by the covariance matrix Ω, is

$$\begin{aligned}\mathcal{R}^2 &= \min_{\boldsymbol{\beta}\in\mathbb{R}^3}(\boldsymbol{y}-F\boldsymbol{\beta})^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta}) \\ &= (\boldsymbol{y}-F\boldsymbol{\beta}^*)^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta}^*).\end{aligned} \tag{14}$$

In addition, the profiled Likelihood function $\mathcal{L}(\boldsymbol{\beta}^*,\sigma,\boldsymbol{s},\theta|\boldsymbol{y})$ enables us to derive the profiled Likelihood estimator of the scale parameter

$$\hat{\sigma}^2 = \frac{\mathcal{R}^2}{n-3}, \tag{15}$$

which is auxiliary for the computation of the maximum of the Likelihood function.

B. Method of Bayesian inversion

In the light of the Bayesian theorem, we can compute the posterior distribution $p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}|\boldsymbol{y})$ of the modeling parameters given the data $\boldsymbol{y}$ from the Likelihood function Eq. (11) by specifying the prior distribution of the parameters $p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s})$, which encodes our belief about the parameters prior to any observation. Since we assume *a priori* no correlations between the parameters, the joint prior distribution can be factorized into the independent parts

$$p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}) = p(\boldsymbol{\beta})\, p(\sigma)\, p(\theta)\, p(\boldsymbol{s}). \tag{16}$$

In general, we do not have any *a priori* knowledge about these hyperparameters and thus we shall use flat and uninformative priors [33,34]

$$p(\boldsymbol{\beta}) \sim 1, \quad p(\theta) \sim 1, \quad p(\boldsymbol{s}) \sim 1. \tag{17}$$

For the scale parameter σ we assume a Jeffrey's prior [35]

$$p(\sigma) \sim \frac{1}{\sigma}. \tag{18}$$

These statistical assumptions enable us to compute the posterior density of the system's parameters given the data $\boldsymbol{y}$ as

$$p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}|\boldsymbol{y}) = C\, \mathcal{L}(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}|\boldsymbol{y})\, \frac{1}{\sigma}. \tag{19}$$

The normalization constant C ensures that the right-hand side actually defines a normalized probability density. From this expression, various marginal posterior distributions may be obtained by integrating over the parameters that shall not be considered. We are mostly interested in the posterior distribution of the possible change point location θ. To produce the posterior distribution of this quantity, we have to marginalize out all other variables. It turns out that all but the integral over the noise slopes $\boldsymbol{s}$ may be carried out explicitly. Thanks to the Gaussian nature of the $\boldsymbol{\beta}$ dependency we obtain

$$p(\sigma,\theta,\boldsymbol{s}|\boldsymbol{y}) \sim \frac{\sigma^{2-n}}{\sqrt{|\Omega||F^T\Omega^{-1}F|}} e^{-\frac{1}{2\sigma^2}\mathcal{R}^2}, \tag{20}$$

and

$$p(\boldsymbol{\beta},\theta,\boldsymbol{s}|\boldsymbol{y}) \sim \frac{[(\boldsymbol{y}-F\boldsymbol{\beta})^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta})]^{-\frac{n}{2}}}{\sqrt{|\Omega|}}. \tag{21}$$

Further marginalization may be performed to yield

$$p(\theta,\boldsymbol{s}|\boldsymbol{y}) = C' \frac{\mathcal{R}^{3-n}}{\sqrt{|\Omega||F^T\Omega^{-1}F|}}, \tag{22}$$

where again C' is a constant that ensures the normalization of the right-hand side to a probability density. Finally the posterior marginal distribution of θ can be computed by numerical evaluation of the following integral:

$$p(\theta|\boldsymbol{y}) = \int d\boldsymbol{s}\; p(\theta,\boldsymbol{s}|\boldsymbol{y}). \tag{23}$$

In the same way the posterior marginal distribution of the heteroscedastic slope parameters $\boldsymbol{s}$ can be obtained by numerical evaluation of the following integral:

$$p(\boldsymbol{s}|\boldsymbol{y}) = \int d\theta\; p(\theta,\boldsymbol{s}|\boldsymbol{y}). \tag{24}$$

Given that a transition θ can be realized by at least three time points, the maximal numerical sampling range of θ depicts the time series itself t_i, $i = 2,\ldots,n-1$, excluding the first t_1 and last t_n observation points. The numerical sampling range of the slope parameters $\boldsymbol{s}$ can roughly be estimated from the readily variability evolution of the data. This clearly is an approximation and its validity has to be reassessed by checking the marginal distribution $p(\boldsymbol{s}|\boldsymbol{y})$ for any artificial cut offs due to an inappropriate choice of the sampling range.

To evaluate the numerical marginalization in general, the numerical sampling grid of the parameters has to be selected based on the apparent behavior of the observations, reassessed by checking the marginal distributions and, if necessary, refined accordingly.

III. VALIDATION OF THE METHOD

In order to validate the performance of our algorithm in an idealized setting we generate time series using the linear mixed model of Eq. (6). We employ artificial data to discuss the method's ability to infer on the parameters of the underlying model and to elaborate the sensitivity of the estimated change points to data loss. Thereby the maxima of the marginal posterior densities can be regarded as the estimators of the most probable parameters, for example, the change point estimator $\hat{\theta} = [p(\theta|\boldsymbol{y})]_{\max}$. In case the observations contain more than one singularity, the posterior distribution will be multimodal. This could therefore be used as an indicator for the existence of secondary change points θ_i, $i = 1,2,\ldots,$ in the time series. Although a more reasonable way would be to consider models with multiple change points, this approach becomes quickly uncomputable due to exploding dimensionality. Thus we propose a local kernel based method to be able to apply our single change point model locally to multi change point data series.

A. Estimation of a single change point

To validate our technique, we apply it on generated time series of $n_{\text{obs}} = 100$ temporally equidistant observations containing a single change point at $\theta = 40$. We compute all relevant marginal posterior distributions of the model's parameters using the formulas of the previous section. The marginal densities provide Bayesian estimates for the change point $\hat{\theta}$, mean behavior $\hat{\boldsymbol{\beta}}$, scale parameter $\hat{\sigma}$, and heteroscedastic behavior $\hat{\boldsymbol{s}}$ of the data as their maxima over the numerical sampling grid.

