,

Automated Generation of Reduced Stochastic Weather Models I: simultaneous dimension and model reduction for time series analysis*

Illia Horenko[**1], **Rupert Klein**[***2], **Stamen Dolaptchiev**[†2], and **Christof Schütte**[‡1]

[1] Institut für Mathematik II, Freie Universität Berlin
Arnimallee 2-6, 14195 Berlin, Germany

[2] Potsdam Institute for Climate Impact Research (PIK),
Telgraphenberg A31, 14473 Potsdam, Germany

This is a preliminary version. Do not circulate!

We present a method for simultaneous dimension reduction, model fitting and metastability analysis of high dimensional time series. The approach is based on the combination of hidden Markov models (HMMs) with localized principal component analysis (PCA) and fitting of multidimensional stochastic differential equations (SDE). We derive explicit estimators for PCA-SDE model parameters and employ the Expectation Maximization algorithm for numerical optimization of HMM-PCA-SDE parameters. We demonstrate the performance of the method by application to historical temperature data in Europe during 1976-2002. In a comparison with the standard SARMA (Seasonal Autoregressive Moving Average Model) technique for time series analysis the HMM-PCA-SDE-method exhibits better numerical performance and efficiency, especially on high-dimensional data sets. We also compare the results of both models w.r.t. errors of one–day temperature predictions.

This is a preliminary version. Do not circulate!

Introduction

In the experimental sciences, recent years have seen a dramatic explosion in the amount and precision of raw data that is available in the form of time series. Due to the development of computational and measuring facilities in geo-sciences (e.g. reanalysis techniques in meteorology) large amounts of measured and simulated information from all kinds of processes have been accumulated. All of these complex processes share the following properties: (i) they are multi-dimensional, i.e. they can be completely described and understood only from the observation or measurement of many of their characteristics simultaneously, (ii) the dynamics of the system is typically non-linear and non-stationary, (iii) many complex systems exhibit (hidden) phases or regimes which are persistent over long periods of time but are not (asymptotically) stable [1]. These properties imply also a classification of data-based methods for time series analysis in two major groups: dimension-reduction methods (aiming at identification of essential dimensions) and model reduction methods (methods for the construction of reduced representations of the dynamics). Let us shortly review both of them in the following.

Dimension reduction methods are aiming at the general task of finding the few most important, i.e., essential, degrees of freedom that can explain most of the observed processes and thus can help to understand the underlying mechanisms. The problem of dimension reduction becomes crucial when dealing with data-bases containing very large data-sets, e.g., libraries of climate data. Recent studies show that even relatively simple linear dimension reduction strategies, such as principal component analysis (PCA), often also referred to as empirical orthogonal function (EOF) technique [2, 3], allow

[*] Supported by the DFG research center MATHEON "Mathematics for key technologies" in Berlin.

[**] E-mail: horenko@math.fu-berlin.de

[***] E-mail: Rupert.Klein@PIK-Potsdam.de

[†] E-mail: stamen@PIK-Potsdam.de

[‡] E-mail: schuette@math.fu-berlin.de

for a significant compression of the information. However, such linear techniques as PCA, if applied to nonlinear phenomena like transitions between the persistent states, can be misleading and produce difficulties in the interpretation [4]; moreover, PCA detects the directions of maximal statistical variance which sometimes may not be the information wanted.

These problems can be circumvented by statistical separation of directions in which the distribution of data is Gaussian from non-Gaussian directions. Such NGCA approaches (non-Gaussian component analysis) are based on the insight that in many complex systems modes with Gaussian distributions can be interpreted as "noise" while such with non-Gaussian distributions contain "signals" [5, 6]. However, all current versions of NGCA are global and rely on a stationarity assumption for the underlying process; they are not yet extended to incorporate hidden phases.

Another dimension reduction approach, optimally persistent patterns (OPP), aims at analyzing the temporal behavior of the system—in contrast to finding the spatial correlations and directions with maximal spatial variation as in PCA-analysis. That is, OPP performs the data–based separation of slow and fast degrees of freedom based on the behavior of the multidimensional autocorrelation-functions [7]. The main problem of the OPP-method is that it is very sensitive w.r.t. the length of the observed series and the upper bound of integration of the autocorrelation function, and has difficulties with systems switching between hidden chaotic regimes [8]. Therefore there is considerable interest in the question of how to localize global linear techniques [9].

One class of approaches to this question considers non–linear generalizations of global/linear approaches (PCA/NGCA/ICA). This has been tried for PCA and has led to a method called NLPCA [10] (non–linear PCA), but not yet for NGCA. However, the NLPCA strategy is numerically expensive and not very robust, thus resulting in restricted applicability [11].

Another possibility to extend linear dimension reduction techniques is contained in the theory of indexing of high dimensional data-bases, where the problem was partially solved by combining correlation analysis with clustering techniques [12, 13], or in the context of the so-called projected clustering methods [14]. But due to the fact that the proposed methods rely on geometrical clustering of possibly high dimensional data–spaces, the resulting algorithms rely on some geometrical framework [15] and scale polynomially w.r.t. the length of the time series.

Alternatively, due to additional information encapsulated in the time component, it is possible to employ machine-learning or statistical techniques which scale linearly w.r.t. the length of the time series: the literature provides statistical approaches to regime switching [16, 17] or Bayesian approaches like hidden Markov models (HMMs) [18, 19, 20]. Recently, two of the authors proposed a method for simultaneous dimension reduction and clustering of time series into persistent phases. The approach is based on the combination of the HMM with PCA but can be also extended to combine HMM and NGCA [21]. The problem of simultaneous dimension reduction and phase identification is solved by optimization of an appropriate log–likelihood functional via the Expectation-Maximization algorithm (EM) [22, 21, 20]. It has been demonstrated that the resulting HMM-PCA algorithm allows the reliable detection of essential dimensions simultaneously with phase identification, leading to "localized essential dimensions" in the sense that they are different for different phases [21].

Methods for model reduction / reduced dynamics aim at finding a dynamical system that approximately describes the observed process in a low dimensional state space (where low means that it is much smaller than the dimension of the data). For our purposes, we can distinguish at least three main classes of related approaches for data-based model reduction: (i) Box-Jenkins identification strategies, (ii) Bayesian models or neuronal networks, and (iii) approaches which are based on fitting of the data with a global dynamical system that at best is "physical"/"interpretable" and usually given in the form of coupled stochastic/deterministic differential equations.

The first group of methods, (i), originated in econometrics at the beginning of 1970s and is also known under the name (S)ARIMA (seasonal autoregressive integrable models with moving average) [23, 24, 25]. The main idea of these methods relies on fitting the observed data with a discrete time stochastic difference scheme. The Box-Jenkins approach is restricted to the analysis of stochastic processes that can be made stationary by some suitable transformation of the data. Under certain circumstances, this can be achieved, e.g., by differencing the time series or subtracting a periodic component and analysing the autocorrelation functions. This is a serious limitation because it implies a large degree of user involvement into the process and prohibits automation [26].

The second group, (ii), is based on dynamical Bayesian networks, such as HMMs or neuronal networks [18, 10]. These are set-oriented approaches, since they decompose the configuration space of the system into several sets, where the dynamics of the system in each of the sets is described by an independent data model. The overall dynamics of the process is then governed by a hidden process switching between those sets. The overall model for the observed process in this case consists of the linear combination of the local models for each of the hidden sets with some time-dependent weights $\gamma_i(t)$ describing the probability for the hidden process to be in the hidden set i at time t. All of the approaches from this group that we are aware of are designed only in the context of *discrete* stochastic systems (e.g., autoregressive moving average models (ARMA) in combination with hidden jump processes [16]) and do not allow a physical understanding of the system. Moreover, and efficient and robust implementation for high-dimensional systems with hidden phases is still lacking.

To the best of our knowledge, the first application of the HMM-approach to low-frequency atmospheric variability, [27], investigates a time series resulting from the barotropic quasi-geostropic equations, an important model for the synoptic and planetary scale dynamics of the atmosphere in the middle latitutes. The authors applied the standard HMM-Gaussian approach to cluster the one-dimensional time series of the reduced variable exhibiting the longest correlation time. They point out the importance of the HMM technique and the construction of reduced stochastic models for extension of the predictability range of weather processes.

