On identification of nonstationary factor models and its application to atmospherical data analysis *

Illia Horenko[†]

Institute of Mathematics

Free University of Berlin

Arnimallee 6, 14195 Berlin, Germany

November 23, 2009

[*] Supported by the DFG SPP 1276 METSTROEM, DFG research center MATHEON "Mathematics for key technologies" in Berlin and Center of Scientific Simulations (FU Berlin).

[†] horenko@math.fu-berlin.de

Abstract

A numerical framework for data-based identification of nonstationary linear factor models is presented. The approach is based on the extension of the recently developed method for identification of persistent dynamical phases in multidimensional time series allowing to identify discontinuous temporal changes in underlying model parameters. Finite element method (FEM) discretization of the resulting variational functional is applied to reduce the dimensionality of the resulting problem and to construct the numerical iterative algorithm. Presented method results in the sparse sequential linear minimization problem with linear constrains. Performance of the framework is demonstrated on two application examples: (i) in context of subgrid scale parameterization for Lorenz96 model with external forcing and (ii) in analysis of climate impact factors acting on the blocking events in the upper troposphere. The importance of accounting for the nonstationarity issue is demonstrated in the second application example: modelling the ERA40 geopotential time series via a single best stochastic model with time-independent coefficients results in the fact that all of the considered external factors are found to be statistically

insignificant, whereas considering the non-stationary model (being also demonstrated to be more appropriate in the sense of information theory) identified by the methodology presented in the paper results in identification of statistically significant external impact factor influences.

Introduction

The parameterization of reduced dynamical models describing the behavior of the observed (measured) multiscale data was a field of intensive research in the last years. Approaches introduced in atmospherical science range from stochastic differential equations (Majda et al. 1999, 2003; Wilks 2005), multiscale schemes (Majda et al. 2003; Fatkullin and Vanden-Eijnden 2004), regression models (Orrell 2003), discrete Markov chain models (Khouider et al. 2003), hidden Markov models (Majda et al. 2006; Horenko et al. 2008b,a) and conditional Markov models (Crommelin and Vanden-Eijnden 2008). In the present paper a purely data-driven approach for parameterization by means of the nonstationary multivariate autoregressive factor models (VARX) is introduced, based on the combination of the stationary VARX models widely used in econometrics (Tsay 2005) with the recently introduced FEM-clustering procedure (Horenko 2009b,c). Resulting numerical strategy is demonstrated to allow the multiscale approximation of the nonstationary dynamical processes via the optimal sequences of locally stationary fast VARX processes and some slow (or *persistent*) *hidden process* switching between them. In atmospherical context it was recently demonstrated that the FEM-clustering framework can be successfully applied to identify the large scale dynamical circulation patterns in realistic AGCM models (Franzke et al. 2009).

In the current manuscript the FEM-clustering framework is extended to allow for discontinuous *hidden processes* via the formulation of the respective variational problem in the space of the functions with bounded variation. Applications of the proposed method are demonstrated in two different scenarios and compared with standard purely data-driven methods (Orrell 2003; Wilks 2005): (i) in context of the subgrid scale parameterization for a Lorenz'96 model (Lorenz 1996) with nonstationary forcing in the right hand side and (ii) in context of climate impact factor analysis for ERA40 historical geopotential data in Europe between 1958 and 2003 (Simmons and Gibson 2000).

The outline of the remainder of this paper is as follows. In section 1 the inverse problem for nonstationary dynamical systems is formulated as a clustering problem and it is demonstrated how the multiscale assumption can be incorporated into the resulting variational formulation via the persistency condition in the space of functions with bounded variation. The finite element method (FEM) is deployed to reduce the dimensionality of the resulting optimization problem and the iterative numerical scheme is introduced. In section 2 the numerical details of the resulting FEM-VARX clustering method are explained. In section 3 different strategies of postprocessing the clustering results are discussed wrt. their insight into the analyzed data. In section 4 the performance of the presented framework is demonstrated on two practical applications and the discussion is presented in section 5.