Since the detection of less evident transitions in time series depicts a crucial concern in real data sets, we analyze the main model modifications. The most significant transition in our modeling approach is realized by a simultaneous change of mean and deviation as presented in Fig. 1. A less evident transition is given if alone the mean or alone the deviation undergoes a singularity. Thus we additionally generate uncorrelated changes in the behavior of the mean

(Fig. 5) and of the deviation (Fig. 6) to validate our algorithm on transitions of different observational evidence.

Note that due to the random nature of the observations, the posterior density too depends randomly on the actual series of observations. It is therefore not surprising that the maxima's locations of the posterior density does not exactly agree with the true parameter values. However, they are within a certain quantile of the posterior distribution. We automatically obtain confidence intervals or regions by considering those level intervals or contour lines that enclose a fixed percentage of the total probability. This yields a natural way of uncertainty quantification.

1. Transition in mean and deviation

The marginal posterior density $p(\theta,\boldsymbol{s}|\boldsymbol{y})$ of Eq. (22) can be derived by pure analytical integration. Thus the distribution depicts the most objective marginalization of the posterior density, since so far no numerical approaches need to be used to evaluate the integration. Consequently, we begin the inference by estimating the change point θ and the slope parameters $\boldsymbol{s}$ of the deviation. The numerical sampling grid of the singularity is chosen as maximal, that means we assume all time points t_i of the series as possible locations of the change point, except the first and last one. To avoid side effects we ignore the first and last five data points, such that the sampling grid for the singularity becomes $\Delta_\theta = [5.0,95.0]$. For the slope parameters we guess a range from the obvious variability of the data as $\Delta_s = [-0.2,0.6]$.

One of the resulting marginal posterior densities $p(\theta|\boldsymbol{y})$ is shown in Fig. 2 representing our degree of belief that a change point θ occurs at a time point t_i. The most likely change point $\hat{\theta} = 40.5$ is given by the position of the distribution's maximum. The estimator $\hat{\theta}$ lies in a relatively symmetric 95% confidence interval [36.0,45.0]. Thus we achieve to restrict the location of a probable singularity to 10% of the assumed sampling range Δ_θ. The other marginal density $p(s_1,s_2|\boldsymbol{y})$ we obtain is shown as a contour plot in Fig. 3. The posterior density in the plane of the slope parameters indicates a heteroscedastic behavior of the time series since the area of the most probable slope combinations (s_1,s_2) does not enclose equal values for both parameters (red area in contour plot of Fig. 3). The analysis of the projections $p(s_1|\boldsymbol{y})$ and $p(s_2|\boldsymbol{y})$ yields unidirectional wider 95% confidence bounds for the estimates $\hat{s}_1 = 0.170$ in [0.050,0.420] and $\hat{s}_2 = 0.080$ in [0.030,0.210] to increasing parameter values.

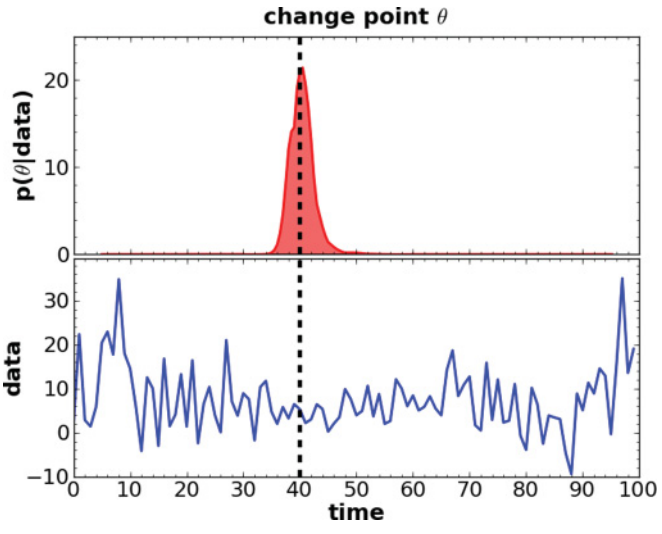

FIG. 2. (Color online) Marginal posterior density $p(\theta|\boldsymbol{y})$ for the time series in Fig. 1, represented in the lower panel. The maximum indicates the estimate of the change point $\hat{\theta} = 40.5$, whereas the dashed line marks the true parameter value of the underlying model.

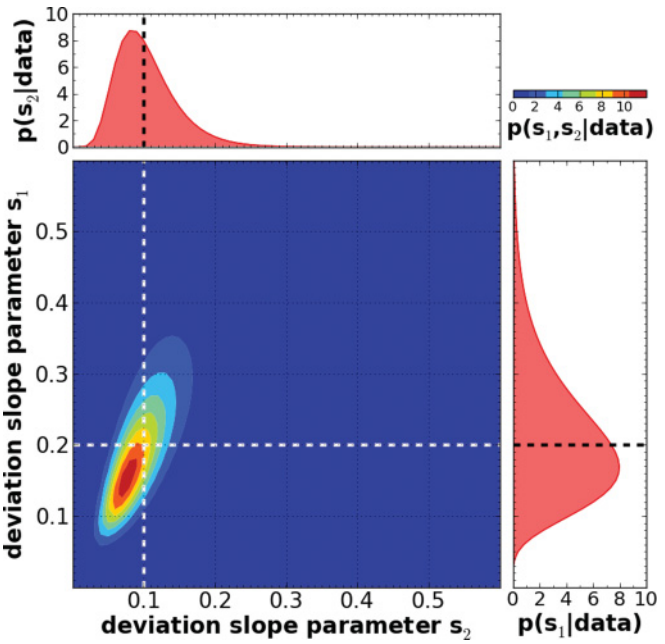

FIG. 3. (Color online) Marginal posterior density $p(s_1,s_2|\boldsymbol{y})$ for the time series in Fig. 1. Alongside are presented the projected posterior densities $p(s_1|\boldsymbol{y})$ and $p(s_2|\boldsymbol{y})$. The maxima indicate the estimates of the slope parameters $(\hat{s}_1,\hat{s}_2) = (0.170,0.080)$, whereas the dashed lines mark the true values of the underlying model.

