On robust estimation of low-frequency variability trends in discrete Markovian sequences of atmospheric circulation patterns

Illia Horenko*

Institute of Mathematics, Free University of Berlin, Berlin, Germany

*Corresponding author address: Illia Horenko, Institute of Mathematics, Free University of Berlin, Arnimallee 2-6, 14195 Berlin, Germany.

E-mail:horenko@math.fu-berlin.de

ABSTRACT

Identification and analysis of temporal trends and low-frequency variability in discrete time series is an important practical topic in understanding and prediction of many atmospheric processes, for example, in analysis of climate change. Widely used numerical techniques of trend identification (like local Gaussian kernel smoothing) impose some strong mathematical assumptions on the analyzed data and are not robust to model sensitivity. The latter becomes crucial when analyzing historical observation data with a short record. Two global robust numerical methods for the trend estimation in discrete non-stationary Markovian data based on different sets of implicit mathematical assumptions are introduced and compared here. The methods are first compared on a simple model example, the importance of mathematical assumptions on the data is explained and numerical problems of local Gaussian kernel smoothing are demonstrated. Presented methods are applied to analysis of the historical sequence of atmospheric circulation patterns over UK between 1946-2007. It is demonstrated that the influence of the seasonal pattern variability on transition processes is dominated by the long-term effects revealed by the introduced methods. Despite of the differences in the mathematical assumptions implied by both presented methods, almost identical symmetrical changes of the cyclonic and anticyclonic pattern probabilities are identified in the analyzed data, with the confidence intervals being smaller then in the case of the local Gaussian kernel smoothing algorithm. Analysis results are investigated with respect to model sensitivity and compared to standard analysis technique based on a local Gaussian kernel smoothing. Fi-

nally, the implications of the discussed strategies on long-range predictability of the data-fitted Markovian models are discussed.

Introduction

Many real life processes can be described as simplified discrete models switching between a finite number of states or regimes. Such processes can be found in biophysics (transitions between different conformations in biomolecules) (Schütte and Huisinga 2003), in computational finance (transitions between different market phases) (Hamilton 1989) and in weather/climate research (transitions between different atmospheric circulation regimes) (Majda et al. 2006; Horenko 2008c; Horenko et al. 2008b,a).

However, most of the available time series data from such process share the two following properties: (i) the data has only a short record (since observations are usually available only on some relatively short time intervals), (ii) the underlying dynamics usually has a temporal trend and can not be assumed stationary (i. e., in the Markovian case, the transition matrix can not a priori be assumed time-independent). Moreover, two important practical problems frequently arise when analyzing non-stationary data: (i) comparison of different trend models and trend hypothesis tests, (ii) influence of the mathematical assumptions associated with the choice of the model on the analysis results (robustness).

One of the most frequently used techniques for analysis of non-stationary observation time series is the local Gaussian kernel smoothing approach (Loader 1996, 1999; Zhao and Wei 2003). The main idea of this approach is based on the locally stationary approximation of estimated quantities (like observable averages) or some estimated model parameters inside of the Gaussian window of a certain width, typically characterized by the respective variance parameter σ^2 (for geophysical applications of the method see, for example, (Anderson 2002; Xiong et al. 2006; Dakos et al. 2008)). The window width parameter σ^2 is typically chosen a

priori to characterize some slow intrinsic time scale of interest. However, the main difficulty of this approach lies in its locality, e. g., the information used to estimate the quantities of interest at certain time point is acquired only locally from the inner part of the Gaussian window around this point.

This paper considers a problem of trend estimation for the dynamical processes with discrete finite state space based on the given observation data. In context of atmospheric applications, such processes can describe the transitions between certain predefined circulation regimes or patterns (like the Lamb circulation index (Lamb 1972) or blocked and unblocked situations in atmosphere (Majda et al. 2006)). The considered processes are assumed to fulfill the Markov-property, e. g., the probability of transition to any other available state at any time is only dependent on the current state and is independent from the previous transition history.

Presented paper introduces two global approaches to trend identification in discrete Markov time series and compares them to the local Gaussian kernel smoothing technique adapted here for Markov chain series estimation. The one of the presented methods is a parametric approach, allowing to test the simple trend models of different types and compare them with each other in Bayesian sense. The second technique is of the non-parametrical form, it allows to estimate the low-frequent trends by solving a metastable clustering problem with a fixed number of cluster states. It is explained how the optimal number of clusters can be identified dependent on the single external scalar regularization parameter ϵ^2. The connection between regularization parameter, the persistence of the resulting decomposition and the Gaussian window width parameter σ^2 is discussed. It is shown that in context of Markovian processes, the presented non-parametric method can be viewed as an adaptive

extension of the local Gaussian kernel smoothing technique. Numerical examples demonstrate that in contrast to the non-parametric local Gaussian kernel smoothing technique (Loader 1996, 1999; Zhao and Wei 2003), both of the presented methods allow for a more robust estimation and comparison of different non-stationary Markov models.

It is shown that the introduced methods can be helpful in analysis of practical applications, e. g., the presented algorithms are applied to analysis of the historical sequence of atmospheric circulation patterns over UK between 1946-2007. The long term effects in transitions between different patterns are systematically investigated and compared.

The paper begins with a short description of the non-parametric local Gaussian kernel smoothing in context of Markov chains, followed by the description of both global methods. Special emphasis is done on the intrinsic mathematical assumptions in each of the methods. The final sections deals with application of the presented techniques to the analysis of illustrative model data and atmospheric data series, interpretation of the obtained results and comparison of different methods with respect to their robustness.

Methods and Approach

In the following we will consider the data series that is discrete in time and space, i. e., $\{X_t\}_{t=1,\ldots,T}$ takes values from some fixed set of m distinct quantities $s_1, \ldots, s_m$. These quantities, for example, can denote the states or configurations of the observed system along the dynamics. The process underlying the observations is called Markovian if the probability P of any current state of the process at time t depends only upon the previous state at time $t-1$ and does not depend on any other previous state. Mathematically this property can

be expressed as $\mathrm{P}\left[X_t = s_j | X_1, X_2, \ldots, X_{t-1} = s_i\right] = \mathrm{P}\left[X_t = s_j | X_{t-1} = s_i\right] = P_{ij}(t)$, where $P_{ij}(t)$ denotes the probability of transition from state s_i to state s_j in one step at time t. These probabilities can be put together into an $m \times m$ stochastic transition matrix $P(t)$, i. e., $\sum_{j=1}^{m} P_{ij}(t) = 1$ for any t and i. In order to be able to estimate the a priori unknown transition probabilities $P_{ij}(t)$ based on the observation data X_t, we first need to introduce the concept of the observation probability or likelihood. For example, consider a simple discrete Markov process with three possible states $s_1 = 1, s_2 = 2$ and $s_3 = 3$ available in the form of the following observed time series:

$$\{X_t\}_{t=1,\ldots,12} = \{1, 1, 2, 1, 3, 2, 3, 1, 3, 3, 2, 2\}. \tag{1}$$

Applying the Markov property, it is easy to demonstrate that the probability of observing this time series can be expressed as a probability of starting in one, then staying in one at $t = 2$, then jumping from state one to state two at $t = 3$, etc. That is:

$$\mathrm{P}\left[X_1 = 1, X_2 = 1, \ldots, X_{12} = 2\right] = \mathrm{P}\left[X_1 = 1\right] \mathrm{P}\left[X_2 = 1 | X_1 = 1\right] \ldots \mathrm{P}\left[X_{12} = 2 | X_{11} = 2\right]. \tag{2}$$

For any given Markovian series $\{X_1, \ldots, X_T\}$, the corresponding observation probability (or likelihood) can be compactly written as $\mathrm{P}\left[X_1, \ldots, X_T\right] = \mathrm{P}\left[X_1\right] \prod_{i,j=1}^{m} \prod_{t \in \{t_{ij}\}} P_{ij}(t)$, where $\{t_{ij}\}$ is the set of all time instances when the transitions between s_i and s_j are observed[1]. This means that if the transition matrix is unknown, it can be found by maximization of the above likelihood function for a fixed observation sequence $\{X_1, \ldots, X_T\}$,e. g., solving

[1]It is assumed that $\{t_{ij}\}$ is not empty $\forall i, j$. If it is not true for a certain pair (i, j), it means that no direct transitions between the states s_i and s_j were observed in the time series and the respective matrix elements $P_{ij}(t)$ can be assumed to be equal zero for all t.

the following maximization problem:

$$P_0(t) \;=\; \arg\max_{P(t)} \mathrm{P}\left[X_1, \ldots, X_T\right]. \tag{3}$$