1. Constrained Clustering Method

a. *Model distance functional*

Let $x_0, \ldots, x_T \in \Psi \subset \mathbf{R}^n$ be the observed n-dimensional time series with $T+1$ snapshots in time interval $[0,T]$. In the following we will assume that in this time interval considered time series x_t is approximated by a time-discrete output of the certain *direct mathematical model*

$$\mathbf{F}\left(x_t, \ldots, x_{t-m\tau}, \theta(t), t\right) = 0, \tag{1}$$

where $\mathbf{F}\left(\cdot\right)$ is the model operator, τ is the model time step, $m\tau$ is the memory depth ($m = 1$ for Markov models) and

$$\theta(t) : [0,T] \rightarrow \Omega \subset \mathbf{R}^d, \tag{2}$$

is a (time-dependent) set of the model parameters (including, if necessary in some model contexts also some initial and/or boundary values) and d is the dimension of a model parameter space. Let

$$g\left(x_t, \theta(t)\right) \quad : \quad \Psi \times \Omega \rightarrow [0, \infty)\,, \tag{3}$$

be a functional (further called *model distance functional*) describing the *distance* between some given x_t at time t and the output of the model (1) calculated for a fixed set of parameters $\theta(t)$. In this case, for a given observation series $x_0, \ldots, x_T$ and some fixed functional form $g\left(\cdot\right)$, the *inverse problem* (or the *parameter identification problem*) can be approached via the solution of the following variational problem:

$$\sum_{t=1}^{T} g\left(x_t, \theta(t)\right) \rightarrow \min_{\theta(t)}, \tag{4}$$

subjected to the constraints (2). Problem (4) is clearly ill-posed if no special assumptions about the temporal dependence of the unknown parameters $\theta(t)$ can be made. In the following we will assume that for any $t \in [0, T]$ model distance functional (3) can be represented as a *convex linear combination* of $K \geq 1$ *stationary* model distance functionals, i. e., model functionals dependent on some constant (time-independent) model parameters $\theta_i \in \Omega, i = 1, \dots, K$:

$$g\left(x_t, \theta(t)\right) = \sum_{i=1}^{K} \gamma_i(t) g\left(x_t, \theta_i\right), \tag{5}$$

with some time-dependent *model affiliations* $\gamma_i\left(t\right)$ fulfilling the convexity condition

$$\begin{aligned} \sum_{i=1}^{K} \gamma_i(t) &= 1, \quad \forall t \in [0, T] & (6) \\ \gamma_i(t) &\geq 0, \quad \forall t \in [0, T], \quad i = 1, \dots, K. & (7) \end{aligned}$$

In another words, we assume here that at any time t the global time-dependent (or nonstationary) model distance functional (3) can be approximated by one of K local time-independent (or stationary) model distance functionals chosen according to some time-dependent probabilities (or model affiliations) $\Gamma(t) = (\gamma_1(t), \dots, \gamma_K(t))$. This idea for the inverse numerical problems is widely used in the context of data clustering (Höppner et al. 1999), in presented general form was introduced in Horenko (2009a) and stems from the classical spline interpolation approach for direct numerical problems (see, for example, Deuflhard (2004)). Inserting the ansatz (5) in (4) results in the minimization of the *average clustering* *functional*

$$\mathbf{L}\left(\Theta,\Gamma(t)\right) \;=\; \sum_{i=1}^{K}\sum_{t=0}^{T}\gamma_i(t)g\left(x_t,\theta_i\right) \rightarrow \min_{\Gamma(t),\Theta}, \tag{8}$$

subject to (2,6,7) with $\Theta = (\theta_1,\ldots,\theta_K)$.

In order to comprehend the above concepts it is instructive to consider a case where the direct model (1) has a following simple form:

$$x_t \;=\; \theta(t) + \epsilon_t, \tag{9}$$

where ϵ_t is some *independent identically distributed* (i.i.d.) stochastic variable with zero expectation $\mathbb{E}\left[\epsilon_t\right] = 0$ and $\theta(t) : [0,T] \rightarrow \mathbf{R}^n$ is a time-dependent parameter describing the evolution of the expectation value of the process x_t. The *model distance functional* (3) in such a case gets the form

$$g\left(x_t,\theta(t)\right) \;=\; \|x_t - \theta_t\|, \tag{10}$$

and the corresponding *average clustering functional* can be numerically minimized applying the standard *K-means-clustering* algorithm (Bezdek 1981; Höppner et al. 1999). This means that with the help of the ansatz (5), the solution of the nonstationary inverse problem for dynamical system (9) can be approached via the iterative clustering algorithm based on the minimization of the *average clustering functional* (8).