The remaining parameters of the mean $\boldsymbol{\beta}$ and the deviation scale σ can be estimated by computing the marginal posterior distributions of Eqs. (20) and (21). Again we assume a range from the obvious behavior of the data's mean and variability to evaluate the numerical integration. By integrating over the parameters that shall not be considered, we derive the most likely scale parameter $\hat{\sigma} = 1.46$ within the 95% confidence interval $[\,0.76\,,\,2.60\,]$ unidirectional wider to growing σ values. The offset of the mean can be estimated as $\hat{\beta}_0 = 4.40$ lying in the 95% confidence interval [2.90,5.90]. The marginal density $p(\beta_1,\beta_2|\boldsymbol{y})$ of the mean slopes is shown as a contour plot in Fig. 4. The posterior density in the plane of the slopes indicates a trend change in the mean of the time series since the area of the most probable slope combinations (β_1,β_2) does not enclose equal values for both parameters (red area in contour plot of Fig. 4). The projections $p(\beta_1|\boldsymbol{y})$ and $p(\beta_2|\boldsymbol{y})$ reveal relatively symmetric 95% confidence bounds for the estimates $\hat{\beta}_1 = 0.198$ in [0.033,0.376] and $\hat{\beta}_2 = 0.088$ in [0.005,0.184].

In conclusion, our Bayesian estimates reproduce the real underlying model convincingly as summarized in Table I and listed with the corresponding confidence intervals based on our analysis.

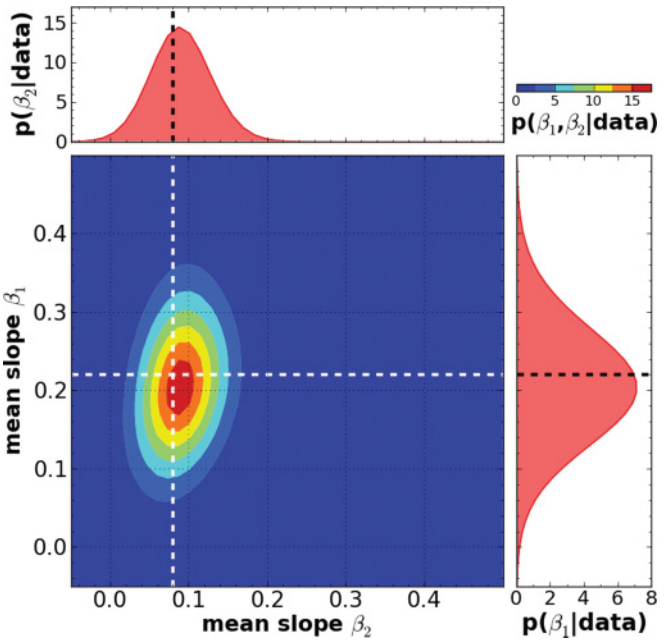

FIG. 4. (Color online) Marginal posterior density $p(\beta_1,\beta_2|\boldsymbol{y})$ for the time series in Fig. 1. Alongside are presented the projected posterior densities $p(\beta_1|\boldsymbol{y})$ and $p(\beta_2|\boldsymbol{y})$. The maxima indicate the estimates of the slopes $(\hat{\beta}_1,\hat{\beta}_2) = (0.198,0.088)$, whereas the dashed lines mark the true values of the underlying model.

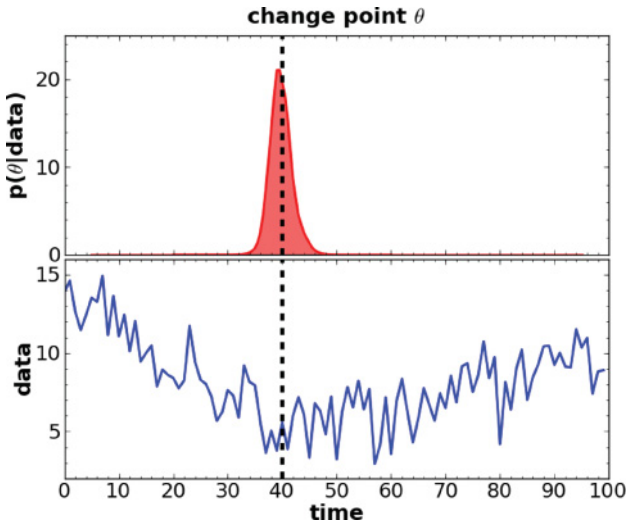

FIG. 5. (Color online) Marginal posterior density $p(\theta|\boldsymbol{y})$ for a time series where only the mean undergoes a transition, represented in the lower panel. The maximum indicates the estimate of the change point $\hat{\theta} = 39.5$, whereas the dashed line marks the true parameter value of the underlying model.

2. *Transition only in mean or only in deviation*

Based on the model of Fig. 1 we modify the parameters to produce uncorrelated transition events in mean or deviation. By setting the deviation's slope parameters to zero, $\boldsymbol{s} = (0,0)$, we realize a time series where only the mean contains a sharp break (Fig. 5). In the same way we produce an artificial time series where only the deviation undergoes a change by keeping the mean trend unchanged over time, $(\beta_1,\beta_2) = (-0.08,0.08)$ (Fig. 6). Note that the alternating sign is contributed to the definition of the ramp functions in Eqs. (2) and (3). We apply our algorithm as described in the previous section and obtain the estimates with corresponding confidence intervals as listed in Table I.

The comparison of the detected change points and their confidence intervals supports the capability and efficiency of our algorithm to perform a convincing estimation of all types of underlying singularities. However, in the setting where alone the deviation is changing the corresponding marginal distribution $p(\theta|\boldsymbol{y})$ is in general broader than in the other settings. This is therefore not surprising since the sole transition in the deviation is less evident in the observations, respectively, more uncertain to detect.

The remaining model's parameters are as well estimated for all transition types convincingly. Moreover, our algorithm infers correctly on a homoscedastic (Fig. 5) or heteroscedastic (Fig. 6) time series by discretizing between equal or unequal estimates of the deviation's slope parameters from analyzing $p(s_1,s_2|\boldsymbol{y})$. The same conclusion follows for trend changes in the mean of the observations by studying the marginal distribution $p(\beta_1,\beta_2|\boldsymbol{y})$. Thus our method enables us to infer substantially on the type of transition underlying the data, that means we are able to distinguish between a singularity alone in the mean or alone in the variability or in both properties of the time series.

TABLE I. Estimated underlying models of the generated time series in Figs. 2, 5, and 6. The asterisk indicates modified parameter values and corresponding estimates due to the different underlying transition models.