From the numerical point of view, instead of solving the above maximization problem, it is much more convenient to maximize the logarithm of the above expression, i. e., the log-likelihood of the Markovian process:

$$\begin{aligned} \mathbf{L}\left(P(t)\right) &= \log \mathrm{P}\left[X_1, \ldots, X_T\right] \\ &= \log \mathrm{P}\left[X_1\right] + \sum_{i}^{m} \mathbf{L}_i\left(P_{i1}(t), \ldots, P_{im}(t)\right) \\ &\rightarrow \max_{P(t)}, \end{aligned} \tag{4}$$

where

$$\mathbf{L}_i\left(P_{i1}(t), \ldots, P_{im}(t)\right) = \sum_{j=1}^{m} \sum_{t \in \{t_{ij}\}} \log P_{ij}(t) \tag{5}$$

is the partial log-likelihood of the state i. To preserve the stochasticity of the resulting matrix $P(t)$, the minimization problem (4) is subjected to the following constraints

$$\begin{aligned} \sum_{j=1}^{m} P_{ij}(t) &= 1, \quad \text{for all } t, i \\ P_{ij}(t) &\geq 0, \quad \text{for all } t, i, j \end{aligned} \tag{6}$$

The main difficulty arising from the formulation (4-6) is that it is not a well-posed problem. If the transition matrix is allowed to be completely time-dependent, than the number of unknowns to be estimated in this case is equal m^2T, whereas the number of observed transitions and equality constraints is $m(T+1)-1$. Optimizer of the problem (4-6) at time t in such a case will be $P_{X_t X_{t+1}}(t) = 1$ and $P_{X_t k}(t) = 0$ for all other elements $k = 1, \ldots, m$ from the same row X_t of the transition matrix $P(t)$. All other elements of the matrix can be set

to any arbitrarily values (satisfying the constraints), resulting in a meaningless estimation. Since the problem (4-6) is ill-posed, one needs to incorporate some additional information into the formulation of the problem (or, in mathematical language, to regularize the problem) in order to make it well-posed. The simplest form of regularization is an additional assumption about the global stationary of the Markov proces, e. g., the transition matrix $P(t)$ is assumed being time independent[2]

One of the straightforward possibilities to relax the aforementioned global stationarity assumption is based on the application of local Gaussian kernel smoothing idea (Loader 1996, 1999; Zhao and Wei 2003) in context of optimization problem (4-6). Assuming a local stationarity of the underlying transition process at time t_0 inside of the window of a certain width σ^2, one can introduce a normalized Gaussian weight function $\gamma(t, t_0) = \frac{1}{c} \exp(-\frac{(t-t_0)^2}{\sigma^2})$ (where c is the normalization constant). Approximating $\mathbf{L}\left(P(t)\right)$ as

$$\begin{aligned} \mathbf{L}\left(P(t_0)\right) &\approx \log \mathrm{P}\left[X_1\right] + \sum_{i,j=1}^{m} \sum_{t \in \{t_{ij}\}} \gamma(t, t_0) \log P_{ij}(t_0) \\ &\rightarrow \max_{P(t)}, \end{aligned} \tag{7}$$

and applying the method of Lagrange multipliers we get

$$P_{ij}(t_0) = \frac{\sum_{t \in \{t_{ij}\}} \gamma(t, t_0)}{\sum_{t \in \{t_i\}} \gamma(t, t_0)}, \tag{8}$$

where $\{t_i\}$ is the set of all time instances when the state s_i was visited.

[2]If $P(t)$ is assumed to be time-independent, optimization problem (4-6) can be solved analytically applying the method of Lagrange multipliers with $P_{ij} = |\{t_{ij}\}| / \sum_k |\{t_{ik}\}|$, where $|\{t_{ij}\}|$ denotes the number of observed transitions between the states s_i and s_j and $\sum_k |\{t_{ik}\}| \neq 0$.

Trend models of discrete Markov chains

Two main disadvantages of the local Gaussian kernel smoothing procedure described above are: (i) assumption about the local stationarity of the transition process and (ii) arbitrariness in the choice of the window width parameter σ. Besides that, this procedure doesn't give a direct possibility to acquire a constructive functional form of the trend in the analyzed data. This means that in order to make predictions based on available time series information, one has to extrapolate the identified transition process $P(t)$ in future. To do that, the identified process $P(t)$ has to be approximated with a time-dependent function of a certain class $\phi(t)$. Instead of that, one can impose the functional form $\phi(t)$ a priori and incorporate it in to the maximization of the log-likelihood in the form of the trend model.

Single trend models have a general form

$$P(t) \;=\; P^{(0)} + P^{(1)}\phi\,(t)\,, \quad \phi : [1,T] \to (-\infty, +\infty)\,, \tag{9}$$

where $\phi = \phi\,(t)$ is some predefined bounded trend function. Inserting (9) to (4) and (6), after some obvious transformations for any $i = 1, \ldots, m$ we get:

$$\sum_{j=1}^{m} \sum_{t \in \{t_{ij}\}} \log\left(P_{ij}^{(0)} + P_{ij}^{(1)}\phi\,(t)\right) \;\to\; \max_{P^{(0)},P^{(1)}}, \tag{10}$$

$$\sum_{j=1}^{m} P_{ij}^{(0)} \;=\; 1, \tag{11}$$

$$\sum_{j=1}^{m} P_{ij}^{(1)} \;=\; 0, \tag{12}$$

$$P_{ij}^{(0)} + P_{ij}^{(1)} \sup_{t\in[1,T]} \phi(t) \;\geq\; 0, \quad \text{for all } j, \tag{13}$$

$$P_{ij}^{(0)} + P_{ij}^{(1)} \inf_{t\in[1,T]} \phi(t) \;\geq\; 0, \quad \text{for all } j. \tag{14}$$

Problem (10) for a fixed observation sequence $\{X_1, \ldots, X_T\}$ and some given trend function ϕ is a $2m$-dimensional concave maximization problem with respect to the elements of the ith row of matrices $P^{(0)}$ and $P^{(1)}$. The maximization is performed on a convex domain defined by the linear equality and inequality constraints (11-14). Therefore, the problem (10-14) has a solution that can be found numerically applying any available nonlinear maximization algorithm with linear constraints, e. g., a Nelder-Mead optimization algorithm (Nelder and Mead 1964). Note that for the *multiple trend model* $P(t) = P^{(0)} + \sum_i P^{(i)}\phi_i(t)$, analogs of the inequality constraints (13-14) will become non-linear. This will make the numerical procedure much more costly, moreover, the convexity of the domain defined by the constraints will not be guaranteed. Therefore, due to these numerical reasons, in the following we will stick to the *single trend model* defined by (9). We will do that even despite of the fact that the *multiple trend models* have more skill in describing the non-trivial scenarios.

Trend identification with hidden states: adaptive FEM-Clustering

Instead of looking for the trend in a certain class of parametric functions (like, e. g., in log-likelihood maximization with a single trend model, see equation (9)) or assuming the local stationarity of the transition process as in the case of the non-parametric Gaussian kernel smoothing, one can assume that the element-wise logarithm of Markovian transition matrix (4) at any time t can be represented as a convex linear combination of $\mathbf{K}$ time-independent logarithms $\log P^i$ with some unknown time-dependent coefficients $\gamma_i(t)$ (where $\mathbf{K}$ is some predefined number). In other words, in the space of the observed transitions x_t, one can look for $\mathbf{K}$ clusters (or hidden states, since they are a priori unknown) characterized

by $\mathbf{K}$ distinct sets of a priori unknown constant transition probability matrices $P^i \in \mathbf{R}^{m\times m}$ with the cluster distance functional of the form

$$g(x_t, P^i) \;=\; -\log P^i_{X_t X_{t+1}}, \quad t = 1, \ldots, T-1. \tag{15}$$