In the following it will be explained in detail how these concepts can be interpreted and applied in context of more general nonstationary multivariate models with external forcing (VARX models).

b. *Non-stationary VARX models and VARX model distance functional*

Stationary VARX model is a widely used dynamical multivariate tool to investigate the time series subject to external forcing (Brockwell and Davis 2002; Tsay 2005). If, in addition to the time series $x_0, \ldots, x_T \in \mathbf{R}^n$ (describing the *internal* degrees of freedom of the considered dynamical system), the time series $u_0, \ldots, u_T \in \mathbf{R}^l$ of the *external* influences (or *forcing*) is available, the non-stationary non-linear VARX model has the following form:

$$x_t \;=\; \mu(t) + \mathbf{A}\left(t\right)\phi_1\left(x_{t-\tau}, \ldots, x_{t-m\tau}\right) + \mathbf{B}\left(t\right)\phi_2\left(u_t\right) + \mathbf{C}\left(t\right)\epsilon_t, \quad (11)$$

where $\phi_1\left(x_{t-\tau}, \ldots, x_{t-m\tau}\right)$ is some (in general non-linear function connecting the previous observations $x_{t-\tau}, \ldots, x_{t-m\tau}$), $\phi_2\left(u\right) = \left(\phi_2^1\left(u(t)\right), \ldots, \phi_2^k\left(u(t)\right)\right)$:

$\mathbf{R}^{nm} \rightarrow \mathbf{R}^{d}$ is some fixed (nonlinear) function of the external factors, $\epsilon_t : [0,T] \rightarrow \mathbf{R}^{h}$ $(h \ll n)$ is a Gaussian process with zero expectation and $\mathbb{E}\left[{\epsilon_t}^{\mathsf{T}}\epsilon_t\right] = Id^{h\times h}$; $\mu(t) : [0,T] \rightarrow \mathbf{R}^{n}$, $\mathbf{A}\left(t\right) : [0,T] \rightarrow \mathbf{R}^{n\times d}$, $\mathbf{B}\left(t\right) : [0,T] \rightarrow \mathbf{R}^{n\times k}$ and $\mathbf{C}\left(t\right) : [0,T] \rightarrow \mathbf{R}^{n\times h}$.

The most simple and straightforward form of the VARX-model used in the literature is the *linear autoregressive factor model* with $\phi_1\left(x_{t-\tau},\ldots,x_{t-m\tau}\right) = \left[x_{t-\tau},\ldots,x_{t-m\tau}\right]$ and a VARX model equation (Brockwell and Davis 2002)

$$x_t \;=\; \mu(t) + \sum_{q=1}^{m}\mathbf{A}_q\left(t\right)x_{t-q\tau} + \mathbf{B}\left(t\right)\phi_2\left(u(t)\right) + \mathbf{C}\left(t\right)\epsilon_t, \qquad (12)$$

We will further assume that the *noise matrix* $\mathbf{C}\left(t\right)$ (describing the coupling between the h-dimensional Gaussian noise process to the analyzed time series x_t) for any $t \in [0,T]$ can be represented as

$$\mathbf{C}\left(t\right) \;=\; \mathbf{P}\left(t\right)\mathbf{\Lambda}\left(t\right), \qquad (13)$$

where $\mathbf{P}\left(t\right) : [0,T] \rightarrow \mathbf{R}^{n\times h}$ is an orthogonal matrix function and $\mathbf{\Lambda}\left(t\right) : [0,T] \rightarrow \mathbf{R}^{h\times h}$ is a diagonal matrix function with nonnegative diagonal elements. Defining $\theta\left(t\right) = \left(\mu\left(t\right),\mathbf{A}\left(t\right),\mathbf{B}\left(t\right),\mathbf{C}\left(t\right)\right)$ and under the assumption (13), the