	Mean & deviation changes (Fig. 2)		Only mean changes (Fig. 5)		Only deviation changes (Fig. 6)	
Parameter	Estimate	Confidence $\geqslant 95\%$	Estimate	Confidence $\geqslant 95\%$	Estimate	Confidence $\geqslant 95\%$
$\boldsymbol{\theta} = 40.0$	40.5	[36.0,45.0]	39.5	[35.5,44.5]	40.5	[36.0,46.0]
$\boldsymbol{\beta}_0 = 5.0$	4.40	[2.90,5.90]	4.90	[4.15,5.65]	4.90	[3.70,6.10]
$\boldsymbol{\beta}_1 = 0.22\ (-0.08)^*$	0.198	[0.033,0.376]	0.228	[0.184,0.271]	-0.063^*	$[-0.200,0.075]$
$\boldsymbol{\beta}_2 = 0.08$	0.088	[0.005,0.184]	0.088	[0.052,0.114]	0.063	[0.000,0.138]
$\boldsymbol{\sigma} = 1.6$	1.46	[0.76,2.60]	1.50	[1.10,2.05]	1.18	[0.58,2.15]
$\boldsymbol{s}_1 = 0.2\ (0.0)^{**}$	0.170	[0.050,0.420]	-0.007^{**}	$[-0.017,0.011]$	0.170	[0.060,0.430]
$\boldsymbol{s}_2 = 0.1\ (0.0)^{**}$	0.080	[0.030,0.210]	-0.001^{**}	$[-0.009,0.013]$	0.090	[0.020,0.220]

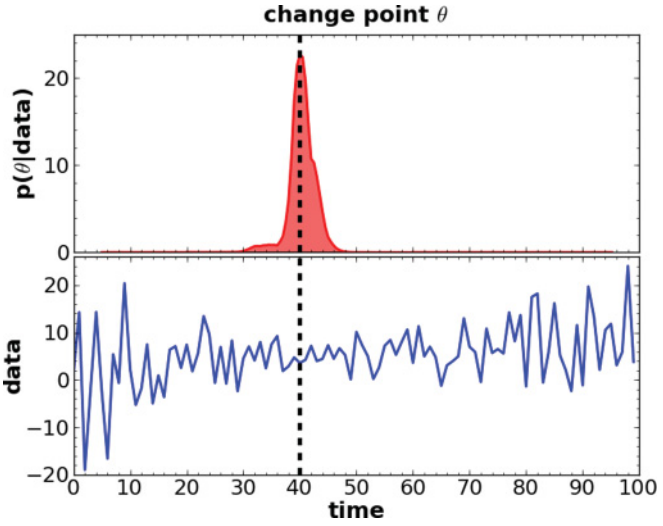

FIG. 6. (Color online) Marginal posterior density $p(\theta|\boldsymbol{y})$ for a time series where only the deviation undergoes a transition, represented in the lower panel. The maximum indicates the estimate of the change point $\hat{\theta} = 40.5$, whereas the dashed line marks the true parameter value of the underlying model.

3. *Sensitivity to data loss*

In real time series, analysts have to deal with sparse and irregularly sampled data. Our technique does not require an uniform sampling grid of data points since from the beginning it employs only the available data. As a validation for the sensitivity of our method to data loss, we generate a temporally equidistant time series modeled by a sequence of $n'_{\text{obs}} = 300$ observations. The observations undergo a transition in mean

$$F_\theta \boldsymbol{\beta} = 12 + 0.24\,\boldsymbol{\zeta}^\theta_- + 0.02\,\boldsymbol{\zeta}^\theta_+ \tag{25}$$

and variance

$$\sigma^2 \Omega_{\theta,s} = [1.2(1 + 0.18\,\boldsymbol{\zeta}^\theta_- + 0.04\,\boldsymbol{\zeta}^\theta_+)]^2 \tag{26}$$

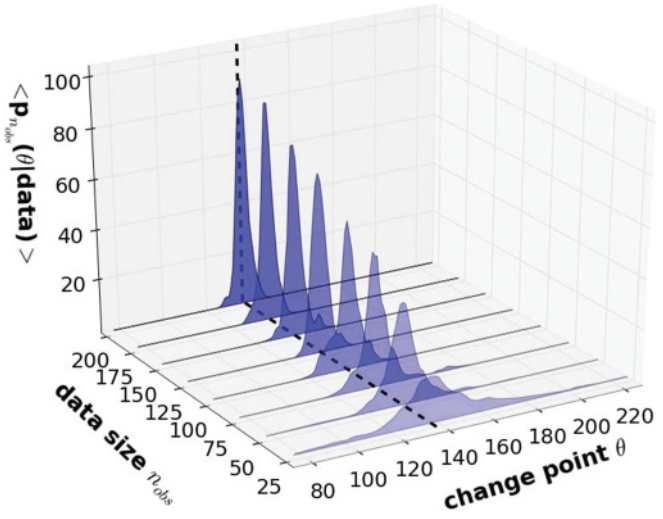

FIG. 7. (Color online) The global maxima of the averaged posterior densities $\langle p_{n_{\text{obs}}}(\theta|\boldsymbol{y})\rangle$ converge for increasing number of data points n_{obs} toward a delta distribution located at the true change point value $\theta = 135$ (dashed line).

at the time point $\theta = 135$. Then we resample the data with two different sampling frequencies changing within the time interval [100,200], but not at the true singularity. The irregularly sampled time series contains $n''_{\text{obs}} = 200$ data points afterwards. We randomly ignore stepwise 0 up to 175 data points of the time series. Leaving out randomly a defined percentage of the observations produces time series with random gaps and irregular sampling steps. For each of these random realizations consisting of $n_{\text{obs}} = 25, \ldots, 200$ data points we compute the posterior densities $p^i_{n_{\text{obs}}}(\theta|\boldsymbol{y})$ for $i = 1, \ldots, 60$ realizations.

The obtained posterior densities $\langle p_{n_{obs}}(\theta|\boldsymbol{y})\rangle$ averaged over the realizations are shown in the plane of the sample size n_{obs} in Fig. 7. Thus their maxima indicate the most probable change points $\langle \hat{\theta}_{n_{\text{obs}}}\rangle$ averaged over the realizations. Apparently the mean of the posterior densities differs from the true value, however still within the width of the distribution. The latter depends inversely proportional on the square root of the sample size

$$\text{width}[\langle p_{n_{\text{obs}}}(\theta|\boldsymbol{y})\rangle] \propto \frac{1}{\sqrt{n_{\text{obs}}}}. \tag{27}$$

At large numbers of sampling points n_{obs} the posterior converges toward a delta distribution located at the true parameter value $\theta = 135$. In any case, even for small data sets, as small as $n_{\text{obs}} = 25$, the nonflatness of the posterior clearly hints toward the existence of a change point in the time series. The investigation of the averaged marginal posterior densities in the plane of the remaining parameters reveals a broadening of the posterior distributions for $n_{\text{obs}} < 200$, as naturally expected due to information loss in the resampled time series considered in the inference process.