The third group of methods, (iii), attempts to fit a global mathematical model in the form of coupled stochastic/deterministic differential equations to observed data [28, 29]. Unfortunately, due to the "curse of dimensionality", the available methods can deal with high-dimensional data only under very specific assumptions (e.g., thermodynamic equilibrium, covariances of correlation matrices assumed to be diagonal, etc.). One way of dealing with high-dimensionality is essentially based on the fitting of the "global" data model in a full dimensionality with some low-dimensional stochastic/deterministic system by minimizing the distance between the solutions of the full and reduced models w.r.t. some (linear) projector defining the degrees of freedom of the reduced model. In the context of reduced models given in the form of ODEs, this technique is known as principal interaction patterns (PIP) in the literature. It was used for analyses of many applied problems in climate research [30, 31, 32, 8].

In this paper we present a novel method for simultaneous dimension reduction, SDE model reduction, *and* clustering of the time series into metastable states. The approach is based on the combination of HMM–PCA [21] with multidimensional SDE parameter fitting [33, 34, 35]. The problem is approached numerically by the optimization of an appropriate log–likelihood functional by means of the Expectation Maximization algorithm (EM), [22]. The presented method is used to analyze historical near-surface air temperatures in Europe betwen 1976 and 2002. For the present demonstration purposes, we have analysed publicly available ERA40 data, [36], onto a 20x29 grid in space and generated daily air temperatures at 2 meters elevation at 12:00h GMT. One quality check for the resulting reduced model consists of a comparison of the temperature predictions resulting from HMM-PCA-SDE with the actual temperature dynamics. Based on this test, we also compare our method with the 580–dimensional SARMA model, [24], for the same data set.

1 Seasonal Autoregressive Moving Average Model (SARMA)

We assume that the measurements of the process are given as discrete time–dependent vector–function $x_t : \mathrm{R}^1 \to \mathrm{R}^n$. The time step of this time series is the distance in time between two observations, it is assumed to be constant and we denote it by τ. In the context of seasonal auto-regressive models with moving average ($SARMA(p,q)$), the measurement x_t is modelled as

$$\begin{aligned} x_t &= K(t) + y_t, \\ y_t &= \sum_{\tau=1}^{p} \alpha_\tau y_{t-\tau} + \sum_{\tau=0}^{q} \beta_\tau W_{t-\tau} \end{aligned} \tag{1}$$

where $K(t) : \mathrm{R}^1 \to \mathrm{R}^n$ is some deterministic periodic vector function, W_t is an n-dimensional vector of white noise, α_τ, β_τ are τ-dependent $n \times n$ matrices and y_t is an n-dimensional stationary process of the type $ARMA(p, q)$ [37, 24]. The parameter p, as it can be seen from the formula (1), describes the depth of a process memory wrt. a state of the system (”internal memory”), whereas q is a depth of the memory wrt. the realizations of the noise-process (”external memory”). For example, a Markov-chain can be understood as ARMA(1,0)-process.

In order to estimate the function $K(t)$ one can for example Fourier-transform the data and filter out the Fourier components which exceed a certain threshold in the spectrum. The applicability of the FFT-technique is restricted to low dimensional cases, while for large n different dimensions of the measured signal x_t can be only treated separately making it difficult to find the high–dimensional periodic patterns in the data. After filtering out $K(t)$, according to the standard Box-Jenkins procedure [23], the resulting signal y_t should be checked with respect to the stationarity assumption, i. e. the decay of the autocorrelation and partial autocorrelation functions should be visually controlled and the time series should be differentiated if necessary. This property seriously restricts the possibility of automatization, since the user is involved in all of the stages of the data–analysis [26]. The problem gets even worse when dealing with high-dimensional data-sets because respective autocorrelation and partial autocorrelation functions also become multidimensional. Having estimated the order (p, q) of the model from the decay of both functions, one can estimate the parameter matrices $(\alpha_\tau, \beta_\tau)$ either through the maximum likelihood method [37] or equivalently through the solution of a system of matrix equations known as multidimensional Yule-Walker equations, [24]. From the numerical point of view this last step becomes expensive with growing number of dimensions n since the Yule-Walker system of equations has n^2 unknowns. This implies in general $\mathbf{O}(n^6)$ operations for the solution of the system (in the case when the resulting matrix is sparse the number of operations scales as $\mathbf{O}\left(n^2 \log(n)\right)$).

2 HMM-PCA-SDE

2.1 Model

For a given n-dimensional time series x_t of the length T we aim at simultaneous: (i) clustering into K metastable sets, (ii) identification of $m << n$ essential dimensions for each of the metastable states (given in the form of local linear filters $(\mu_i, \mathbf{T}_i)$, where $\mu_i \in \mathrm{R}^n$ and $\mathbf{T}_i \in \mathrm{R}^{n\times m}$, $i = 1, \dots, K$), and (iii) fitting of K optimal m-dimensional SDE models for each of the metastable states describing the reduced dynamics of the system in essential degrees of freedom. We assume that the essential dynamics in the metastable state i can be parametrized with the help of the linear m-dimensional SDE of the form

$$\dot{z}(t) \quad = \quad F_i \left(z(t) - \bar{\mu}_i\right) + \Sigma_i \dot{W}(t) \tag{2}$$

where $F_i, \Sigma_i \in \mathrm{R}^{m\times m}, \bar{\mu}_i \in \mathrm{R}^m, i = 1, \dots, K$ and $z(t) = \mathbf{T}_i^\top \left(x_t - \mu_i\right)$. We choose this form of the SDE model because of two main reasons: (1) as shown in [35], for the SDE of this type it is possible to derive *explicit* formulas for estimating the parameters that are optimal for each individual time step τ, and (2) application of the HMM-framework to these equations allows to describe complex multidimensional multi-well structures without the exponential growth of the underlying model parameters (“curse of dimensionality”) [38].

The formal solution to (2) on the time interval $[t, t + \tau]$ is given by

$$z(t+\tau) \quad = \quad \bar{\mu}_i + e^{\tau F_i}\left(z(t) - \bar{\mu}_i\right) + \int_0^\tau e^{(\tau - s)F_i} \Sigma_i dW(s). \tag{3}$$

Thus, the probability density $\rho_\lambda(z_{k+1}|z_k)$ of observation of z_{k+1} at time $k + 1$ under the condition of observation of z_k at k is proportional to

$$\exp\left[-\frac{1}{2}\xi_k^\top R_i^{-1}(\tau)\,\xi_k\right],$$

where

$$\begin{aligned} \xi_k &= z_{k+1} - \bar{\mu} - e^{\tau F_i}\left(z_k - \bar{\mu}_i\right) &(4)\\ R_i(\tau) &= \int_0^\tau e^{sF_i}\Sigma_i\Sigma_i^\top e^{sF_i^\top}\,ds. &(5)\end{aligned}$$

The subscript λ, which denotes the complete set of parameters defining the reduced dynamics of the observed system, will be defined in (9) below.

The integral in (5) can be solved by partial integration resulting in the following linear matrix equation

$$R_i(\tau)F_i^\top + F_i R_i(\tau) = e^{\tau F_i}\Sigma_i\Sigma_i^\top e^{\tau F_i^\top} - \Sigma_i\Sigma_i^\top\,. \tag{6}$$

Thus, we can express the joint conditional probability density of the observation of the time series (z_k) from our model by

$$\begin{aligned} p(z|\lambda) &= \prod_{k=1}^{T-1}\rho_\lambda(z_{k+1}|z_k) = \prod_{k=1}^{T-1}\rho_0(\tau)\exp\left[-\frac{1}{2}\xi_k^\top R_i^{-1}(\tau)\,\xi_k\right] &(7)\\ \rho_0(\tau) &= (2\pi)^{-n}/\sqrt{\det\left(R_i(\tau)\right)}. &(8)\end{aligned}$$

As it can be seen from (7), for fixed τ, a given observation sequence x_t and projector parameters $(\mu_i, \mathbf{T}_i)$ it is sufficient to define $((\bar{\mu}_i, \exp\tau F_i, R_i(\tau))$ in order to do a maximizitation of (7) for a given observation sequence z_t. Finding the optimal values of this objects allows to calculate the parameters of (2). So the complete set of parameters describing the reduced dynamics of the observed system in the metastable set i is

$$\lambda_i = (\bar{\mu}_i, \exp\tau F_i, R_i(\tau), \mathbf{T}_i, \mu_i). \tag{9}$$