This cluster distance functional describes the quality of explaining the observed transition $x_t : X_t \to X_{t+1}$ at time t with the help of the transition matrix P^i. For a given cluster distance functional (15), under data clustering we will understand the problem

$$\mathbf{L} \;=\; \sum_{i=1}^{\mathbf{K}} \sum_{1}^{T-1} \gamma_i(t) g(x_t, P^i) \to \min_{\gamma_i(t), P^i}, \tag{16}$$

subject to the constraints on $\gamma_i(t)$:

$$\sum_{i=1}^{\mathbf{K}} \gamma_i(t) \;=\; 1, \quad \forall t \in [1, T-1]\,, \tag{17}$$

$$\gamma_i(t) \;\geq\; 0, \quad \forall t \in [1, T-1]\,, \quad i = 1, \ldots, \mathbf{K}, \tag{18}$$

and on $P^1, \ldots, P^{\mathbf{K}}$ (6). As was demonstrated in (Horenko 2008a,b), in order to get persistent (or metastable clusters) from this minimization procedure, it is enough to impose some metastability assumptions in the space of functions $\gamma_i(\cdot)$) and then apply a finite Galerkin projection of this finite-dimensional Hilbert space. For example, one can restrain the weak discrete differentiability of functions γ_i, i. e.:

$$\begin{aligned} |\gamma_i|_{h^1(1,T-1)} \;&=\; \| \frac{\gamma_i(t+1) - \gamma_i(t)}{\Delta t} (\cdot) \|_{l_2(1,T-1)} = \sum_{t=1}^{T-1} \frac{\left(\gamma_i\left(t+1\right) - \gamma_i\left(t\right)\right)^2}{\Delta t} \\ &\leq\; C^i_\epsilon < +\infty, \quad i = 1, \ldots, \mathbf{K}. \end{aligned} \tag{19}$$

For a given observation time series, the above constraint limits a total number of transitions between the clusters and is connected to the metastability of the hidden process $\gamma_i(t), i = 1, \ldots, \mathbf{K}$ (Horenko 2008a).

One of the possibilities to incorporate the constraint (19) into the minimization of (16) is to introduce the Lagrange-multiplier ϵ^2

$$\begin{aligned} \mathbf{L}^{\epsilon} &= \mathbf{L} + \epsilon^2 \sum_{i=1}^{\mathbf{K}} \sum_{t=1}^{T-1} \frac{\left(\gamma_i\left(t+1\right) - \gamma_i\left(t\right)\right)^2}{\Delta t} \\ &\rightarrow \min_{\gamma_i(t), P^i} . \end{aligned} \tag{20}$$

This form of penalized regularization was first introduced by A. Tikhonov for solution of ill-posed linear least-squares problems (Tikhonov 1943) and was frequently used for non-linear regression analysis in context of statistics (Hoerl 1962) and multivariate spline interpolation (Wahba 1990). In contrast to the aforementioned application of Tikhonov-type regularization to interpolation problems (where the regularization is controlling the smoothness of some non-linear functional approximation of the given data), presented form of the regularized averaged clustering functional (20) has a completely different mathematical structure due to the form of the functional (16). This specific formulation of the optimization problem with constrains allows one to control the metastability of the assignment $\Gamma(t)$ of the analyzed data to $\mathbf{K}$ distinct a priori unknown clusters (Horenko 2008a).

Let $\{1 = t_1, t_2, \ldots, t_{N-1}, t_N = T-1\}$ be a finite subdivision of the time interval $[1, T-1]$ with uniform time step δ_t. We can define a set of continuous functions $\{v_1(t), v_2(t), \ldots, v_N(t)\}$ called hat functions or linear finite elements with compact support (i. e., such that each function $v_i(t)$ is taking positive values in the time interval (t_{i-1}, t_{i+1}) and is zero outside) (Braess 2007). Assuming that $\gamma_i \in h^1\left(1, T-1\right)$ (i. e., functions with the first discrete derivative being square integrable functions in discrete sense, cf. (Braess 2007)) we can

write

$$\begin{aligned} \gamma_i &= \tilde{\gamma}_i + \chi_N \\ &= \sum_{k=1}^{N} \tilde{\gamma}_{ik} v_k + \chi_N, \end{aligned} \tag{21}$$

where $\tilde{\gamma}_{ik} = \sum_{t=1}^{T-1} \gamma_i(t)\, v_k(t)$ and χ_N is a discretization error (it becomes zero if $\delta_t = \Delta t$). Optimal N can be chosen adaptively to guarantee that χ_N doesn't exceed a certain discretization error threshold. Inserting (21) into functional (20) and constraints (17,18) we get

$$\tilde{\mathbf{L}}^{\epsilon} = \sum_{i=1}^{\mathbf{K}} \left[a(P^i)^{\mathbf{T}} \tilde{\gamma}_i + \epsilon^2 \tilde{\gamma}_i^{\mathbf{T}} \mathbf{H} \tilde{\gamma}_i \right] \to \min_{\tilde{\gamma}_i, P^i}, \tag{22}$$

$$\sum_{i=1}^{\mathbf{K}} \tilde{\gamma}_{ik} = 1, \quad \forall k = 1, \ldots, N, \tag{23}$$

$$\tilde{\gamma}_{ik} \geq 0, \quad \forall k = 1, \ldots, N, \quad i = 1, \ldots, \mathbf{K}, \tag{24}$$

$$\sum_{j=1}^{m} P_{ij}^{l}(t) = 1, \quad \text{for all } l, i \tag{25}$$

$$P_{ij}^{l}(t) \geq 0, \quad \text{for all } l, i, j \tag{26}$$

where $\tilde{\gamma}_i = (\tilde{\gamma}_{11}, \ldots, \tilde{\gamma}_{iN})$ is the vector of discretized affiliations to cluster i, and

$$a(P^i) = \left(\sum_{t=t_1}^{t_2} v_1(t) g(x_t, P^i) \delta_t \Delta t, \ldots, \sum_{t_{N-2}}^{t_N - 1} v_N(t) g(x_t, P^i) \delta_t \Delta t \right), \tag{27}$$

is a vector of discretized cluster distances and $\mathbf{H}$ is the symmetric tridiagonal stiffness-matrix of the linear finite element set with $2/t\delta_t$ on the main diagonal, $-1/\delta_t$ on both secondary diagonals and zero elsewhere. (22-26) is a non-linear minimization problem with linear equality and inequality constraints, imposed on both $\tilde{\gamma}_i$ and $P^i, i = 1, \ldots, \mathbf{K}$.

If $\epsilon^2 = 0$, then the above minimization problem (22-24), can be solved analytically wrt. $\tilde{\gamma}_i^{(L)}$ for a fixed set of transition matrices $P^1, \ldots, P^{\mathbf{K}}$ (where L denotes the index of current iteration) resulting in

$$\gamma_i^{(L)}(t_j) \quad = \quad \begin{cases} 1 & i = \arg\min a_j\left(P^i\right), \\ \\ 0 & \text{otherwise}, \end{cases} \tag{28}$$

If $\epsilon^2 > 0$, for a fixed set of transition matrices $P^1, \ldots, P^{\mathbf{K}}$, the minimization problem (22-24), reduces to a sparse quadratic optimization problem with linear constraints which can be solved by standard tools of sparse quadratic programming (sQP) with computational cost scaling as $\mathcal{O}\left(N \log\left(N\right)\right)$ (Gill et al. 1987).

The minimization problem (22) with transition matrix constraints (25-26) can be solved analytically wrt. the parameters $P^1, \ldots, P^{\mathbf{K}}$ for a fixed set of discretized cluster affiliations $\tilde{\gamma}_i$. Applying the method of Lagrange multipliers results in the following expression for the transition matrices:

$$P_{ij}^l = \frac{\sum_{t \in \{t_{ij}\}} \gamma_l(t)}{\sum_{t \in \{t_i\}} \gamma_l(t)}, \tag{29}$$

where $\{t_{ij}\} \subset [1, T-1]$ are the time instances when the transitions between the states s_i and s_j are observed and $\{t_i\}$ is a set of all time instances when the state s_i is visited.

Note the obvious similarity between the expression (29) and the local transition matrix estimator (8) in context of Gausian kernel smoothing. Despite of this similarity and the fact that both of the methods rely on some form of stationarity assumption, statistical weights $\gamma(t, t_0)$ are fixed and localized by the choice of the Gaussian curve of the certain width, whereas hidden probabilities $\gamma_i(t)$ are defined adaptively, dependent on the whole set of available data and not just on the part of it from inside the window. In another words,

expression (29) gets use of the global information contained in the cluster affiliation functions $\gamma_i(t)$ (later on in the text it is explained how predefined discretization error threshold and regularization parameter ϵ^2 influence these values). Globality of the adaptive FEM-clustering procedure gives an advantage when analyzing the data with a short time record (since it allows to tighten the confidence intervals around the estimated parameters and can help to obtain more robust numerical results).

Similarly to the algorithms described in (Horenko 2008a,b), minimization of the functional (22) can be implemented as an iterative numerical scheme. The iterations can be repeated until the change of the functional value doesn't exceed some predefined threshold for the change of the functional value (22).