Additionally, we point out the efficiency of our method to infer on the explicit location of a singularity $\hat{\theta}^i_{n_{\text{obs}}}$ for every single time series of the previous setting. In Fig. 8 the histograms of the global point estimators $\hat{\theta}^i_{n_{\text{obs}}}$ for every single realization $i = 1, \ldots, 60$ are presented. We observe that the particular global estimators $\hat{\theta}^i_{n_{\text{obs}}}$ are relatively robust to data loss and enable us to infer convincingly on the location of the singularity. Even by considering only one third of the full time series, that is, $n_{\text{obs}} = 100$, our algorithm performs global estimates lying in a narrow interval [121,148]. This time interval represents <18% of the chosen time range $\Delta_\theta = [75,225]$ of the numerical marginalization.

However, for such a data-poor situation, local additional, less dominant maxima are likely to appear due to random fluctuations in the posterior distribution, and more sophisticated techniques are needed to assess the existence of single or multiple change points. One approach to clarify multimodal posterior densities is the computation of local posterior densities within a sliding window as presented in the following.

B. Estimation of multiple change points

Long data sets are likely to contain more than one change point and using our algorithm globally may not be justified. However, locally our model assumption may still be valid. For this reason, we propose the following kernel based local posterior method. In addition, this method allows us to treat

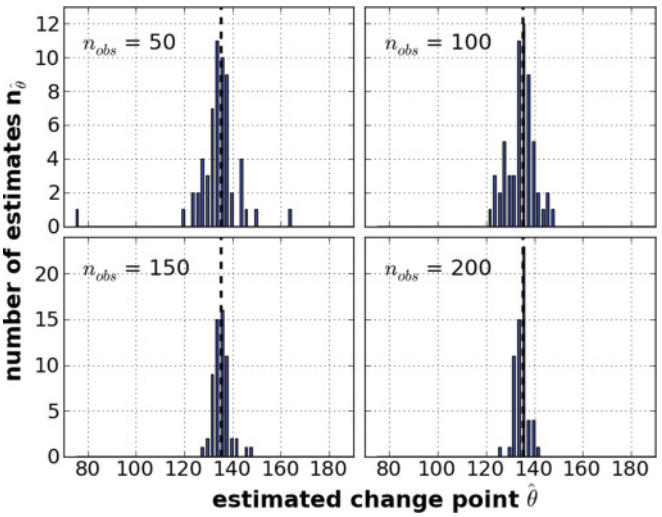

FIG. 8. (Color online) Histograms of the global change point estimators $\hat{\theta}^i_{n_{\rm obs}}$ for $i = 1, \ldots, 60$ realizations and with respect to $n_{\rm obs}$ data points from the setting of Fig. 7. Even for $n_{\rm obs} = 100$ all global estimates $\hat{\theta}^i_{100}$ lie in the relatively narrow interval $[121, 148]$ around the true singularity $\theta = 135$ (dashed lines).

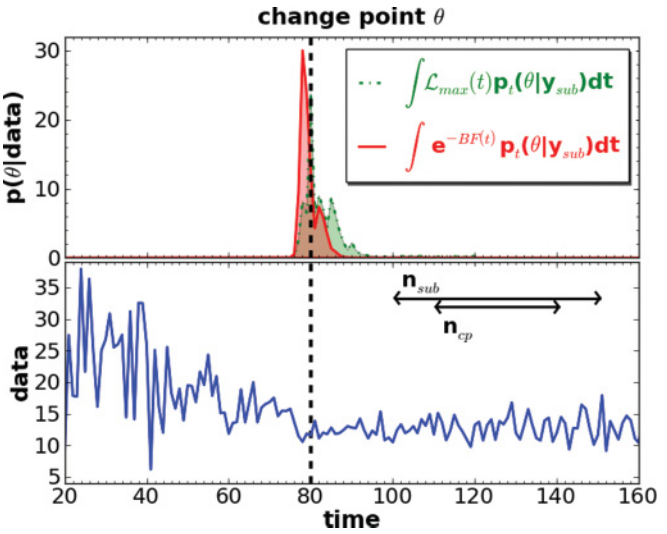

FIG. 9. (Color online) Sum of local posterior densities weighted by the local Likelihood maxima (dash-dotted) and with respect to the Bayes factor (solid), computed for windows of $n_{\rm sub} = 50$ data points and $n_{\rm cp} = 30$ change points on uniform sampling grids. The dashed line marks the true singularity $\theta = 80$.

very long data sets numerically more efficient since the computation scales with the third power of the number of employed data points.

Around each time point t we choose a data window $I_t = [t - \frac{T}{2}, t + \frac{T}{2}]$ of length T. Inside this window we take as prior distribution for the change point location $p_t(\theta)$ a flat prior inside some subinterval of length a:

$$p_t(\theta) = \begin{cases} \frac{1}{a} & \text{for } t - \frac{a}{2} \leqslant \theta \leqslant t + \frac{a}{2} \\ 0 & \text{else} \end{cases}, \quad 0 < a < T. \tag{28}$$

The resulting data windows $\boldsymbol{y}_{|I_t}$ may be interpreted as kernels of neighborhood I_t around the target point t and the weighting function $p_t(\theta)$ (based on [29]).