VARX *model distance functional* (3) of dynamical system (11) can be written as

$$g\left(x_t,\theta(t)\right) \;=\; \|x_t-\mu(t)-\mathbf{A}\left(t\right)\phi_1\left(x_{t-\tau},\ldots,x_{t-m\tau}\right)-\mathbf{B}\left(t\right)\phi_2\left(u(t)\right)\|_{\mathbf{P}(t)}, \tag{14}$$

where the $\mathbf{P}\left(t\right)$-weighted norm $\|\cdot\|_{\mathbf{P}(t)}=\sqrt{\left(\cdot\mathbf{P}^{\dagger}\left(t\right),\mathbf{P}\left(t\right)\cdot\right)_2}$ is used and † denotes the matrix transposition. The main advantage of the above definition of the *least squares residual norm* (14) (compared to the standard Gaussian norm based on the $\mathbf{C}\left(t\right)$-weighted norm $\|\cdot\|_{\mathbf{C}(t)}=\sqrt{\left(\cdot,\mathbf{C}\left(t\right)\cdot\right)_2}$ is that it preserves the norm of the original residuals of the least-squared problem in the *essential noise dimension*.) If the aforementioned assumptions are fulfilled, then it is easy to demonstrate that the time series of model distances $g\left(x_t,\theta(t)\right), t=0,\ldots,T$ is a χ^2-process. In the application examples below it will be demonstrated how this property of the process can be used to estimate the confidence intervals of the model parameters. In context of stationary VARX models (e. g., in the case of the time independent parameter matrices $\mathbf{A},\mathbf{B}$ and $\mathbf{C}$) this property can be deployed to *a posteriori* check the model assumption (11) and to demonstrate the *asymptotical normality* of the resulting parameter estimates (i. e., to demonstrate that for $T\rightarrow\infty$ the parameter estimates are distributed according to the multivariate

Gaussian) (Reinsel 1993). Applying the ansatz (5), the nonstationary parameter identification problem can be approached via the solution of the minimization problem (8) (with $\theta_i = (\mu^i, \mathbf{A}^i, \mathbf{B}^i, \mathbf{C}^i)$). This can be done by means of the iterative clustering algorithm (in the same way as it was described for the K-Means clustering) (Höppner et al. 1999; Horenko 2009b).

c. *Incorporation of additional information*

Direct numerical treatment of the problem (8) is hampered by the following problems: (i) the problem is *ill-posed* since the number of unknowns can be higher then the number of known parameters, and (ii) because of the non-linearity of g the problem is in general *non-convex* and the numerical solution gained with some sort of *local minimization algorithm* depends on the initial parameter values (Deuflhard 2004). Perhaps it is even more important that the solution Γ of the above constrained minimization task might be an irregular function: To see this let us assume that we already know the minimizing values Θ_* for Θ. Then, the minimizer Γ_* for the affiliation vector Γ has the following form:

$$\gamma_{*,i}(t) = \begin{cases} 1 & \text{if } i = \arg\min_j g(x_t, \theta_{*,j}) \\ 0 & \text{otherwise} \end{cases}, \tag{15}$$

thus the datum x_t has perfect affiliation with state i if the model distance functional for x_t is minimal in state i. That is, if the process exhibits strong variability then the affiliations are rather non-smooth functions. Whenever the affiliation functions just take values 0 or 1 we will call them *deterministic* in the following, which is meant in the sense that then for every datum in the time series it is certain to which cluster it belongs.

As was demonstrated in the literature (Horenko 2009b,c), one of the possibilities to approach the two aforementioned problems simultaneously is first to incorporate some *additional information* about the regularity of the observed process (e.g., in the form of *smoothness assumptions* in space of time-continuous functions $\Gamma(\cdot)$) and then to apply a finite Galerkin-discretization (e. g, FEM-discretizatization) of this infinite-dimensional Hilbert space. In context of Tykhonov-based FEM-clustering methods, this was done assuming the *weak differentiability* of the time-continuous functions γ_i, i. e.:

$$|\gamma_i|_{W_{1,2}(0,T)} \quad = \quad \| \, \partial_t \gamma_i \left(\cdot \right) \|_{\mathcal{L}_2(0,T)} = \int_0^T \left(\partial_t \gamma_i \left(t \right) \right)^2 \delta t \leq C < +\infty, \quad i = 1, \ldots, K, \tag{16}$$

where $W_{1,2}\left(0, T\right)$ is the Sobolev space of weakly differentiable functions, e. g., functions with the $\mathcal{L}_2\left(0,T\right)$ integrable first derivatives.