Estimation of optimal $\mathbf{K}$ *dependent on* ϵ^2*:*

The upper bound for the number of statistically distinguishable cluster states for each value of ϵ^2 can be algorithmically estimated in the following way: starting with some a priori chosen (big) $\mathbf{K}$ one solves the optimization problem (22-26) for a fixed value of ϵ^2 and calculates the confidence intervals of the resulting local transition matrices $P^i, i = 1, \ldots, \mathbf{K}$ (this can be done applying the standard sampling procedures, see Noe (2008)). If two of the estimated matrices have the confidence intervals that are overlapping in all components, this means that respective clusters are statistically indistinguishable and the whole procedure must be repeated for $\mathbf{K} = \mathbf{K} - 1$. If at a certain point all of the matrices are statistically distinguishable the procedure is stopped and $\mathbf{K}_{max}(\epsilon^2) = \mathbf{K}$.

Robustness Estimation

It is intuitively clear that the quality of the resulting reduced model is very much dependent on the original data, especially on the length of the available time series. The shorter the observation sequence, the bigger the uncertainty of the resulting parameters. Therefore global methods presented above, like single trend model and FEM-clustering should be more robust then local methods, like Gaussian kernel smoothing, as the analyzed time series become longer. That is, for a short time series, everything is local, but as the time series becomes long (in the sense that $T >> \sigma$, with σ defining the width of a Gaussian moving window) the global methods then have the advantage of assimilating more information.

Let $P^* = P^M_{\{X_t\}}(t)$ be the Markovian transition matrix, estimated with one of the methods M described above from a given observation sequence $\{X_t\}_{t=1,\ldots,T}$. As $\{X^M_t\left(P(t),\omega\right)\}_{t=1,\ldots,T}$ we denote a single realization of the Markov process $P(t)$ with the model M. To compare the robustness of different models for a given data, we introduce the following estimate deviation process $R^M_{\{X_t\}}(f,\omega)$ for a given function f

$$R^M_{\{X_t\}}(f,\omega) \;\; = \;\; f\left(P^*\right) - f\left(P^M_{\{X^M_t(P^*,\omega)\}}(t)\right), \tag{30}$$

The Gaussian assumption for the stochastic process $R^M_{\{X_t\}}(f,\omega)$ gives an opportunity to estimate its confidence intervals for some given function f straightforwardly. This can be done in a standard way by using multivariate statistical analysis, i. e., by the Monte Carlo sampling from the respective distribution and calculating expectation values wrt. ω (Mardia et al. 1979). The confidence intervals calculated for the variable $R^M_{\{X_t\}}(f,\omega)$ give a possibility to estimate the intrinsic robustness of the chosen model M for a give observation $\{X_t\}_{t=1,\ldots,T}$.

Application I: Instructive Model Example

To illustrate an application of the methods described above, it is instructive to investigate the following simplified model scenario: consider a process X_t with three directly observed possible states $\{1, 2, 3\}$ and transition probabilities taking values from one of the two following transition matrices

$$A_1 = \begin{pmatrix} 0.82 & 0.12 & 0.06 \\ 0.08 & 0.87 & 0.05 \\ 0.07 & 0.09 & 0.84 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0.16 & 0.70 & 0.14 \\ 0.52 & 0.37 & 0.11 \\ 0.30 & 0.27 & 0.43 \end{pmatrix}. \tag{31}$$

The choice of the transition matrix, taken to calculate a probability of the next transition at time t, depends on another discrete hidden process Y_t with two possible states $\{1, 2\}$. If the hidden process Y_t takes the value 1, then matrix A_1 is chosen to calculate transition probabilities of the observed process X at time t, otherwise transition matrix A_2 is taken. An example of the observation series X_t resulting from this computation procedure is demonstrated in the right panel of Figure 1. Hidden process Y_t used in this calculation is shown in the left panel of Figure 1. The log-likelihood of the generated time series in terms of the switching Markov model (31) is $\mathbf{L}_{real} = -708.67$.

Next, generated Markov time series from Figure 1 is taken to compare three presented methods wrt. the identification of the non-stationary trend of the Markov transition probabilities (driven by the hidden variable Y_t). We start with the local Gaussian kernel smoothing procedure (8), repeat the estimation with different values of the Gaussian window width parameter σ^2 and keep the parameter value with the highest log-likelihood (which in this case turns out to be $\sigma^2_{opt} = 1296$, that corresponds to effective window width of approx. 70 time

units). The log-likelihood of the analyzed time series in terms of the local Gaussian kernel smoothing model with $\sigma^2_{opt} = 1296$ turns out to be $\mathbf{L}_{Gauss} = -717.34$ (this is just 0.58% less then the log-likelihood of the original switching Markov model). Estimation of the single trend model (9) with $\phi(t) = t^\alpha$ as expected results in almost stationary transition matrix estimates. This is simply explained by the fact that models with polynomial trend is a bad choice for approximating non-stationary processes where the change of the transition probability matrices is described by the hidden process like the one from Figure 1. Finally, we apply the adaptive FEM-Clustering procedure with different values of regularization parameter ϵ^2. The tolerance threshold χ_N (21) is chosen to be 0.0001 in all cases.

Figure 2 demonstrates an influence of the regularization parameter ϵ^2 on the identified hidden process (described by the variables $\gamma_i(t)$ resulting from the numerical solution of the optimization problem (22-26)). Whereas for $\epsilon^2 = 0$ (correspondent to the unregularized optimization) we get an arbitrary hidden process, for $\epsilon^2 = 0.2$ the identified hidden process almost exactly reproduces the original process Y_t used in generation of the time series. Hidden processes identified with bigger values of ϵ^2 become more and more smooth until the point where the regularization part of the functional (20) starts to dominate the optimization producing constant hidden process with no switches.

Figure 3 shows a comparison of the original transition probability trends as functions of time with the results obtained by the local Gaussian kernel smoothing model with $\sigma^2_{opt} = 1296$ (dotted, the robustness intervals are in light grey) and the trends obtained by the adaptive FEM-Clustering procedure with $\epsilon^2 = 0.2$ and $\mathbf{K} = 2$ (dashed). The log-likelihood of the analyzed time series in terms of the adaptive FEM-Clustering procedure turns out to be $\mathbf{L}_{FEM} = -703.21$ (this is 0.67% bigger then the log-likelihood of the original switching

Markov model (31)). As it can be seen from Figure 3, trends resulting from the adaptive FEM-Clustering procedure reproduce the original trends much better then the local Gaussian kernel smoothing trends do. This is not a big surprise since the internal structure of the analyzed time series is much better reproducible with the help of the adaptive FEM-Clustering procedure. Figure 3 also demonstrates the effect of locality of the Gaussian smoothing procedure: since the optimal Gaussian window turns up to become relatively narrow, only a relatively small amount of information is getting incorporated in the trend estimation at any point. This result in huge confidence intervals for the estimated matrix elements. Comparison of log-likelihood values $\mathbf{L}_{real} = -708.67, \mathbf{L}_{Gauss} = -717.34, \mathbf{L}_{FEM} = -703.21$ and inspection of the respective probability trends in the Figure 3 reveals, that very small relative changes in the value of the log-likelihood can be induced by very significant changes of trend. This means that the relative value of the log-likelihood function alone can not be considered as a measure for the model quality.

Finally, the procedure of $\mathbf{K}_{max}(\epsilon^2)$ identification described above is exemplified for this model data example. Figure 4 demonstrates that the maximal number of statistically distinguishable clusters decreases with increasing value of ϵ^2. Moreover, described procedure allows a correct identification of hidden states Y in a wide range of regularization parameters ($0.2 \leq \epsilon^2 \leq 0.8$). Note that due to the form of the regularized functional (20), the absolute value of the regularization parameter ϵ^2 necessary to achieve a desired persistence of the hidden process $\gamma_i(t), i = 1, \ldots, \mathbf{K}$ will be different for different applications. This value is influenced by the norm and the properties of the original cluster distance functional (16) evaluated for a given time series, i. e., the magnitude of ϵ^2 and numerical investigations analogous to the one presented in Figure 4 should be performed in each data-analysis case.