We then compute the local posterior $p_t(\theta|\boldsymbol{y}_{|I_t})$ around t based on the subseries $\boldsymbol{y}_{|I_t}$. This yields a posterior distribution of a possible change point within each kernel under the assumption that each kernel actually contains a singularity. In order to compare different kernel locations, we need to quantify the credibility that there exists a change point. Therefore we compute the maximum of the Likelihood within each kernel

$$f(t) = \max_{\theta \in [t - \frac{a}{2}, t + \frac{a}{2}], s_1, s_2 \in \mathbb{R}} \mathcal{L}(\boldsymbol{\beta}^*, \hat{\sigma}; \boldsymbol{y}_{|I_t}), \tag{29}$$

where $\hat{\sigma}$ and $\boldsymbol{\beta}^*$ are the estimators given by Eqs. (15) and (13). The global distribution of change points θ given the full time series is therefore obtained as a weighted superposition in form of

$$p(\theta|\boldsymbol{y}) = C \int f(t)\, p_t(\theta|\boldsymbol{y}_{|I_t}) dt, \tag{30}$$

whereas the constant C ensures the normalization to a probability density. In subdata sets with no change point, the credibility of the model fit is very low, in conclusion the Likelihood maxima is of very small value and local estimates are judged as negligible. By construction the method works for multiple change points as soon as they are separated by at least one data window. We demonstrate this by applying our algorithm first on a generated single change point time series of $n_{\rm obs} = 200$ temporally equidistant data points. The underlying model is parametrized through Eqs. (25) and (26), whereas the change point occurs at $\theta = 80$. In Fig. 9 the sum of the local posterior densities weighted by the maxima of the local Likelihood (dash-dotted curve) is shown. The applied window size $n_{\rm sub} = 50$ and the sampling range of the change points $n_{\rm cp} = 30$ are presented for comparison. The sum of local posterior densities indicates the best model fit for windows covering the real change point $\theta = 80$ but is nonzero even between $[100, 121]$ suggesting that a change point model might be suitable for these singularity values as well.

A second quantity that may be used to produce relative credibility weights for the windows is given by the Bayes factor [36]. Besides the goodness of fit, the complexity of the assumed model has to be taken into account to assess the most capable model describing the data and thus performing the estimation. Therefore we test the hypothesis of no change point, respectively, a linear model $\mathcal{M}_{\rm lin}$, against a change point model $\mathcal{M}_{\rm cp}$ in form of the Bayes factor

$$BF(t) = \frac{p(\mathcal{M}_{\rm lin}|\boldsymbol{y}_{|I_t})}{p(\mathcal{M}_{\rm cp}|\boldsymbol{y}_{I_t})}. \tag{31}$$

The dependency of the Bayes factor on a logarithmic scale is shown in Fig. 10 for the generated time series of Fig. 9. The Bayes factor in this test case favors the change point over the linear model for all windows, for which the true change point is in the support of the inner prior distribution of θ. This local Bayes factor itself can be used as a diagnostic tool like the Likelihood weighted posterior, but we may also combine the techniques by using the Bayes factor as a kernel weighting function by setting $f(t) = e^{-BF(t)}$ in Eq. (30). In this form Eq. (30) corresponds essentially to the total

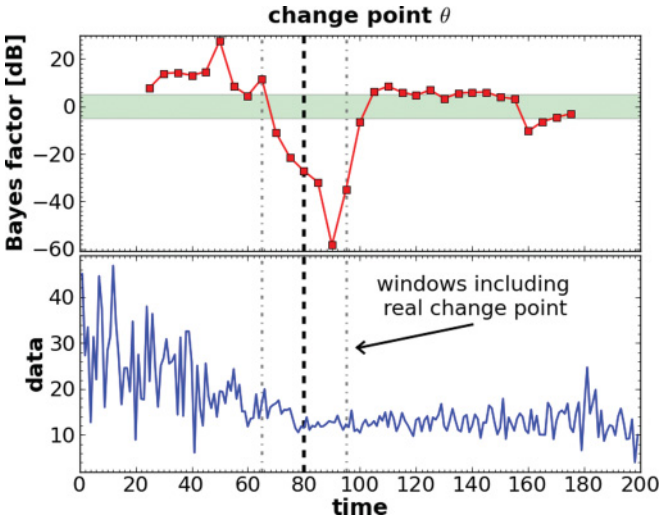

FIG. 10. (Color online) Local Bayes factor (squares) scaled in deciban and obtained for the time series in Fig. 9. The shaded area encloses values whose support for none of the models is substantial (based on [36]). Values underneath this area strongly support a change point against a linear model, and vice versa.

probability decomposition of the change point (cp)

$$\sum_{\text{windows}} p(\theta|\text{cp in window})\, p(\text{cp exists in window}). \qquad (32)$$

For comparison of both kernel approaches, we present in Fig. 9 additionally the sum of local posterior densities weighted by $e^{-BF(t)}$ (solid curve). The distribution weighted with respect to the Bayes factor is nonzero in the range between [78,89], whereas the one weighted by the maxima of the Likelihood is nonzero in [78,121]. The long tail of the latter hints to less probable change point locations which are automatically rejected in the Bayes factor weighting. Finally we exemplify the algorithm on a generated multi change point time series shown in Fig. 11. For clarity of presentation we plot the sum of posterior distributions weighted with the plain Bayes factor BF. We are able to infer on the true change point values $(\theta_1,\theta_2,\theta_3) = (40,100,160)$ via the estimators $(\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_3) = (38.9,93.0,162.9)$ within their intervals ([33.9,47.8],[87.5,109.1],[158.9,167.0]) of about 90% confidence. We obtain these intervals from a more detailed analysis of the partial sums of local posterior densities weighted by the factor e^{-BF} enclosing the estimated singularity locations.

The main advantage of this localization approach even in a single change point context is however the enormous speedup of the computations. For instance, for a time series of $n_{\text{obs}} = 2000$ data points we pass from a global computation of the marginalized posterior density in 3 h 41 min 40 s to a local one divided into 40 overlapping subdata sets of $n_{\text{sub}} = 100$ in 7 min 44 s, a reduction in computation time by about 95% respectively. This is achieved using Python 2.6.5 on a Supermicro Intel(R) Core(TM)i7 CPU 920 @ 2.68 GHz with 12 GB RAM. In the context of complex multiple change point scenarios, as real time series mostly are, the localization approach of the posterior density $p(\theta|\boldsymbol{y})$ combined with the Bayes factor realizes a powerful tool to scan the data separately for single change points, as implemented in the following.