2.2 Likelihood function

In order to solve the problems (i)-(iii) simultaneously, we combine the results of [33, 34] and [21] and construct a likelihood function as a linear combination of two functionals

$$\mathbf{L} = \beta\mathbf{L}_{\mathrm{HMM-SDE}} - \alpha\mathbf{L}_{\mathrm{HMM-PCA}}, \tag{10}$$

where $\mathbf{L}_{\mathrm{HMM-PCA}}$ and $\mathbf{L}_{\mathrm{HMM-SDE}}$ are the functionals for dimension-reduction and SDE model-reduction, respectively, and the scalar parameters (α, β) are the corresponding user-defined weights of both functionals. $\mathbf{L}_{\mathrm{HMM-PCA}}$ is the least-squares residuum-functional describing the distance from the original n-dimensional data–set to the reduced m-dimensional linear manifold

$$\mathbf{L}_{\mathrm{HMM-PCA}}(x_t, \mathbf{T}_i, \mu_i) = \sum_{i=1}^{K}\sum_{t=1}^{T}\gamma_t(i)\left\|(x_t - \mu_i) - \mathbf{T}_i\mathbf{T}_i^\top(x_t - \mu_i)\right\|_2^2. \tag{11}$$

Here $\gamma_t(i)$ is the probability for the system to be in the metastable state i at time t.The functional $\mathbf{L}_{\mathrm{HMM-PCA}}$ depends on the projector matrices $\mathbf{T}_i$ and *center vectors* $\mu_i \in \mathrm{R}^n$. Moreover, the projectors $\mathbf{T}_i$ are subjected to the orthogonality condition:

$$\mathbf{T}_i^\intercal\mathbf{T}_i = Id^{m\times m}, \tag{12}$$

The functional (11) can be equivalently written as

$$\begin{aligned} \mathbf{L}_{\mathrm{HMM-PCA}} &= \sum_{i=1}^{K}\sum_{t=1}^{T}\gamma_t(i)\Big((x_t-\mu_i)^\intercal(x_t-\mu_i)\\ &- \operatorname{tr}\left[\mathbf{T}_i^\intercal(x_t-\mu_i)(x_t-\mu_i)^\intercal\mathbf{T}_i\right]\Big) + \sum_{i=1}^{K}e^T\Lambda_1^i\cdot\left(Id^{m\times m} - \mathbf{T}_i^\intercal\mathbf{T}_i\right)e &(13)\end{aligned}$$

where $e = (1, 1, ..., 1) \in \mathrm{R}^m$, $\Lambda_1^i \in \mathrm{R}^{mxm}$ being the matrix of Lagrange multipliers and $\cdot$ denoting the component-wise multiplication of two matrices. It is important to mention that no assumptions about the statistics of the observation sequence x_t are necessary for construction of the optimal minimizer of (13), and that the error of the dimension reduction in l_2-norm is bounded by the actual value of the functional. This allows to control the quality of dimension reduction since the reduced dimensionality m can be chosen such that the corresponding value $\mathbf{L}_{\mathrm{HMM-PCA}}$ is less then a predefined threshold.

The $\mathbf{L}_{\mathrm{HMM-SDE}}$ is the corresponding log-likelihood SDE-functional. It is the logarithm of the probability of the observed data sequence (7) conditioned on the parameters $\lambda_i = (\bar{\mu}_i, \exp \tau F_i, R_i(\tau), \mathbf{T}_i, \mu_i)$ of the SDE

$$\begin{aligned}
\mathbf{L}_{\mathrm{HMM-SDE}} &= a - \frac{1}{2}\sum_{k=1}^{T-1} \gamma_t(i) \log\det R_i(\tau) \\
&- \frac{1}{2}\sum_{k=1}^{T-1} \gamma_t(i) \left(z_{k+1} - \bar{\mu}_i - e^{\tau F_i}\left(z_k - \bar{\mu}_i\right)\right)^\top R_i^{-1}(\tau) \left(z_{k+1} - \bar{\mu}_i - e^{\tau F_i}\left(z_k - \bar{\mu}_i\right)\right) \\
&+ \sum_{k=1}^{T-1} \gamma_t(i)\Lambda_2^i(k) \left(z_k - \mathbf{T}_i^\top \left(x_k - \mu_k\right)\right) \qquad (14)
\end{aligned}$$

where a is some constant and $\Lambda_2^i(k) \in \mathrm{R}^m$ is the vector of Lagrange multipliers.

Putting (13) and (14) together into (10) we can start fitting the parameters by maximizing the resulting functional with the help of the *maximum likelihood principle* applied to the functional $L(\lambda, x_t)$, i.e., we consider the observation sequence x_t as being given and ask for the variation of the probability in terms of the parameters and the hidden sequence of metastable sets X_t. Since the maximum of (7) coincides with the maximum of its logarithm (14), the maximum likelihood principle then simply states, that the optimal parameters maximize the observation probability of the time series x_t and are given by the absolute maximum of $\mathbf{L}$.

In order to apply the HMM-framework to maximization of (10), we first have to make an assumption that the hidden process switching between the metastable states is Markovian, i.e., the probability of the state change depends on the current state only. This assumption is connected to the characteristic timescale at which the memory kernel of the system is decaying, is satisfied for a wide class of applications and allows to use the standard Expectation-Maximization algorithm [39], often also called the Baum-Welch algorithm [40, 19] for the maximization. The Expectation-Maximization (EM) algorithm is a maximum likelihood approach that improves iteratively an initial parameter set, and converges to a local maximum of the likelihood functional (10). Its two steps, the E- and M-steps, are iteratively repeated until the improvement of the likelihood becomes smaller than a given limit. In all other details the EM algorithm used herein follows standard procedures.

To apply the EM algorithm to a given observation sequence, we have to set up a HMM by assuming a finite number K of hidden states and initial values for all remaining parameters.

EM steps. There is no known way to analytically determine the model parameters that globally maximize the probability of the given observation sequence. We can, however, estimate λ such that it locally maximizes the conditional probability to observe a certain sequence x_t, where the conditional probability function is interpreted as $P(x_t \mid \lambda) = \exp \mathbf{L}(\lambda)$. Since

$$\begin{aligned}
P(x_t \mid \lambda) &= \exp\left(\beta \mathbf{L}_{\mathrm{HMM-SDE}}\right) \exp\left(-\alpha \mathbf{L}_{\mathrm{HMM-PCA}}\right) \\
&= \exp\left(\beta + \alpha\right) p(z_t|\lambda) \exp\left(-\mathbf{L}_{\mathrm{HMM-PCA}}\right), \qquad (15)
\end{aligned}$$

with the observation probability $p(z_t|\lambda)$ as defined in (7). In order to be able to interpret the function P as the probability we have to impose $(\alpha + \beta) \leq 0$. Since, given the orthogonality constraint in (12), we have $0 \leq \mathbf{L}_{\mathrm{HMM-PCA}}$, the expression $\exp\left(-\mathbf{L}_{\mathrm{HMM-PCA}}\right)$ can also be interpreted as a probability density. This also implies the probabilistic understanding of the exponent of the functional (10). We employ the

EM algorithm to maximize both likelihood and log-likelihood functions simultaneously. Starting with some initial model λ_0, we iteratively refine the model within two steps:

- The Expectation-step: In this step the state occupation probability $\gamma_t(i) = P(X_t = i \mid x_t, \lambda)$, and the transition probability $\xi_t(i,j) = P(X_t = i, X_{t+1} = j \mid x_t, \lambda)$, are calculated for each time t in the sequence, given the observation x_t and the current model λ. The calculation of these variables can be done in a standard HMM way [22].
- The Maximization-step: This step finds a new model $\hat{\lambda}$ via a set of reestimation formulas. The maximization guarantees that the likelihood does not decrease in each single iteration.