Application II: Analysis of Lamb Circulation Index Series for UK

a. Data Sources

Analyzed data (electronically retrievable at *http://www.cru.uea.ac.uk/ftpdata/lwtjenk.dat*, provided by the *Climatic Research Unit of the University of East Anglia*) represents a sequence of atmospheric circulation patterns over UK. In the original time series, each of the days between 1.Jan.1880 and 31.Jul.2007 is assigned to one of the 28 basic patterns (cyclonic, anticyclonic, western, etc.) based on the daily grid-point mean sea level pressure data. The assignment of the circulation regimes in the considered time series is done according to the modified Lamb classification scheme proposed by Jenkinson and Collison (Lamb 1972; Jenkinson and Collison 1977; Jones et al. 1993). In the following, the analysis is restricted to the part of the series after the second world war, i. e., between the 1.Jan.1946 and 31.Jul.2007. To simplify the analysis and to make the graphical representation of results more comprehensible, a total number of the analyzed circulation patterns is reduced to three, i. e., the anticyclonic pattern is denoted as state 1, the cyclonic pattern is denoted as state 3 and all other patterns are assigned to the collective state 2.

b. Trend Discrimination with Maximum Likelihood Approach

Before demonstrating the applicability of the strategies described in this paper to the resulting time series (switching between the discrete states 1, 2 and 3), it is necessary first

to demonstrate the Markovian property of the respective transition process. For a pity, there is no standard strategy to test the Markov-assumption in non–stationary case. In the following we will use the test strategy applicable to stationary case and will assume that the deviations from stationarity are not very significant for the analyzed data so the stationary tests are still valid in some sence. Development of non-stationary Markov tests is a matter of future research. For the stationary case (where the underlying transition matrix is assumed to be time independent), the Markovian property can be verified applying some standard tests, e. g., one can check the generator of the process, see (Crommelin and Vanden-Eijnden 2006; Metzner et al. 2007). Application of this test confirms the Markov hypothesis for the analyzed data in stationary case and reveals that the underlying process can be represented by a rapidly mixing Markov chain (i. e., the probabilities to stay in the same state are comparable to the probabilities to change the state so the respective process does not get stuck in the same state for a long time).

As was mentioned above, identification of multiple trend models, due to the non-convexness of the respective inequality constraints, is a problem that is unfeasible from the numerical point of view. To choose the most probable single trend model for a given time series of atmospheric circulation patterns, different trend function classes $\phi(t)$ can be parameterized by numerical solution of the problem (10-14) and then compared with the help of the standard Bayesian hypothesis test approach (Gelman et al. 2004).

One of the most intuitive single trend models in meteorological context is the seasonal trend model of the form

$$P(t) = P^{(0)} + P^{(1)} \sin(\frac{2\pi}{T}t + \phi), \tag{32}$$

where $T = 365.24$ days is a seasonal period and $\phi \in [0, T]$ is a phase factor. The maximization problem (10-14) is independently solved for each of the rows of the transition matrix. Optimization is repeated for various values of ϕ and for each matrix row the parameters $P^{(0)}, P^{(1)}, \phi$ with the highest value of the partial log-likelihood are kept (see the dashed lines in Figure 4). The same kind of procedure is performed for the single trend model of the polynomial form

$$P(t) = P^{(0)} + P^{(1)} t^{\alpha}. \tag{33}$$

Statistical hypothesis tests can be performed to decide, which of the single trend models can better explain the observed data. The log-likelihood of the respective partial log-likelihood maxima (see Figure 5) can be calculated and the a posteriori model probabilities can then be acquired from the Bayes formula. It shows up that the polynomial trend model $P(t) = P^{(0)} + P^{(1)} t^{\alpha}$ and the non-stationary hidden states model (estimated with adaptive FEM-clustering algorithm for 3 hidden states) have the highest probability to describe the analyzed date. Moreover, different values of exponent α in the polynomial trend model are optimal for different atmospheric circulation patterns ($\alpha = 0.8$ for anticyclonic, $\alpha = 0.3$ for cyclonic and $\alpha = 0.5$ for the combination of all other patterns). Inspection of the resulting trend derivatives $\dot{P}(t) = \alpha P^{(1)} t^{\alpha - 1}$, shows that according to the analyzed data, the speed of climate change visible in transition probability trends was higher at the beginning of the analyzed period as compared to the current period of time (since for all of the identified trends ($\alpha < 1$)). This finding may represent a local effect and must be verified on the global data. This is a matter of further research.

Figure 5 demonstrates that the maximal log-likelihood values for optimal polynomial

trend and hidden states models are higher than the optimal values for the seasonal trend model (dashed lines) and the local Gaussian kernel smoothing result (dash-dotted, with effective Gaussian window width of 10 years, corresponds to the Gaussian window width used in the original work (Jones et al. 1993) considering the analysis of the UK Lamb index series). This finding means that the influence of seasonal pattern variability on transition processes is dominated by the long-term effects induced by the single polynomial trend and hidden states models. Moreover, the optimal hidden states model with $\mathbf{K} = 3$ has a highest probability in the case of states 1 and 2. In the case of the state 3 (describing the cyclonic pattern), the single polynomial trend model $P_{ij}(t) = P_{ij}^{(0)} + P_{ij}^{(1)} t^{0.3}$ can explain the observed data better then any other tested model. However, this finding should be handled with care since the number of parameters used in the hidden state model for $\mathbf{K} = 3$ is higher then the number of parameters involved in the single polynomial trend model.

Comparison of Robustness and Predictability

In order to interpret the obtained results, estimated transition matrices $P(t)$ and instantaneous statistical weights $\pi_i(t), i = 1, 2, 3$

$$\pi(t) P(t) = \pi(t), \quad \pi(t) = (\pi_1(t), \ldots, \pi_m(t)) \tag{34}$$

can be compared, both wrt. their qualitative temporal behavior and robustness. Instantaneous statistical weights are the components of the steady state PDF of the stationary Markov process, e. g., in non-stationary cases these quantities describe a time evolution of an equilibrium PDF. Figure 6 demonstrates a good agreement of the resulting parameters for the single polynomial trend model and the hidden states model. It can clearly be seen that

in both cases the probability to stay in the anticyclonic pattern decreases, the probability to go from the anticyclonic to the cyclonic pattern increases and the probability to go from anticyclonic to any other circulation pattern stays almost constant. For the cyclonic pattern transitions the situation is reversed: the probability to stay in the cyclonic pattern increases, the probability to go from the cyclonic to the anticyclonic pattern decreases and the probability to go from cyclonic to any other circulation pattern stays almost constant. As can be seen from Figure 7, this tendency results in increasing instantaneous statistical weight of the cyclonic pattern and in symmetric anticyclonic pattern weight decreasing. Moreover, Figures 2 and 3 demonstrate the higher robustness of results obtained by the global methods compared to the local Gaussian kernel smoothing method.

In order to make predictions based on available time series information, one has to extrapolate the identified transition process $P(t)$ to the future. As was mentioned above, single trend models give a direct possibility to generate the long-term predictions since the exact functional form (in some predefined class of functions, e. g., polynomials) of the transition process $P(t)$ is estimated explicitly from the observation data. For any given time $1 < \tau < T$, $P^{[1,\tau]}(t)$ will denote the Markovian transition matrix estimated only from part of the available time series $X_1, X_2, \ldots, X_\tau$. In order to quantify the prediction quality based on this estimate, mean log-likelihood of prediction $\bar{\mathbf{L}}_\tau^{\Delta t}$ can be used:

$$\bar{\mathbf{L}}_\tau^{\Delta t} = \frac{1}{\Delta t} \sum_{i=1,j=1}^{m} \sum_{t \in \{t_{ij}^{[\tau,\tau+\Delta t]}\}} \log P_{ij}^{[1,\tau]}(t) \tag{35}$$

where $\{t_{ij}^{[\tau,\tau+\Delta t]}\} \subset [\tau+1, \tau+\Delta t]$ are the time instances between $\tau+1$ and $\tau+\Delta t$, when the transitions between s_i and s_j are observed. Figure 8 illustrates that predictions based on single trend polynomial model are more probable then the predictions based on the

stationary Markovian model without trend. This means that the long term effects explained by the polynomial trend are significant and can be reliably identified even using a part of the original data.

Concluding Discussion

Three numerical methods for analysis of non-stationary Markovian data have been presented and compared here. In contrast to the standard approach (being a local non-parametric technique, i.e., getting use only of the local information inside of the moving Gaussian window), two presented methods acquire the information globally, therefore, under predefined mathematical assumptions, they allow a more reliable estimate of the underlying model parameters. This feature is very important for analysis of a short record series. Both presented global methods demonstrated a more reliable trend discrimination with more narrow robustness intervals compared with the results of the local Gaussian kernel smoothing. It was exemplified how the new methods can help to perform a robust identification of transition probabilities and stationary weights in the analyzed circulation data. Both methods, despite the difference in the mathematical assumptions implied on the data (the one being parametric, the second being non-parametric), revealed the same robust weight increase of the cyclonic circulation pattern (absolute increase of $(6.3 \pm 0.5)\%$) and symmetrical decrease of the anticyclonic pattern weight (absolute decrease of $(5.5 \pm 0.5)\%$) over UK between 1945-2007. Moreover, the results of the single trend model analysis indicated that the speed of climate change identified from transition probability trends was higher at the beginning of the analyzed period as compared to the current period of time.