C. Annual Nile flow from 1871 to 1970

We demonstrate our technique by applying it on a hydrological time series including a known significant change point. For this purpose, we analyze the annual Nile River flow measured at Aswan from 1871 to 1970 [37]. Several investigation methods have verified a shift in the flow levels starting from the year 1899 [4,21,37]. Historical records provide the fact that this shift is attributed partly to weather changes and partly to the

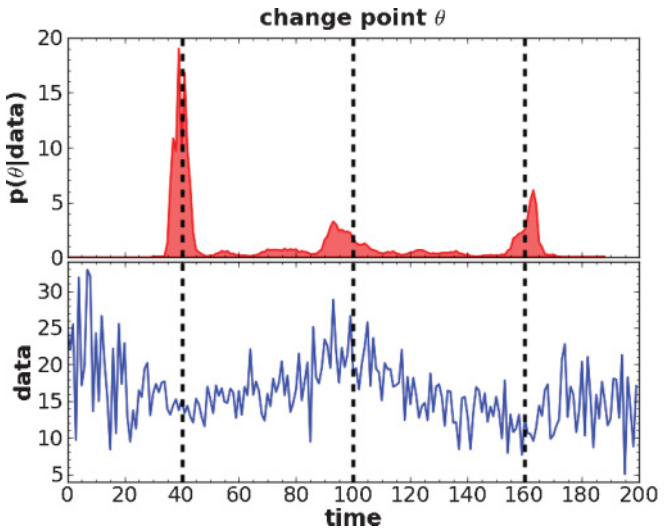

FIG. 11. (Color online) Sum of local posterior densities weighted by the Bayes factor, computed for sub time series of $n_{\text{sub}} = 50$ data points and $n_{\text{cp}} = 30$ change points on uniform sampling grids. The parametrization of the mean is defined as $F\boldsymbol{\beta} = 14 + 0.2\,\zeta_-^{40} + 0.1\,\zeta_+^{40} - 0.25\,\zeta_+^{100} + 0.3\,\zeta_+^{160}$ and the variance as $\sigma^2\Omega = [1.6(1 + 0.2\,\zeta_-^{40} + 0.03\,\zeta_+^{40} - 0.05\,\zeta_+^{100} + 0.1\,\zeta_+^{160})]^2$.

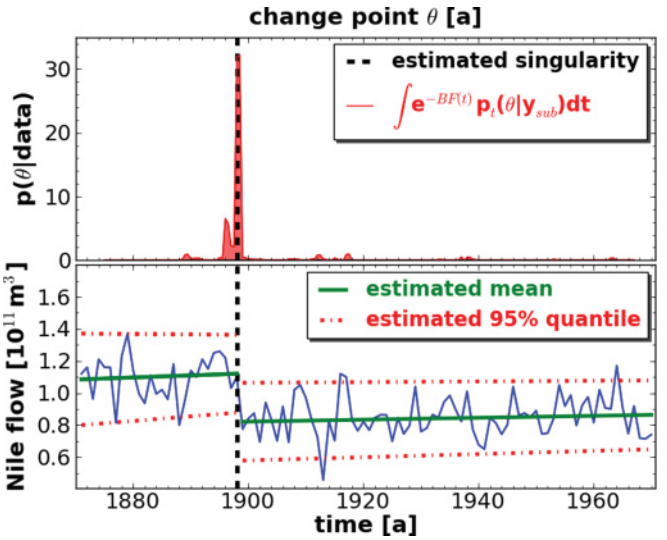

FIG. 12. (Color online) Annual Nile flow containing a known change point at $\theta = 1899$. The sum of localized posterior densities weighted with respect to the Bayes factor indicates a change point at $\hat{\theta} = 1898$ within its confidence interval [1895,1901] of about 90%. The estimated underlying model reveals the most dominant transition in the behavior of the mean.

TABLE II. Estimated underlying model of the annual Nile flux based on the inversion of the global posterior distribution.

Parameter	Estimate	Confidence $\geqslant 90\%$
$\hat{\theta}$	1898	[1895,1901]
$\hat{\beta}_0$	1.12	[1.01,1.22]
$\hat{\beta}_1$	−0.0013	[−0.0082,0.0057]
$\hat{\beta}_2$	0.0006	[−0.0011,0.0024]
$\hat{\beta}_3$	0.82	[0.76,0.90]
$\hat{\sigma}$	0.124	[0.094,0.160]
$\hat{s}_1$	0.0065	[−0.0190,0.0450]
$\hat{s}_2$	−0.0016	[−0.0065,0.0855]

start of construction work for a new dam at Aswan. Since we expect a natural behavior of the underlying mean we generalize our previous model to undergo besides trend changes as well a sharp shift in the mean offset at the singularity θ. Therefore we modify the system matrix according to

$$F_\theta = \begin{pmatrix} (\varphi_-^\theta)_1 & (\zeta_-^\theta)_1 & (\zeta_+^\theta)_1 & (\varphi_+^\theta)_1 \\ \vdots & \vdots & \vdots & \vdots \\ (\varphi_-^\theta)_n & (\zeta_-^\theta)_n & (\zeta_+^\theta)_n & (\varphi_+^\theta)_n \end{pmatrix}, \qquad (33)$$

whereas we define piecewise constant basis functions $\boldsymbol{\varphi}_\pm^\theta$ referring to Eqs. (2) and (3) not as linear but as constant, that is, Heaviside functions, respectively. However, the formulas of the Bayesian inference remain the same.

First of all, we compute the global posterior density $p(\theta,\boldsymbol{s}|\boldsymbol{y})$ as presented in Eq. (22). By initially guessing a reasonable sampling grid for the change point θ and the slope parameters $\boldsymbol{s}$ from the observations, we clearly obtain significant maxima in the posterior projections $p(\theta|\boldsymbol{y})$ and $p(\boldsymbol{s}|\boldsymbol{y})$. Therefore we adjust the sampling grid to obtain finer posterior structures around the evident maxima. We estimate the change point as $\hat{\theta} = 1898$ within a 95% confidence interval [1895,1901]. The slope parameters of the deviation are estimated as $(\hat{s}_1,\hat{s}_2) = (0.0065, -0.0016)$ within the 90% confidence intervals $\hat{s}_1$ in [−0.0190,0.0450] and $\hat{s}_2$ in [−0.0065,0.0855]. Prior the estimators $\hat{\theta}$ and $\hat{\boldsymbol{s}}$ we compute the posterior projections $p(\boldsymbol{\beta},\theta,\boldsymbol{s}|\boldsymbol{y})$ and $p(\sigma,\theta,\boldsymbol{s}|\boldsymbol{y})$ formulated in Eqs. (21) and (20). By reducing the numerical sampling grid of θ and $\boldsymbol{s}$ to its confidence intervals we are able to speed up the computation and to estimate the remaining parameters $\boldsymbol{\beta}$ and σ. We reveal from the global posterior distribution the most probable model plotted in Fig. 12 (lower panel) and listed in Table II. The estimated underlying model indicates that the most dominant transition in the observations occurs in the behavior of the mean, whereas the change in the deviation trend is negligible.