In order to apply the EM-algorithm, we need to reestimate parameters λ_i describing the local SDE models and essential dimensions via their maximum likelihood estimator. Hereby, the observations x_t have to be weighted with the probability $\gamma_t(i)$ to be in the hidden state i. In order to calculate this reestimation formulas we fix the sequence X_t of the hidden states (this means also keeping the sequence of $\gamma_t(i)$ fixed) and calculate the derivatives of the functional (10) wrt. the parameters λ_i. Setting all of the partial derivatives to zero for some fixed reduced dimensionality m we get a coupled system of nonlinear algebraic equations for the parameters which can be solved numerically with the Newton–method [41]. However, the numerical effort of the Newton-method in this case will scale as $\mathcal{O}(n^6 m^6)$ resulting in long computation times for a large number of dimensions. For high-dimensional data, instead of the optimization of (10), we suggest first to optimize the functional (13) (since it is independent of the SDE–parameters and it is possible to calculate the explicit optimum of this functional and to get the explicit expressions for μ_i and $\mathbf{T}_i$, for details see [21]) and then to use the derived expressions for the optimal projector parameters in an optimization of (14) [35] which gives the following explicit estimators of parameter λ_i

$$\begin{aligned}
\mu_i &= \frac{1}{\sum_{t=1}^{T} \gamma_t(i)} \sum_{t=1}^{T} \gamma_t(i) x_t, && (16)\\
C_i &= \sum_{t=1}^{T} \gamma_t(i) \left(x_t - \mu_i\right) \left(x_t - \mu_i\right)^{\mathsf{T}}, && (17)\\
C_i \mathbf{T}_i &= \mathbf{T}_i \max_m \left(spec(C_i)\right), && (18)\\
z(t) &= \mathbf{T}_i^{\mathsf{T}} \left(x_t - \mu_i\right), && (19)\\
e^{\tau \hat{F}_i} &= \mathrm{Cor}_i \mathrm{Cov}_i^{-1}, && (20)\\
\bar{\mu}_i &= \bar{z} - (\mathrm{Id} - e^{\tau \hat{F}_i})^{-1} \delta_i, && (21)\\
R_i &= \frac{1}{\sum_{t=1}^{T-1} \gamma_t(i)} \sum_{t=1}^{T-1} \gamma_t(i) d_t {d_t}^{\mathsf{T}}. && (22)
\end{aligned}$$

where $\max_m \left(spec(C_i)\right)$ denotes m dominant eigenvalues of the covariance matrix C_i and

$$\begin{aligned}
\bar{z}_i &= \frac{1}{\sum_{t=1}^{T-1} \gamma_t(i)} \sum_{t=1}^{T-1} \gamma_t(i) z_t, && (23)\\
d_t &= z_{t+1} - \bar{\mu}_i - e^{\tau \hat{F}_i} (z_t - \bar{\mu}_i), && (24)\\
\mathrm{Cor}_i &= \frac{1}{\sum_{t=1}^{T-1} \gamma_t(i)} \sum_{t=1}^{T-1} \gamma_t(i) (z_{t+1} - \bar{z}_i)(z_t - \bar{z}_i)^{\mathsf{T}}, && (25)\\
\mathrm{Cov}_i &= \frac{1}{\sum_{t=1}^{T-1} \gamma_t(i)} \sum_{t=1}^{T-1} \gamma_t(i) (z_t - \bar{z}_i)(z_t - \bar{z}_i)^{\mathsf{T}}, && (26)\\
\delta_i &= \frac{\sum_{t=1}^{T-1} \gamma_t(i) \left(z_{t+1} - z_t\right)}{\sum_{t=1}^{T-1} \gamma_t(i)}. && (27)
\end{aligned}$$

As we can see, for a fixed sequence of hidden states X_t and a given observation sequence x_t, expressions (16-22) straightforwardly provide the unique explicit estimators for unknown parameter set λ_i. Note that these estimators are independent from the parameters α and β (15) which represent the weights of the dimension reduction and model reduction functionals in the optimization of $\mathbf{L}$. Moreover, from (15) it becomes clear that both of these parameters build a constant exponential factor in front of the observation probability and hence are not influencing the optimization procedure wrt. λ and γ_t. If wanted, we can use the so-defined estimators as initial values of the Newton method for determining the optimal parameter λ.

The E- and M-steps are iteratively repeated until a predetermined maximal number of iterations is reached or the improvement of the likelihood becomes smaller than a given limit. The entire EM algorithm has the nice property that the likelihood function is non-decreasing in each step, i.e., we iteratively approximate local maxima. As for the scaling of numerical effort, the HMM-PCA-SDE method is linear in the length of the observation series x_t, quadratic in the number of hidden Markov states (essentially since the transition matrix elements of the hidden Markov chain should be estimated), has order $\mathbf{O}(n)$ in the dimensionality of x_t (since m dominant eigenvectors in n dimensions can be computed with a Lanczos method) and cubic in the reduced dimensionality m (since the Cov_i matrix in (20) has to be inverted and the log of $e^{\tau \hat{F}_i}$ has to be computed, if this matrix can be assumed to be sparse the corresponding number of iterations should be $\mathbf{O}(m \log(m))$). Compared to SARMA where the number of operations needed for the solution of the Yule-Walker system is proportional to n^6 in the general case, the HMM-PCA-SDE-approach is applicable to systems with much higher dimensionality, even in the case of no dimension reduction (i. e. $m = n$). This feature is demonstrated in the next section where both of the methods are used for analysis of the same multidimensional data-set.

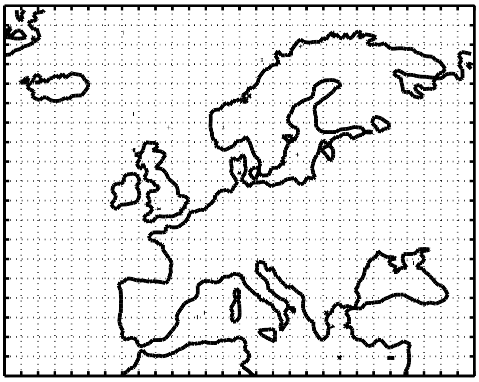

Fig. 1 Spatial grid where the temperature data are available. Each of the temperatures at each of the 580 grid points was given in a form of the time series over 9736 days between 1976 and 2002.

3 Analysis of historical temperatures over Europe between 1976 and 2002.

In order to demonstrate the performance of the presented HMM-SDE-PCA-method we take for x_t daily mean values of the 2 m surface air temperature from the ERA 40 reanalysis data [36]. We consider a region with the coordinates: 27.0 W – 45.5 E and 33.0 N – 73.5 N (see Fig. 1), which includes Europe and a part of the Eastern North Atlantic. The original spatial resolution of the data was reduced to approximately $2.0^\circ \times 2.5^\circ$ latitude and longitude by spacial averaging. Thus we have temperature values on a grid of 20×29 points. The time record is from 1976 till 2002 and it includes 9736 daily values from 12:00h GMT observations.

Before we apply the HMM-PCA-SDE setting to the data we first check whether they can be described by the SDE-model with additive noise. The form of the likelihood function (7) indicates that

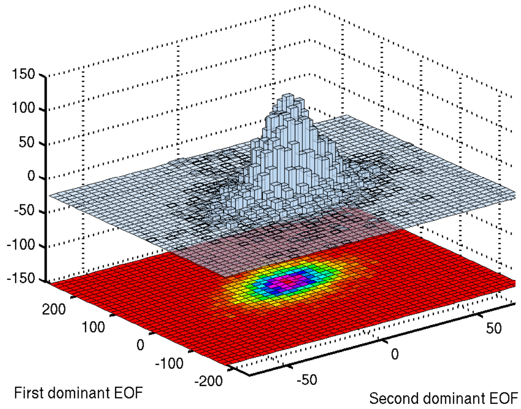

Fig. 2 Distribution of the temperature increments after the subtraction of the deterministic part of the dynamics defined by $\dot{x}_t = F_i(x_t - \mu_i)$. Here is shown the projection of the resulting distribution on two dominant EOFs describing $> 90\%$ of the dynamics. Standard statistical tests (like χ^2–test, Kolmogorov–Smirnov–test and Shapiro–Wilks–test) for the probability of error of type I $\alpha = 0.05$ show that the resulting distribution is quite close to a Gaussian and indicate that a data-model with additive noise is reasonable.