As was demonstrated in (Horenko 2009a), one possibility to incorporate this *a priori information* from (16) into the optimization is to modify the functional (8) and to write it in the *Tykhonov-regularized* form

$$\mathbf{L}^{\epsilon}(\Theta,\Gamma,\epsilon^2) \;\; = \;\; \mathbf{L}(\Theta,\Gamma)+\epsilon^2\sum_{i=1}^{K}\int_0^T\left(\partial_t\gamma_i\left(t\right)\right)^2\delta t \rightarrow \min_{\Gamma\in\mathcal{W}_{1,2}(0,T),\Theta}. \quad (17)$$

However, introduction of the ϵ^2-dependent penalty term in the formulation of the Tykhonov-regularized problem (17) changes the functional form of the original clustering problem (8) and biases the position of the solution of the respective minimization problem with growing ϵ^2, e. g., the solution of the regularized problem may have a significant deviation from the global minimum of the original problem. As was demonstrated in ((Horenko 2009b,c)), increasing the ϵ^2 leads to the "smoothing out" of the sharp transitions between the cluster states[1]. Moreover, the formulation (16) of the persistency condition in $W_{1,2}$ sense relies on the differentiability and *continuity* of the underlying *cluster affiliation functions* $\gamma_i(t)$. This can not in general be assumed to be granted in the cases where the transitions

[1] The influence of the Tykhonov regularization in $W_{1,2}$ Sobolev norm is equivalent to the action of the diffusion operator $\epsilon^2\Delta_t$ on $\gamma_i(t)$, e. g., increase of the regularization parameter ϵ^2 results in the stronger diffusion of the affiliation function $\gamma_i(t)$ through the interface separating the clusters in time. For a detailed discussion of the $W_{1,2}$ regularization effects and its influence in different clustering scenarios see Horenko (2009b,c).

between the cluster states are sharp (e. g., when the function $\gamma_i(t)$ has jumps and is discontinuous).

In the following, an alternative way to incorporate the persistency assumption into the original clustering problem (8) (avoiding the two above mentioned problems) will be presented, based on the formulation of the persistency condition in a functional space allowing for a discontinuity of its elements (functions with bounded variation, $BV\,(0,T)$).

d. *Persistence in* $BV\,(0,T)$*-sence: constrained* BV*-clustering method and FEM-discretization*

Instead of limiting $\partial_t \gamma_i$ in $\mathcal{L}^2$ sense (which relies on the differentiability and *continuity* assumption for the cluster affiliation function $\gamma_i(t)$), we will consider the time-discrete functions $\gamma_i(t)$ defined only at the time instances where the observations x_t are available. We will formulate the persistency condition in the time-discrete $BV\,(0,T)$ sense

$$|\gamma_i|_{BV(0,T)} \;=\; \sum_{t=0}^{T-1} |\gamma_i(t+1) - \gamma_i(t)| \leq C, \tag{18}$$

where the persistency parameter C defines the maximal number of transition between the cluster state i and all other states in time interval $(0, T)$. Note that since at least in the time-continuous case $\mathcal{W}_{1,2}(0, T) \in BV(0, T)$ (Moreau 1988), this kind of persistency condition will also allow to preserve the "smooth" $\mathcal{W}_{1,2}$-transitions between the cluster states in the continuous limit by an appropriate choice of the persistency parameter C.

Let D be the discrete difference operator (the right-hand derivatives) wrt. t

$$D = \begin{bmatrix} -1 & 1 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & -1 & 1 \end{bmatrix}. \qquad (19)$$

For a given $\Theta = (\theta_1, \dots, \theta_K)$ let

$$\begin{aligned} g_i(\theta_i) &= [g(x_0, \theta_i), \dots, g(x_T, \theta_i)] \in \mathbf{R}^{T+1}, \\ \gamma_i &= [\gamma_i(0), \dots, \gamma_i(T)] \in \mathbf{R}^{T+1} \end{aligned} \qquad (20)$$