One of the most important practical issues in the area of time series analysis is construction of dynamical models able to predict the future. The single trend models have obvious shortcomings predicting the situations where the functional and parametric form of the trend function $\phi(t)$ is changed, e. g., in the case of the regime change. On the other hand, adaptive FEM-clustering method has a potential to cope with this problem. However, in the current setting of the algorithmic procedure, hidden state probabilities $\gamma_i(t)$ are identified without making any assumptions about their dynamics (in contrast to the widely used HMM-strategy where these quantities are taken a priori to be Markov processes). It means that in order to be able to predict the phase transitions in realistic applications, dynamics of hidden state probabilities $\gamma_i(t)$ has to be further investigated a posteriori. Moreover, robust statistical techniques of change point analysis have to be developed. These problems are the matters of further research.

For the analyzed practical application, it was demonstrated that the polynomial parameterization of the non-stationary Markov model enables a better quality of predictions as compared to the stationary case. Further development of the presented data-analysis techniques can help to acquire a better understanding of the low-frequency variability and dynamics of processes switching between metastable weather and climate regimes.

Acknowledgments

The author would like to thank Andrew Majda (Courant Institute, NYU), Eric Vanden-Eijnden (Courant Institute, NYU), Rupert Klein (FU Berlin) and Dirk Becherer (HU Berlin) for intensive discussions and valuable comments.

The author thanks the unknown referees who's valuable comments helped to make the paper more readable and the Climatic Research Unit of the University of East Anglia for the time series data. The work was partially supported by the DFG SPP 1276 MetStroem "Meteorology and Turbulence Mechanics" and DFG Research Center "Matheon".

REFERENCES

Anderson, J., 2002: A local least squares framework for ensemble fitting. *Month. Weath. Rev.*, **6131(4)**.

Braess, D., 2007: *Finite Elements: Theory, Fast Solvers and Applications to Solid Mechanics*. Cambridge University Press.

Crommelin, D. and E. Vanden-Eijnden, 2006: Fitting time series by continuous-time Markov chains: A quadratic programming approach. *J. Comp. Phys.*, **217(2)**.

Dakos, V., M. Scheffer, E. van Nes, V. Brovkin, V. Petukhov, and H. Held, 2008: Slowing down as an early warning signal for abrupt climate change. *Proc. Nat. Acad. Sci*, **105(38)**.

Gelman, A., J. Carlin, H. Stern, and D. Rubin, 2004: *Bayesian Data Analysis*. Chapman and Hall.

Gill, P., W. Murray, M. Saunders, and M. Wright, 1987: A Schur-complement method for sparse quadratic programming. *Technical report, STANFORD UNIV CA SYSTEMS OPTIMIZATION LAB*.

Hamilton, J., 1989: A new approach to the econometric analysis of nonstationary time series and the bysiness cycle. *Econometrica*, **57**, 357–384.

Hoerl, A., 1962: Application of ridge analysis to regression problems. *Chemical Engineering Progress*, **58**, 54–59.

Horenko, I., 2008a: Finite element approach to clustering of multidimensional time series. *submitted to SIAM J. of Sci. Comp., (available via biocomputing.mi.fu-berlin.de).*

Horenko, I., 2008b: On clustering of non-stationary meteorological time series. *submitted to J. of Cli., (available via biocomputing.mi.fu-berlin.de).*

Horenko, I., 2008c: On simultaneous data-based dimension reduction an hidden phase identification. *J. of Atmos. Sci.*, **65(6)**.

Horenko, I., S. Dolaptchiev, A. Eliseev, I. Mokhov, and R. Klein, 2008a: Metastable decomposition of high-dimensional meteorological data with gaps. *J. of Atmos. Sci.*, **65(11)**.

Horenko, I., R. Klein, S. Dolaptchiev, and C. Schuette, 2008b: Automated generation of reduced stochastic weather models i: simultaneous dimension and model reduction for time series analysis. *SIAM MMS*, **6(4)**.

Jenkinson, A. and F. Collison, 1977: An initial climatology of gales over the North Sea. *Synoptic Climatology Branch Memorandum*, **62**.

Jones, P., M. Hulme, and K. Briffa, 1993: A comparison of Lamb circulation types with an objective classification scheme. *International Journal of Climatology*, **13**.

Lamb, H., 1972: British Isles weather types and a register of daily sequence of circulation patterns 1861-1971. *Geophysical Memoir*, **116**.

Loader, C., 1996: Local likelihood density estimation. *The Annals of Statistics*, **24**.

Loader, C., 1999: *Local Regressions and Likelihood*. Springer, New Yorck.

Majda, A., C. Franzke, A. Fischer, and D. Crommelin, 2006: Distinct metastable atmospheric regimes despite nearly gaussian statistics : A paradigm model. *PNAS*, **103 (22)**, 8309–8314.

Mardia, K., J. Kent, and J. Bibby, 1979: *Multivariate Analysis*. Academic Press.

Metzner, P., I. Horenko, and C. Schuette, 2007: Generator estimation of Markov jump processes based on incomplete observations nonequidistant in time. *Physical Rewiev E*, **227(1)**.

Nelder, J. and R. Mead, 1964: A simplex method for function minimization. *The Computer Journal*, **7**.

Noe, F., 2008: Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.*, **128**.

Schütte, C. and W. Huisinga, 2003: Biomolecular conformations can be identified as metastable sets of molecular dynamics. *Handbook of Numerical Analysis*, P. G. Ciaret and J.-L. Lions, Eds., Elsevier, Vol. X, 699–744.

Tikhonov, A., 1943: On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, **39 (5)**, 195–198.

Wahba, G., 1990: *Spline Models for Observational Data*. SIAM.

Xiong, L., S. Guo, and K. O'Connor, 2006: Smoothing the seasonal means of rainfall and runoff in the linear perturbation model (LPM) using the kernel estimator. *J. of Hydrology*, **324**.

Zhao, S. and G. Wei, 2003: Jump process for the trend estimation of time series. *Comp. Stat. and Data Anal.*, **42**.

List of Figures

5 Maximal partial log-likelihoods (5) of anticyclonic (left), cyclonic (right) and all other (middle) circulation patterns. Dotted lines represent the partial log-likelihoods estimated for non-stationary Markovian models $P(t) = P^{(0)} + P^{(1)}t^{\alpha}$ as functions of exponent α. Dashed lines mark the maximal values of partial loglikelihoods for the seasonal Markovian trends of the form $P(t) = P^{(0)} + P^{(1)}\sin(\frac{2\pi}{T}t + \phi)$ (the maxima are calculated over all possible $\phi \in [0, T]$, where $T = 365.4$ days). Also shown are the partial log-likelihoods derived with the help of: adaptive metastable FEM-clustering with $\mathbf{K} = 3, \mathbf{N} = 50, \epsilon = 50000$ (bars), and local kernel smoothing with the Gaussian window width of 10.0 years (dash-dotted lines). The log-likelihood maxima for the polynomial trend are achieved at $\alpha_1 = 0.8$ (left), $\alpha_2 = 0.5$ (middle) and $\alpha_3 = 0.3$(right). 39

6 Comparison of Markovian transition probabilities $P_{ij}\,(t)$ estimated with: log-likelihood maximization with polynomial trend $P_{ij}(t) = P_{ij}^{(0)} + P_{ij}^{(1)}t^{\alpha_i}$ for $\alpha_1 = 0.8, \alpha_2 = 0.5, \alpha_3 = 0.3$ (black solid lines), local Gaussian kernel smoothing for the window width of 10.0 years (gray solid lines) and FEM-clustering for $\mathbf{K} = 3, \mathbf{N} = 50, \epsilon = 50000$ (dashed lines). Black dotted lines mark the robustness intervals for the parameters estimated by the log-likelihood maximization with polynomial trend and gray dotted lines are the robustness intervals of the local Gaussian kernel smoothing. The robustness intervals of the FEM-clustering method have almost the same size as the confidence intervals of the single trend model (since both methods are global approaches.) 40