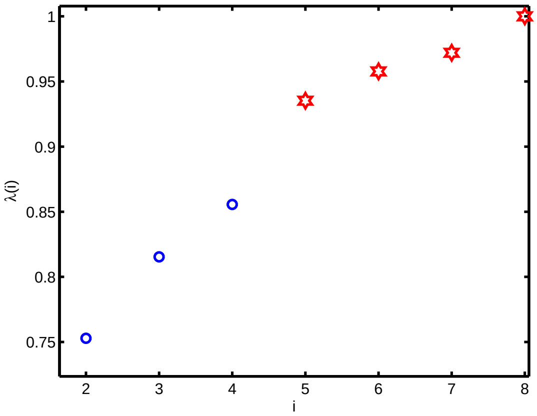

Fig. 3 Spectrum of the transition matrix of the HMM-PCA-SDE hidden Markov chain indicates the presence of 4 hidden metastable states.

after substraction of the deterministic trend the distribution of increments of the SDE-dynamics (2) should be Gaussian. The two–dimensional histogram of the projection of this distribution on two dominant EOFs is shown in Fig. 2. Qualitatively, the distribution in two dominant local EOFs is very much reminiscent of a Gaussian. This hypothesis is quantitatively verified by applying some standard statistical tests (like χ^2–test, Kolmogorov–Smirnov–test and Shapiro–Wilks–test). Application of these tests to marginal statistical distributions of the reduced trend–free dynamics in 20 dimensions affirms the Gaussian hypothesis (for the probability of error of type I being $\alpha = 0.05$). This feature indicates the data-model with additive noise [37] and so far validates the application of the HMM-PCA-SDE model with SDE-part (2).

Application of the HMM–PCA–SDE method for $m = 20$ (describing the 99.2% of the data variance) with 7 hidden states to the time series indicates the presence of four metastable states (see the gap in the spectrum of the resulting transfer operator in Fig. 3). Application of the HMM–PCA–SDE–approach in the case of 4 hidden states and reduced dimension $m = 20$ results in the hidden probability functions $\gamma_i(t)$ shown in Fig. 4. The function $\gamma(t)$ describes the temporal dynamics of probabilities for the observed

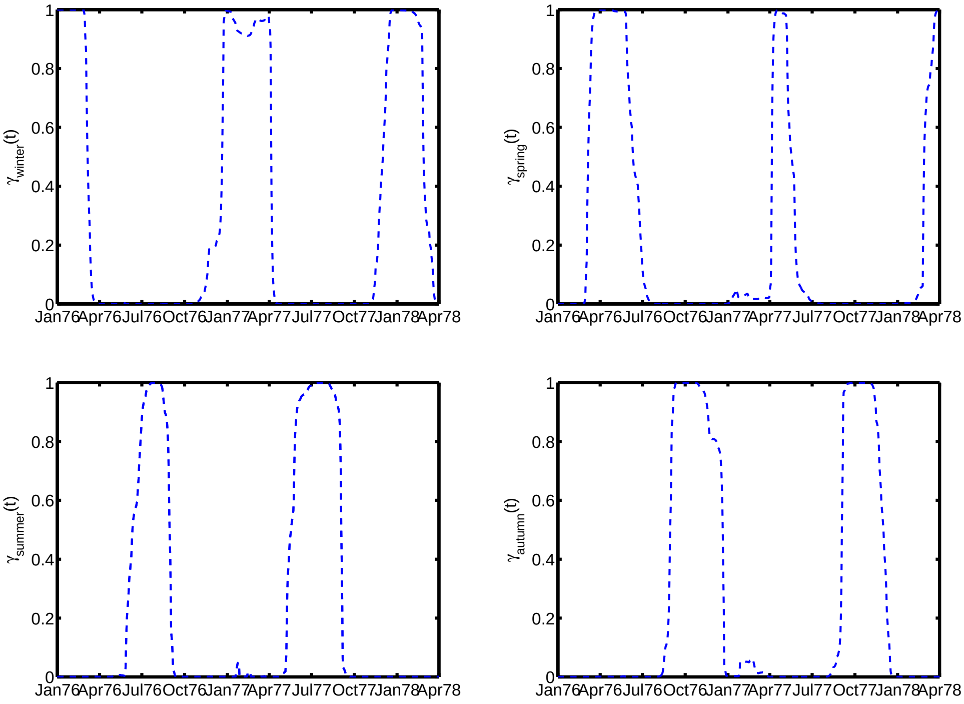

Fig. 4 Part of the identified $\gamma_i(t)$ function describing the probability of the hidden weather process to be in the state i at time t. At each moment t all of these 4 probabilities sum to 1.0.

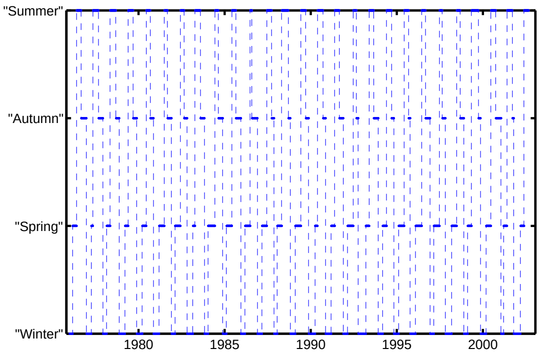

Fig. 5 Viterbi path of the hidden Markov chain.

temperature data to be described by each of the four local weather models ("winter", "spring", "summer" and "autumn"). This function can be understood as optimal low–dimensional coarse–grained description of the full dynamics in many dimensions and can be further analyzed itself with standard techniques of time series analysis (like Fourier– or ARMA–analysis). Notice that the individual probabilities for the hidden Markov states representing the four seasons exhibit sharp transitions, and that they are practically zero everywhere outside the season they are supposed to represent. This indicates that the identification of the states is unambiguous. This is remarkable, in particular, for spring and autumn, as one could intuitively guess that they are more of a superposition of "winter" and "summer" with a little bit of an individual ingredient added. In contrast, we find clear distinctions between all four seasons in the data.

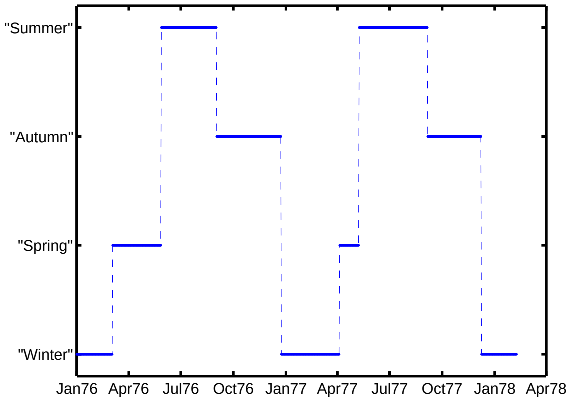

Fig. 6 Part of the Viterbi path of the hidden Markov chain from Fig. 5. 4 hidden states can be roughly identified with 4 seasons.

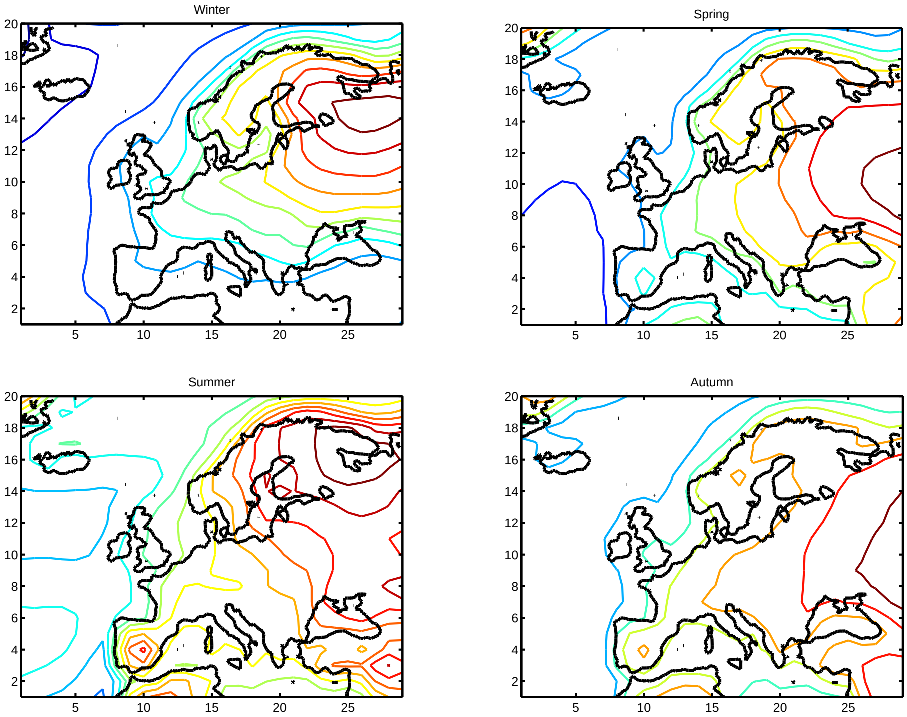

Fig. 7 Dominant dimensions of identified metastable sets as contour plots (contain > 90% of the dynamics in the corresponding sets).