Then the problem (8) transforms to

$$\mathbf{L} \;=\; \sum_{i=1}^{K} \gamma_i g_i^{\dagger}\left(\theta_i\right) \rightarrow \min_{\Gamma(t),\Theta}, \qquad (21)$$

where † denotes the transposition operation. The above problem (21) is subject to the constraints

$$\begin{aligned} \|D\gamma_i^{\dagger}\|_1 &\leq C \quad \forall i, & (22)\\ \sum_{i=1}^{K}\gamma_i(t) &= 1 \quad \forall t, & (23)\\ \gamma_i(t) &\geq 0 \quad \forall t, i. & (24) \end{aligned}$$

The direct numerical optimization of the problem (21-24) is hampered by the fact that the persistency constrain (22) makes the overall problem *non-differentiable*. In the following, an adequate transformation of the above problem to the *differentiable* formulation will be introduce . This will allow us to use the standard numerical optimization methods in the context of (21-24). We define $v_i := D\gamma_i^T$ and split this into non-positive and non-negative parts (following the theorem about the unique representation of the BV-functions, cf. Moreau (1988))

$$v_i \;=\; v_i^+ - v_i^-, \quad v_i^+ = \max(v_i, 0), v_i^- = \max(-v_i, 0). \qquad (25)$$

Moreover let D^{-1} be the discrete integration operator, e. g.

$$D^{-1}v_i = \underbrace{\begin{bmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 1 & \dots & 1 & 1 \end{bmatrix}}_{=:\mathcal{D}^{-1}} v_i + \gamma_{00}^i \tag{26}$$

for some variable γ_{00}^i (defining the initial value for the cluster affiliation i at time 0). Now we can express γ_i^T as

$$\begin{aligned} \gamma_i^T &= D^{-1}(v_i^+ - v_i^-) \\ &= \mathcal{D}^{-1}(v_i^+ - v_i^-) + \gamma_{00}^i \mathbb{1}. \end{aligned} \tag{27}$$

Defining

$$\tilde{x}_i = \begin{bmatrix} v_i^+ \\ v_i^- \\ \gamma_{00}^i \end{bmatrix}, \tilde{x} = [\tilde{x}_1, \dots, \tilde{x}_K], c_i(\theta_i) = \begin{bmatrix} (g_i(\theta_i)\mathcal{D}^{-1})^T \\ -(g_i(\theta_i)\mathcal{D}^{-1})^T \\ g_i(\theta_i)\mathbb{1} \end{bmatrix} \tag{28}$$

we can express the original *non-differentiable* optimization problem (21-24) as the following *differentiable* optimization problem, (for more details see the Appendix)

$$\min_{\tilde{x},\Theta} c^T(\Theta)\,\tilde{x}, \quad \text{subject to } A_{\text{eq}}\tilde{x} = b_{\text{eq}}, A_{\text{neq}}\tilde{x} \geq b_{\text{neq}}, \tag{29}$$

in the vector space of the *higher dimensionality* (since by construction, the dimension of variable $\tilde{x}_i$ defined in (28) is almost twice as high as the dimension of the original variable γ_i). The solution of the above minimization problem can be approached via the *subspace iteration* procedure, e. g., via the solution of the restrained optimization problems in parameter subspaces $\tilde{x}$ and θ subsequently. Completely analogously to the Tykhonov-regularized FEM-clustering case (Horenko 2009b,c), it can be demonstrated that this *subspace iteration* procedure converges towards the (local) minimum of the problem (29) if some appropriate assumptions (convexity and differentiability) of the model distance functional (3) are fulfilled.

However, since the dimensionality of the variable $\tilde{x}$ is growing as $K\,(2T+1)$ with the length of the analyzed time series, the numerical solution of the restrained problem (29) for a fixed value of Θ can become increasingly expensive for long time series. In the following the Finite Element Method (FEM) will be deployed to reduce the dimensionality of the above problem.