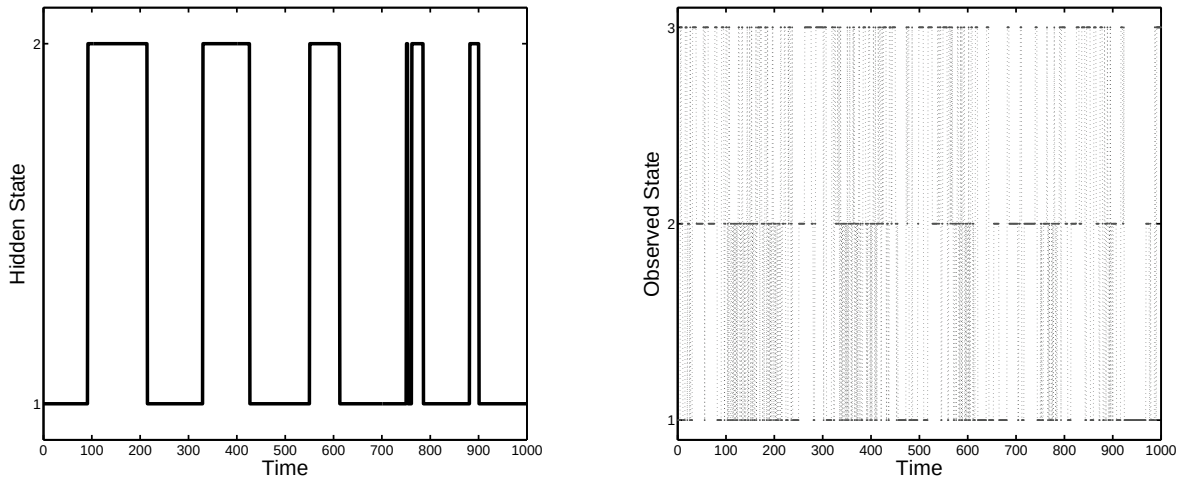

FIG. 1. Hidden process Y_t switching between two different transition matrices (31) (left) and an example of the resulting discrete Markovian series X_t (right).

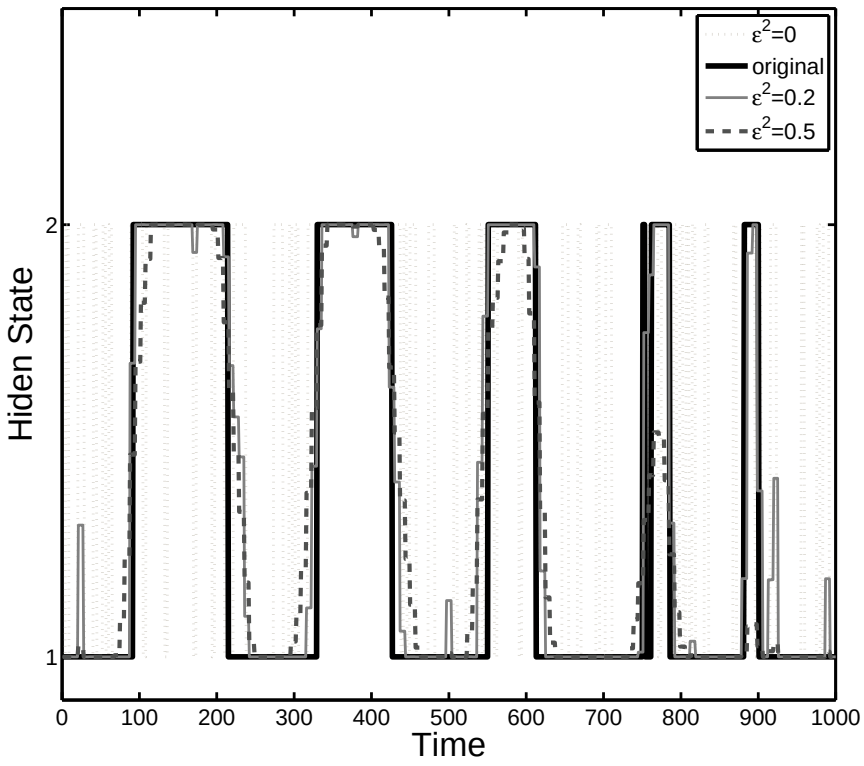

FIG. 2. Comparison of the original hidden process from the left panel of Figure 1 (solid) with the hidden processes identified from the respective observed sequence (see the right panel of Figure 1) by adaptive FEM-Clustering ($\mathbf{K} = 2$, with tolerance threshold $\chi_N = 0.0001$). The minimization is performed with different values of ϵ^2.

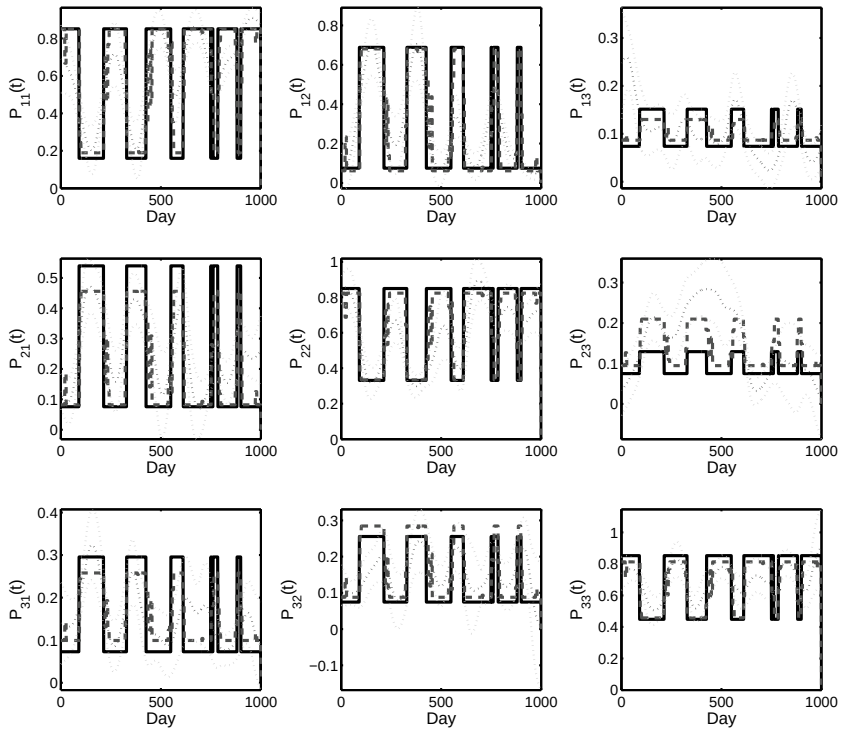

FIG. 3. Comparison of Markov transition probabilities as functions of time: original transition probabilities (solid) used in generation of the time series; probabilities estimated by the local Gaussian kernel smoothing with the value of σ^2 defined by the variation of the log-likelihood maximization procedure (dark grey, dotted) together with its robustness intervals (light grey, dotted); probabilities estimated for $\epsilon^2 = 0.2$ ($\mathbf{K} = 2$, with tolerance threshold $\chi_N = 0.0001$) (dark grey, dashed). The robustness intervals for the FEM-clustering method estimates are of the order of ± 0.06.

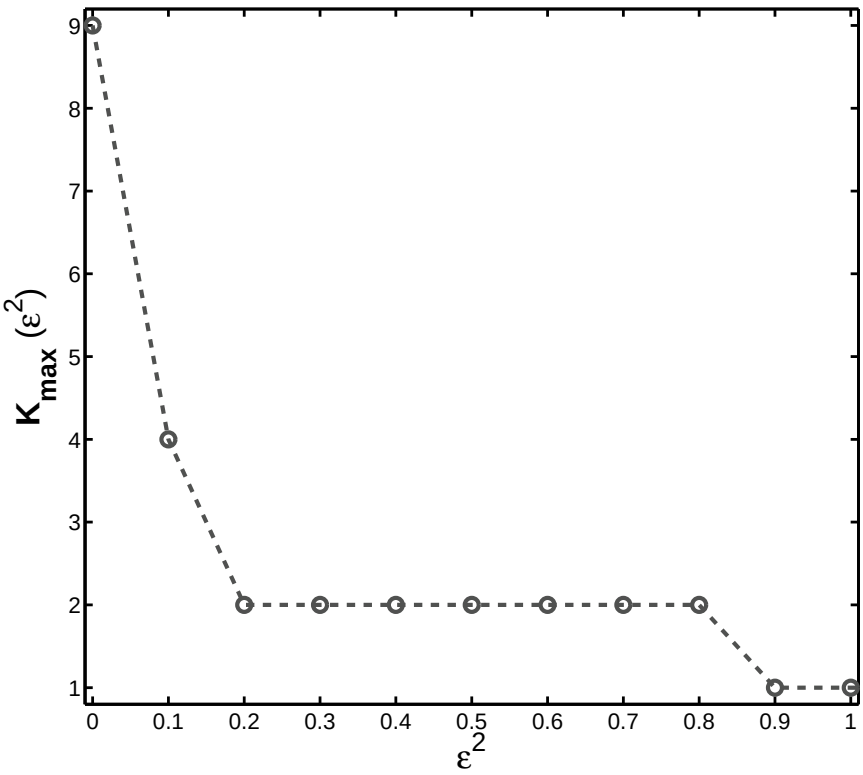

FIG. 4. Maximal number of the statistically distinguishable cluster states $\mathbf{K}_{max}$ as a function of ϵ^2.