Applying the Viterbi–algorithm [18] to the probabilities $\gamma_i(t)$ and to the identified local models we can find *the most probable discrete hidden path* of the system (also called Viterbi–path) which is shown in Figs. 5 and 6. In this graph, each of the time instances in the temperature data series is now assigned to one of the four seasons. In contrast to a "common sense" approach, in which the seasons would be defined as a priori known intervals of time (e.g., from December 22 to March 21 for astronomical winter), here this assignment is not predefined but rather depends on the corresponding probabilities

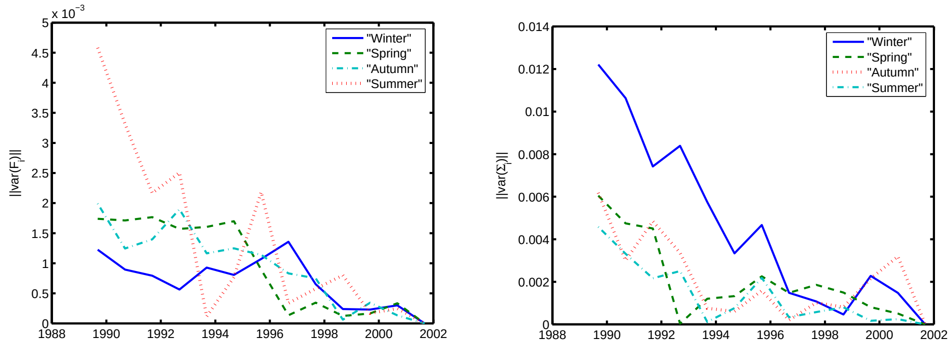

Fig. 8 Relative variation of the difference between the parameter matrices estimated from the complete data vector (i.e. temperature measurements from 1976 to 2003, approximately 2300 data–points for each season) and the parameters estimated from merely part of the data (i.e. temperature measurements from 1976 to the year shown on the abscissa of the graph.)

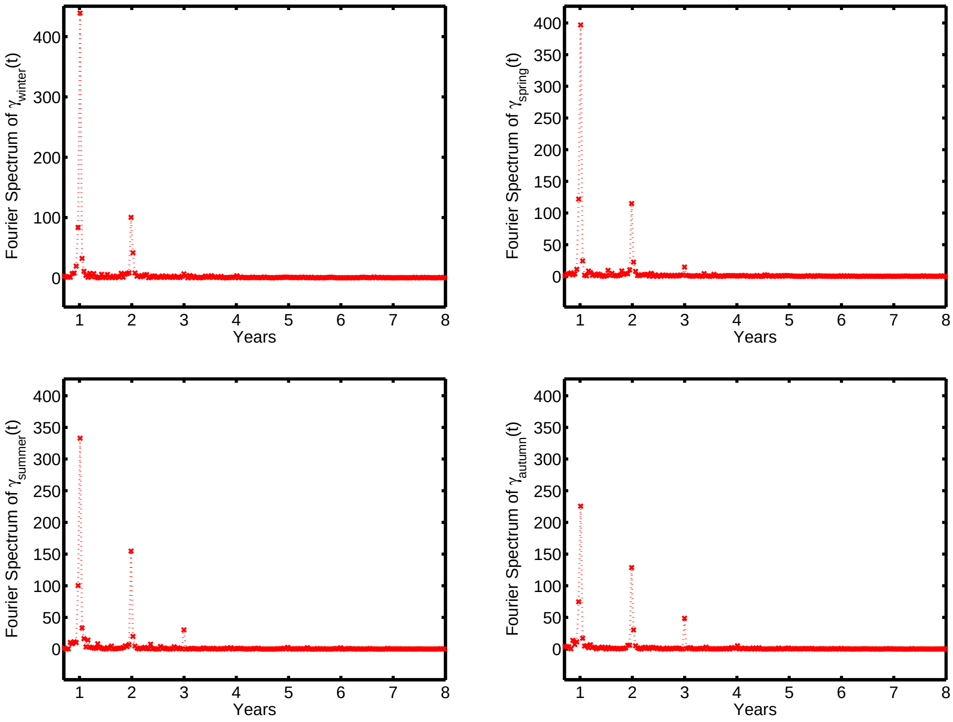

Fig. 9 Fourier spectra of the components of $\gamma(t)$-function from Fig. 4. They indicate the presence of one–, two– and three–year cycles in the data.

$\gamma_i(t)$. They describe a certain probability distribution of temperatures in Europe at a given time being properly represented by a certain seasonal SDE–model.

The dominant model dimensions in the seasonal states (the columns of T_i which correspond to directions with maximal variance) are presented in Fig. 7 and give an idea about the correlation patterns in the data. Positive and negative values of dominant eigenvectors correspond to correlating and anti–correlating geographical areas, respectively. That is, when the reduced one–dimensional temperature trajectory in the metastable state goes up, the temperature in areas corresponding to positive components of the dominant eigenvector increases and at negative ones decreases. As it can be deduced from

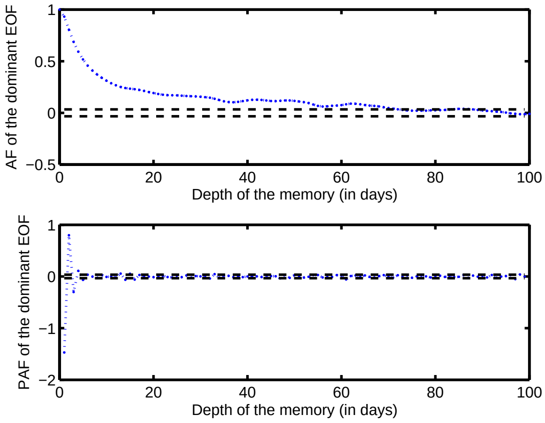

Fig. 10 Autocorrelation function (upper panel) and the partial autocorrelation function (lower panel) as being determined in the course of the Box–Jenkins procedure of SARMA-model for the dominant EOF (after filtering out the periodical component and stationarizing the data). Black dashed lines mark the level of statistical significance of the analyzed time series.

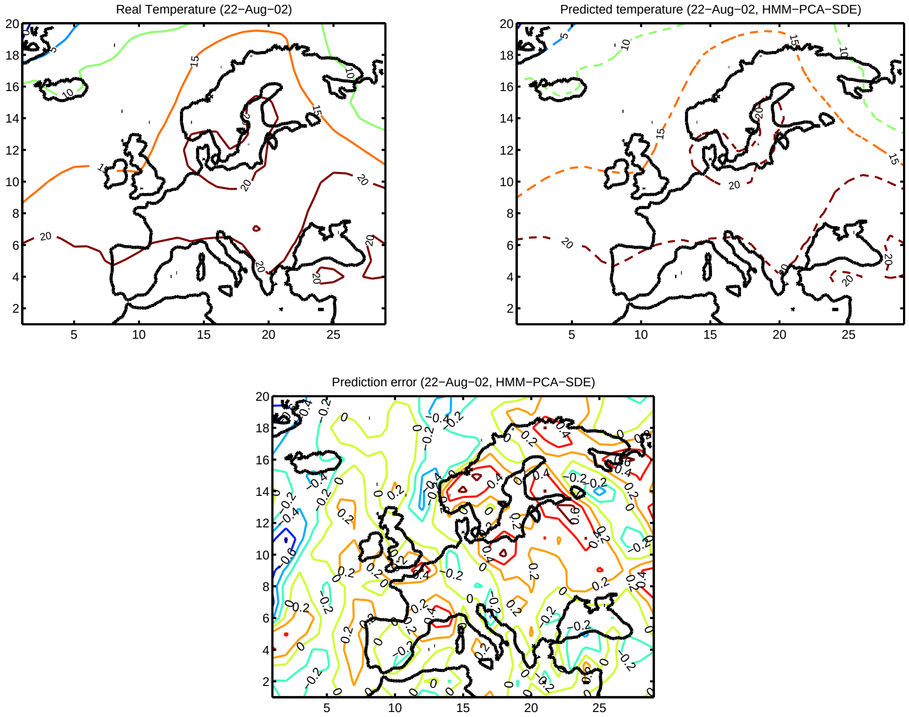
Real Temperature (22−Aug−02)
Predicted temperature (22−Aug−02, HMM−PCA−SDE)
Prediction error (22−Aug−02, HMM−PCA−SDE)
20
18
16
14
12
10
8
6
4
2
5
10
15
20
25

Fig. 11 Comparison of the real temperature (upper left) and the corresponding prediction with the HMM-PCA-SDE model (for $m = 20$ and 4 hidden states) (upper right). The error of prediction is shown in the lower panel.

the comparison of eigenvectors in Fig. 7, the correlation patterns are changing in the time together with the Viterbi–path from Fig. 5. The patterns found here are strongly reminiscent of those found in the established literature.