Finally we investigate the time series for local singularities by computing the sum of local posterior densities weighted with respect to the Bayes factor as e^{-BF} [displayed in Fig. 12 (upper panel)] for the kernel size $50a$. The change point sampling grid contains $n_{\mathrm{cp}} = 30$ steps in a resolution of $\Delta\theta = 0.5a$. Since all secondary maxima are about a factor 30 smaller than the global maxima we ignore them and therefore conclude on one global change point at $\hat{\theta} = 1898$ in the interval [1895,1901] of about 90% confidence. Note that we interpret the splitting of the global maximum as an artifact from the high resolution of the numerical change point sampling $\Delta\theta = 0.5a$.

In conclusion, we are able to confirm previous investigation techniques and reveal auxiliary information from the parameter space of the multidimensional posterior density of the applied linear mixed model.

IV. CONCLUSIONS

We introduce a general Bayesian algorithm for the detection of trend changes in heteroscedastic time series by describing the observations as a linear mixed model. The change point is thereby considered as an isolated singularity in a regular background of a signal, assuming partial linear mean and deviation in the first-order approach. By addressing the framework of linear mixed models we simplify the explicit computation of the marginal posterior distributions and thus reduce the computational time considerably. The formulation of the marginalized posterior densities of the model's parameters enables us to obtain *inter alia* the probability density of a change point given the data. Therefore our technique yields an insight in the parameter space of the underlying model by estimating the parameters and intrinsically providing their confidence intervals.

We elaborate our technique for single change point models of different observational transition evidence, infer on the model parameters, and discuss the sensitivity of the singularity estimator with respect to data loss. Additionally, we present a kernel based approach to investigate more complex time series containing multiple change points by localizing the posterior density and using the Bayes factor as a weighting function.

Moreover, we apply our algorithm on the annual flow volume of the Nile River at Aswan from 1871 to 1970. We confirm a well-established transition in the year 1899 by the estimated change point at 1898 within the interval [1895,1901] of about 90% confidence. We specify the underlying model and identify the mean as the statistical property undergoing the most significant transition.

We conclude by emphasizing that our algorithm realizes a powerful tool in estimating the location of transitions in heteroscedastic time series and inferring on the underlying behavior in a partial linear approach, meanwhile reducing the computational time.

ACKNOWLEDGMENTS

We thank M. H. Trauth for fruitful discussions and gratefully acknowledge financial support by DFG (GRK1364) and the University of Potsdam.

[1] M. H. Trauth, J. C. Larrasoaña, and M. Mudelsee, Quat. Sci. Rev. **28**, 399 (2009).
[2] M. Mudelsee and M. E. Raymo, Paleoceanography **20**, PA4022 (2005).
[3] M. P. Girardin *et al.*, Global Change Biol. **15**, 2751 (2009).
[4] P. Jong and J. Penzer, J. Am. Stat. Assoc. **93**, 796 (1998).
[5] V. N. Minin, K. S. Dorman, F. Fang, and M. A. Suchard, Bioinformatics **21**, 3034 (2005).

[6] J. S. Liu and C. E. Lawrence, Bioinformatics **15**, 38 (1999).
[7] P. Li and B. H. Wang, Physica A (Amsterdam) **378**, 519 (2007).
[8] D. W. K. Andrews, Econometrica **61**, 821 (1993).
[9] M. Mudelsee, Eur. Phys. J. Special Topics **174**, 49 (2009).
[10] R. Lund and J. Reeves, J. Clim. **15**, 2547 (2002).
[11] Y. Zhou, A. T. K. Wan, S. Xie, and X. Wang, J. Econometrics **159**, 183 (2010).
[12] L. R. Olsen, P. Chaudhuri, and F. Godtliebsen, Comput. Stat. Data Anal. **52**, 3310 (2008).
[13] A. Antoniadis and I. Gijbels, Nonparametric Stat. **14**, 7 (2002).
[14] F. Abramovich, T. C. Bailey, and T. Sapatinas, The Statistician **49**, 1 (2000).
[15] J. B. Gao, Yinhe Cao, Lingyun Gu, J. G. Harris, and J. C. Principe, Phys. Lett. A **317**, 64 (2003).
[16] R. V. Donner, Y. Zou, J. F. Donges, N. Marwan, and J. Kurths, New J. Phys. **12**, 033025 (2010).
[17] N. Marwan, J. F. Donges, Y. Zou, R. V. Donner, and J. Kurths, Phys. Lett. A **373**, 4246 (2009).
[18] E. Moreno, G. Casella, and A. Garcia-Ferrer, Stoch. Environ. Res. Risk Assess **19**, 191 (2005).
[19] Y. C. Tai, M. N. Kvale, and J. S. Witte, Biometrics **66**, 675 (2010).
[20] H. Lian, e-print arXiv:0709.1309v1.
[21] A. B. Downey, e-print arXiv:0812.1237.
[22] P. Fearnhead, Stat. Comput. **16**, 203 (2006).
[23] G. D'Agostini, Rep. Prog. Phys. **66**, 1383 (2003).
[24] D. M. Bates and S. DebRoy, J. Multivariate Anal. **91**, 1 (2004).
[25] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd edition (Chapman & Hall/CRC, New York, 2004).
[26] C. Yau, O. Papaspiliopoulos, G. O. Roberts, and C. Holmes, J. R. Stat. Soc. **73**, 37 (2011).
[27] M. Yuan and C. Kendziorski, J. Am. Stat. Assoc. **101**, 1323 (2006).
[28] M. E. McCulloch, S. R. Searle, and J. M. Neuhaus, *Generalized, Linear, and Mixed Models*, 2nd ed. (Wiley, New York, 2008).
[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer, New York, 2001), pp. 117–120 and pp. 165–183.
[30] J. H. Friedman, Annals Stat. **19**, 1 (1991).
[31] G. K. Robinson, Stat. Sci. **6**, 15 (1991).
[32] P. C. Mahalanobis, Proc. Natl. Inst. Sci. **2**, 49 (1936).
[33] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science* (Cambridge University Press, 2003), pp. 149–179.
[34] G. Wahba, J. R. Stat. Soc. Ser. B **40**, 364 (1978).
[35] H. Jeffreys, Proc. R. Soc. London Ser. A **186**, 453 (1946).
[36] R. E. Kass and A. E. Raftery, J. Am. Stat. Assoc. **90**, 773 (1995).
[37] G. W. Cobb, Biometrika **65**, 243 (1978).