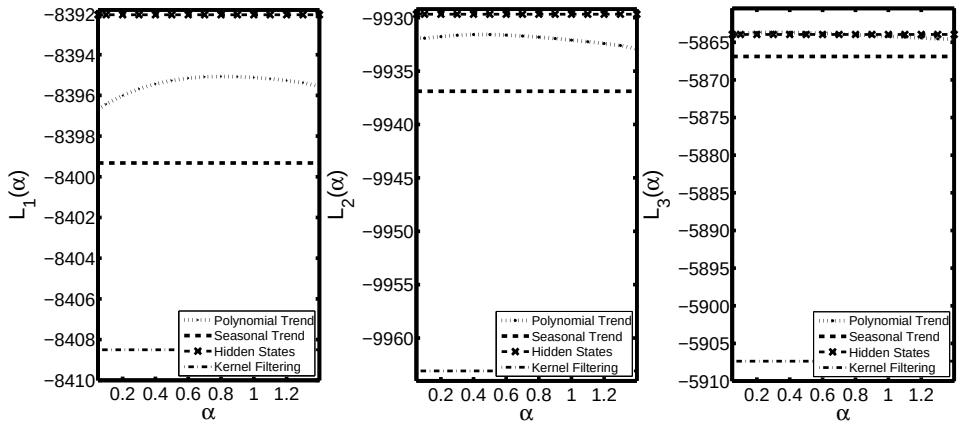

FIG. 5. Maximal partial log-likelihoods (5) of anticyclonic (left), cyclonic (right) and all other (middle) circulation patterns. Dotted lines represent the partial log-likelihoods estimated for non-stationary Markovian models $P(t) = P^{(0)} + P^{(1)}t^{\alpha}$ as functions of exponent α. Dashed lines mark the maximal values of partial loglikelihoods for the seasonal Markovian trends of the form $P(t) = P^{(0)} + P^{(1)}\sin(\frac{2\pi}{T}t + \phi)$ (the maxima are calculated over all possible $\phi \in [0, T]$, where $T = 365.4$ days). Also shown are the partial log-likelihoods derived with the help of: adaptive metastable FEM-clustering with $\mathbf{K} = 3, \mathbf{N} = 50, \epsilon = 50000$ (bars), and local kernel smoothing with the Gaussian window width of 10.0 years (dash-dotted lines). The log-likelihood maxima for the polynomial trend are achieved at $\alpha_1 = 0.8$ (left), $\alpha_2 = 0.5$ (middle) and $\alpha_3 = 0.3$(right).

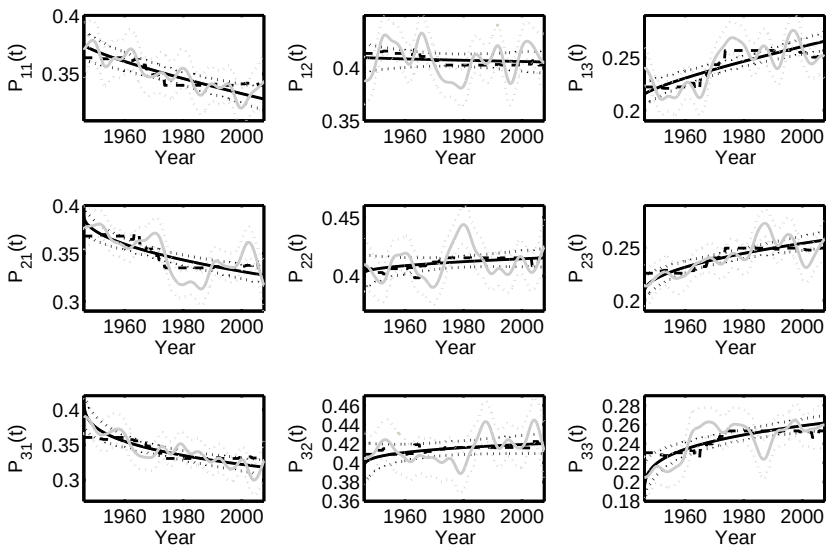

FIG. 6. Comparison of Markovian transition probabilities $P_{ij}(t)$ estimated with: log-likelihood maximization with polynomial trend $P_{ij}(t) = P_{ij}^{(0)} + P_{ij}^{(1)} t^{\alpha_i}$ for $\alpha_1 = 0.8, \alpha_2 = 0.5, \alpha_3 = 0.3$ (black solid lines), local Gaussian kernel smoothing for the window width of 10.0 years (gray solid lines) and FEM-clustering for $\mathbf{K} = 3, \mathbf{N} = 50, \epsilon = 50000$ (dashed lines). Black dotted lines mark the robustness intervals for the parameters estimated by the log-likelihood maximization with polynomial trend and gray dotted lines are the robustness intervals of the local Gaussian kernel smoothing. The robustness intervals of the FEM-clustering method have almost the same size as the confidence intervals of the single trend model (since both methods are global approaches.)

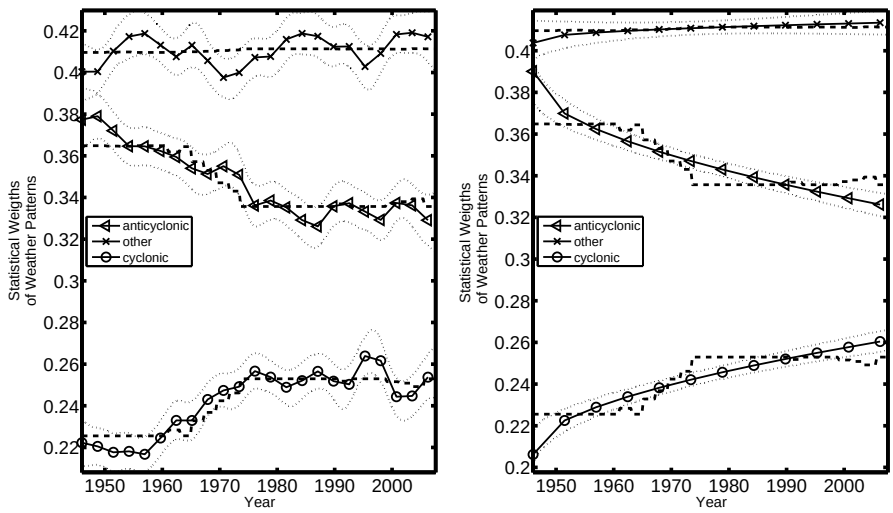

FIG. 7. Instantaneous statistical weights (34) and their robustness intervals (dotted) calculated with the local Gaussian kernel smoothing transition matrix(left) and with the polynomial trend transition matrix from Fig. 6 (right). Dashed lines denote the respective statistical weights calculated with the FEM-clustering procedure. The robustness intervals of the FEM-clustering method have almost the same size as the confidence intervals of the single trend model (since both methods are global approaches.

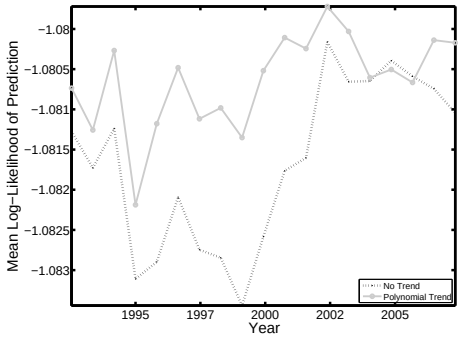

FIG. 8. Mean log-likelihood of predictions (35) as a function of τ (in years, $\Delta t = 6000$ days) for $P^{[1,\tau]}(t)$ estimated as: a stationary Markovian process without trend (dotted line), and a non-stationary Markovian process with polynomial trend (solid line).