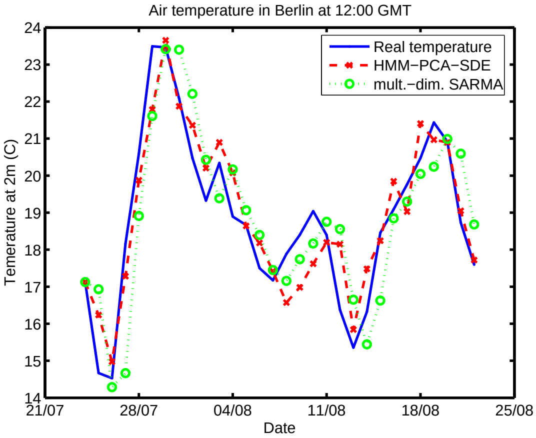

Fig. 12 Temperature time series for Berlin in summer 2002 (blue solid) compared to HMM-PCA-SDE predicted temperature (red dashed, crosses) and SARMA(4,0) (green dotted, circles). The mean prediction error for 1-day HMM-PCA-SDE is 0.7C, for SARMA(4,0) 1.2C.

As mentioned above, the locally linearized form of the SDE allows for explicit and time step independent estimation of the parameters, which is in contrast to estimators based on some sort of numerical discretization of the underlying SDE. An issue to be checked is the reliability of the local SDE model estimation in the four hidden states wrt. the length of the data series. In Fig. 8 we compare the estimated parameter matrices obtained from subsets of the total data set with those obtained from the full set of data. The relative variation of the estimated parameters is quite low, so that the length of the data set appears to be sufficient for the estimation. The application of more advanced variance estimators is subject of future investigation.

Figure 9 shows the Fourier spectra of the occupation probabilities $\gamma_i(t)$. Besides the obvious one-year periodicity mode, we clearly identify further strong perdiodic contributions at two- and three-year periods. This is strongly reminiscent of related observations in the meteorological literature, [42, 43, 44, 45, 46].

The next task is to compare the HMM-PCA-SDE model with one of the existing data-based models used for predictions in time series analysis. As a benchmark model we took a 580–dimensional $SARMA(4,0)$ model. In order to stationarize the data and to determine the periodic trend $K(t)$ in (1), we applied a Fourier filter to each of the systems dimensions (i.e. to the temperature time series at each of the grid-points) separately. As mentioned earlier, this makes it difficult to get an impression about the periodicity of the process as a whole, since we obtain only the Fourier spectra of some low-dimensional projections). In contrast, the HMM-PCA-SDE method provides direct access to the global behavior of the multidimensional system through the $\gamma(t)$–function (see Figs. 5, 9).

Visual control of the decay of 580 autocorrelation functions indicated that the data becomes stationary to a good approximation and the partial autocorrelation indicates that the data can be described with AR (auto regressive) model of order 4 (see Fig. 10)[23]. In order to fit the parameters of the 580-dimensional process, we first calculated the 580×580-autocorrelation functions and inserted them into the Yule-Walker system of equations [24]. The solution of the Yule-Walker system provides the estimators for SARMA parameters α_τ. One immediate result of this analysis is the temporal decorrelation of the dominant EOF as shown in Fig. 10. The characteristic "memory" for this mode is found to be about 20 days, which is in good agreement with earlier suggestions in [47], which were obtained by very different means.

In order to produce the temperature prediction at time t for both of the models (HMM-PCA-SDE ($m = 20$) and SARMA(4,0)) we took the actual temperature data from the beginning of the observation (Jan. 1, 1976) to the time ($t - 1day$) and fitted both of the models to the resulting time series. Then we calculated a temperature prediction for time t from these models, including the expectation value and the variance for both of them. In the case of HMM-PCA-SDE, the prediction is first made for the hidden process $\gamma_i(t)$. As a first attempt, we constrained the prediction to a Markovian case assuming that the underlying hidden process is Markovian and the predicted value $\gamma_i(t)$ is dependent only on the value of $\gamma_i(t-1)$. Generally, this assumption is not necessary, and one could as well fit a SARIMA(p,q) process to the series of $\gamma_i(0), \ldots, \gamma_i(t-1)$ with $(p > 1, q > 0)$ in order to account for the memory effects in the data. We intend to explore these options in more detail in a future publication. Given the Markovian assumption, the temperature prediction at time t is a linear combination of the predictions based on the individual SDE models in the hidden states weighed by the predicted hidden probabilities $\gamma_i(t)$. Fig. 11 shows a comparison of a typical temperature distribution in summer 2002 together with the 1-day predictions from both of the models.

The 20-dimensional HMM-PCA-SDE dynamics reasonably captures the temperature dynamics in the 580-dimensional space with a maximal prediction error of 0.8 °C. The maximal error occurs predominantly near the edges of the map. This can be explained by the influence of the regions lying outside of our grid which are not covered by the measurement data and are therefore not described by the model). Fig. 12 shows the comparison of the actual air temperature in Berlin with the predictions calculated from both models. The 20-dimensional HMM-PCA-SDE predicts the temperature somewhat more reliable than SARMA, the mean prediction error being 0.7 °C for HMM-PCA-SDE compared to 1.2 °C for SARMA.

A more dramatic difference arises wrt. computing times. The SARMA approach is much more expensive than HMM-PCA-SDE, which is crucial when dealing with high-dimensional data as in the present case. After filtering out the periodic components, the next step in the SARMA approach is the calculation of $336400 = 580 \times 580$ cross-correlation functions. Then a system of 336400 linear equations is to be solved. Despite the fact that the resulting system is sparse, so that efficient numerical solvers can be applied, the fitting of the SARMA model parameters took 16 hours on a modern PC as compared to 6 minutes for HMM-PCA-SDE.

4 Conclusion

We present here a novel hidden Markov model (HMM)–based method for simultaneous dimension reduction, stochastic differential equation (SDE)–fitting and clustering of the time series data. The method is based on a combination of the HMM approach, SDE parameter fitting and local principal component analysis (PCA). Incorporation of the local PCA analysis helps to map the clustering problem into a low dimensional space. We have demonstrated the application of the method for a historical temperature time series and tested the quality of the resulting model for stochastic predictions. We also compared the presented approach to the existing multidimensional SARMA-approach based on the fitting of the observed data with a multidimensional discrete autoregressive process. We have demonstrated that the quality of the temperature prediction in our sample meteorological application and the prediction horizon in both models depend on the size of the underlying area, such that reliable predictions over longer periods of time should be based on data covering much larger areas. This will be accounted for in future, more exhaustive meteorological and climatological investigations based on the new technology developed here.

The HMM-PCA-SDE–approach for analysis and predictions of high-dimensional stochastic dynamics is more attractive then SARMA strategy, because besides much less computational effort, HMM-PCA-SDE provides insight into the dynamics: (a) in the form of the hidden Viterbi–path allowing to assign the data to each of the local models, see Figs. 5 and 6, (b) in the form of local dominant PCA dimensions which allow the interpretation in the sense of different correlation patterns for each of the hidden states or regimes, see Fig. 7). In our opinion, the most important feature of the algorithm is its possibility to extract a low–dimensional description of the dynamics (function $\gamma(t)$) and further analyze it by means of standard time series methods. For example, one can find temporal patterns in the process of the seasonal change, investigate memory effects and find some general principles in the process of seasonal transitions. These aspects will be explored in future publications.

One interesting issue yet to be approached is the analysis of other geographical regions and comparison of the local principal dimensions identified with the help of HMM-PCA-SDE to the results produced by other dimension reduction approaches like principal interaction patterns (PIP) or optimally persistent patterns (OPP) [32, 7]. As was already demonstrated by A. Majda and co-workers, the PIP-method produces a more reliable optimal basis wrt. the reproduction of the switching behavior than the EOFs or the OPP-method [8]. However, the PIP-method relies on the availability of the full mathematical model, given in the form of the system of ODEs, in all of the observed degrees of freedom, which is not always the case. On the other hand, besides the fact that OPP is a purely data–based method and does not need any assumptions about the model, the procedure is very sensitive to numerical errors arising in the integration of multidimensional autocorrelation functions. Therefore, it will be interesting to compare all these approaches localizing the dimension reduction to each of the identified hidden states. This is another topic of ongoing research.

Acknowledgements

We are thankful to H. Oesterle who provided us with the ERA 40 reanalysis data from the European Center for Medium-Range Weather Forcasting. We are also indebted to A. Majda for valuable hints and suggestions concerning the topic of the manuscript. Thanks also to Vladimir Petukhov for constant encouragement and for providing meteorological background information.

References

[1] P. Metzner, Ch. Schuette, and E. Vanden-Eijnden. Illustration of transition path theory on a collection of simple examples. *submitted to J. Chem. Phys.*, 2006. (available via biocomputing.mi.fu-berlin.de).

[2] U. Achatz and G. Branstator. A two-layer model with empirical linear corrections and reduced order for studies of internal climate variability. *J. Atmos. Sci.*, 56:3140–3160, 1999.

[3] I.T. Jolliffe. *Principal Component Analysis.* Springer, 2002.

[4] I.T. Jolliffe. A cautionary note of artificial examples of eofs. *J. Climate*, 16:1084–1086, 2003.

[5] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Annals of Statistics*, 29(6):1537–1566, 2001.

[6] E. Diederichs, C. Vial, A. Juditsky, J. Polzehl, V. Spokoiny, and Ch. Schuette. Sparse non-gaussian component analysis. *manuscript in preparation*, 2006. (Preprint to appear October 2006).

[7] T. DelSole. Optimally persistent patterns in time-varying fields. *J. Atmos. Sci.*, 58:1341–1356, 2001.

[8] D. Crommelin and A. Majda. Strategies for model reduction: comparing different optimal bases. *J. Atmos. Sci.*, 61:2206–2217, 2004.

[9] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[10] A.H. Monahan. Nonlinear principal component analysis by neural networks: Theory and application to the lorenz system. *J. Climate*, 13:821–835, 2000.

[11] B. Christiansen. The shortcomings of NLPCA in identifying circulation regimes. *J. Climate*, 18:4814–4823, 2005.

[12] P. Zhang, Y. Huang, S. Shekhar, and V. Kumar. Correlation analysis of spatial time series datasets: A filter-and-refine approach. *the Proc. of the 7th Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD 2003)*, 2003.

[13] K. Chakrabarti and S. Mehrotra. Local dimensionality reduction: A new approach to indexing high dimensional spaces. In *Proceedings of the 26th VLDB Conference*, pages 98–115. Cairo,Egypt, 2000.

[14] C. Agarrval, J. Wolf, P. Yu, C. Procopiuc, and J. Park. Fast algorithms for projected clustering. *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 1999.

[15] V. de Silva, J.B. Tenenbaum, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

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

[17] C.R. Nelson and C.J. Kim. *State–space models with regime switching: classical and Gibbs–sampling approaches.* MIT Press, 1999.

[18] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Informat. Theory*, 13:260–269, 1967.

[19] L.E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[20] G. McLachlan and D. Peel. *Finite mixture models.* Wiley, New–York, 2000.

[21] I. Horenko, J. Schmidt-Ehrenberg, and Ch. Schuette. Set-oriented dimension reduction: Localizing principal component analysis via hidden markov models. In *LNBI: Proceedings of the 2nd International Symposium on Computational Life Science*, volume 4216, pages 74–85, 2006.

[22] J.A. Bilmes. *A Gentle Tutorial of the EM Algorithm and its Applications to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Thechnical Report.* International Computer Science Institute, Berkeley, 1998.

[23] G. Box and G. Jenkins. *Time Series Analysis, Forecasting, and Control.* Holden–Day, 1976.

[24] S.M. Kay. Vector space solution to the multi dimensional yule-walker equations. In *IEEE International Conference on Acoustics, Speech and Signal Proceedings.*, volume 3, pages 289–292, 2003.

[25] P. Premakanthan and W.B. Mikhael. Multidimensional model based speech signal representations for automatic speaker identification. In *Proceedings of Circuits, Signals, and Systems*, 2004.

[26] S. Papadimitriou, A. Bockwell, and C. Faloutsos. Adaptive, unsupervised stream mining. *VLDB J.*, 13(3):222–239, 2004.

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

[28] G. Branstator and S.E. Haupt. An empirical model of barotropic atmospheric dynamics and its response to tropical forcing. *J. Climate*, 11:2645, 1998.

[29] A. Stuart and P. Wiberg. Parameter estimation for partially observed hypoelliptic diffusion. 2004. (available via www.Maths.Warwick.ac.uk/ stuart).

[30] U. Achatz and J.D. Opsteegh. Principal interaction patterns in baroclinic wave life cycles. *J. Atmos. Sci.*, 52:3201–3213, 1995.

[31] F. Kwasniok. The reduction of complex dynamical systems using principal interaction and oscillation patterns. *Physica D*, 92:28–60, 1996.

[32] F. Kwasniok. Empirical low-order models of barotropic flow. *J. Atmos. Sci.*, 61:235–245, 2004.

[33] I. Horenko, E. Dittmer, A. Fischer, and Ch. Schuette. Automated model reduction for complex systems exhibiting metastability. *to appear in Mult. Mod. Sim.*, 2006. (available via biocomputing.mi.fu-berlin.de).

[34] I. Horenko, E. Dittmer, and Ch. Schuette. Reduced stochastic models for complex molecular systems. *Comp. Vis. Sci.*, 9(2):89–102, 2006.

[35] I. Horenko and C. Schuette. Likelihood-based estimation of Langevin models and its application to biomolecular dynamics. *submitted to MMS*, 2006. (available via biocomputing.mi.fu-berlin.de).

[36] A. Simmons and J. Gibson. The ERA 40 project plan. In *ERA 40 Project Rep. Ser. 1*, 2000. European Center for Medium-Range Weather Forcasting, Reading.

[37] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting.* Springer, Berlin, 2002.

[38] I. Horenko and C. Hartmann. Blind model reduction for high-dimensional time-dependant data. *submitted to Physica D*, 2005. (available via biocomputing.mi.fu-berlin.de).

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39(1):1–38, 1977.

[40] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41(1):164–171, 1970.

[41] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, volume 35 of *Computational Mathematics.* Springer, Heidelberg, 2004.

[42] R.J. Reed. The structure and dynamics of the 26-month oscillation. In *Proc. Intern. Symp. "Dynamics of large-scale processes in the atmosphere", Moscow*, pages 376–387, 1967.

[43] R.S. Lindzen and J.R. Holton. A theory of the quasi-biennial oscillation. *J. Atmos. Sci.*, 25:1095–1107, 1968.

[44] V.K. Petokhov. Two mechanisms of temperature oscillations in a thermodynamical model of the troposphere-stratosphere system. *Izvestiya, Atmos. Ocean. Phys.*, 18(2):126–136, 1982.

[45] J.W. Hurrell. Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science*, 269:676–679, 1995.

[46] J.W. Hurrell, M.P. Hoerling, A.S. Phillips, and T. Xu. Twentieth century north atlantic climate change. part 1: assessing determinism. *Clim. Dyn.*, 23:371–389, 1995.

[47] I.N. James. Some aspects of the global circulation of the atmosphere in january and july 1980. In *Large-scale dynamical processes in the atmosphere. Hoskins B, Pearce R (edts.), Academic Press, New York*, pages 5–26, 2001.