e. *FEM-discretization*

Let $\{0 = \tau_0, \tau_1, \tau_2, \ldots, \tau_{N-1}, \tau_N, \tau_{N+1} = T\}$ be a finite subset of the interval $[0, T]$ with uniform time steps δ_t. We can define a set of $N \ll (T+1)$ time-discrete functions with bounded variation $\{f_1(t), f_2(t), \ldots, f_N(t)\}$ defined at $T+1$ time points of the observation series x, where each function $f_i(t)$ shall take positive values at the observation time instances of x in time interval (τ_{i-1}, τ_{i+1}) and be zero at the time instances outside this interval[2]. Now we represent the v's from (25) by these functions, thus

$$\begin{aligned} v_i^+(t) &= \sum_{k=1}^{N} \tilde{v}_{ik}^+ f_k(t) + \chi_N^+ \\ v_i^-(t) &= \sum_{k=1}^{N} \tilde{v}_{ik}^- f_k(t) + \chi_N^- \end{aligned} \tag{30}$$

where χ_N^+ and χ_N^- are discrete discretization errors. As δ_t goes to 1, the discretization errors become zero. This develops into a reduced discrete representation of

[2] For practical examples of standard *finite element function sets* $f_i(t)$ (like linear finite elements) in discrete time see Horenko (2009b) and Braess (2007)

the discrete problem (29) by using

$$v_i^+ = \underbrace{\begin{bmatrix} f_1(0) & \dots & f_N(0) \\ \vdots & \ddots & \vdots \\ f_1(T) & \dots & f_N(T) \end{bmatrix}}_{=:W_N} \underbrace{\begin{bmatrix} \tilde{v}_{i1}^+ \\ \vdots \\ \tilde{v}_{iN}^+ \end{bmatrix}}_{=:\tilde{v}_i^+} + \chi_N^+ \mathbb{1}, \tag{31}$$

where $W_N \in \mathbf{R}^{(T+1)\times N}$ is a FEM-basis matrix and $\tilde{v}_i^-$ is defined analogously. Then the previously defined $\tilde{x}_i$ can be approximated as

$$\tilde{x}_i = \underbrace{\begin{bmatrix} W_N & 0 & 0 \\ 0 & W_N & 0 \\ 0 & 0 & \mathrm{Id} \end{bmatrix}}_{=:\hat{\omega}} \underbrace{\begin{bmatrix} \tilde{v}_i^+ \\ \tilde{v}_i^- \\ \gamma_{00}^i \end{bmatrix}}_{=:\bar{x}_i}. \tag{32}$$

Defining

$$\omega = \left.\begin{bmatrix} \hat{\omega} & 0 & \dots & 0 \\ 0 & \hat{\omega} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \hat{\omega} \end{bmatrix}\right\} (K \text{ times}), \tag{33}$$

results in the *FEM discretized problem in BV sense*:

$$\min_{\bar{x},\Theta} \mathbf{L}_0 = \min_{\bar{x},\Theta} c^T(\Theta)\,\omega\bar{x}, \quad \text{subject to } A_{\text{eq}}\omega\bar{x} = b_{\text{eq}}, A_{\text{neq}}\omega\bar{x} \geq b_{\text{neq}}. \qquad (34)$$

Note that $\bar{x}$ has a dimensionality of $K(2N+1)$ (which is much less then the dimension $K(2T+1)$ of the original variable $\tilde{x}$ if $N \ll T$). Analogously to the Tykhonov-regularized FEM-clustering problem described in (Horenko 2009c), *adaptive FEM techniques* can be deployed to find the *optimal set of time intervals* δ_t for a given total discretization error $\tilde{\delta}$[3].

For a *fixed* Θ the above minimization problem is a linear minimization wrt. $\bar{x}$ with linear equality and inequality constraints. This problem can be solved by means of some standard numerical methods of linear programming like simplex method or interior point methods. For the fixed parameter $\bar{x}$, the above problem is an unconstrained minimization problem that can be solved *analytically* if the model distance functional g is convex wrt. θ and the problem $\frac{\partial g(x_t,\theta)}{\theta} = 0$ has a unique analytical solution wrt. θ (as would be demonstrated in the following, it is the case for the VARX models). The iterations can be repeated until the change of the functional value doesn't exceed some predefined *threshold for the change of*

[3] This can be done in a standard way, controlling the norm of the disretization errors χ_N^+, χ_N^- locally and applying the multigrid approach to guarantee that $\|\chi_N^{\pm}\| \leq \tilde{\delta}$ (Braess 2007)

the functional value.

```
Algorithm:
choose an arbitrary $\bar{x}^0$
set s=1
while not converged repeat
    step 1: solve (34) wrt. $\Theta$ for a fixed $\bar{x}^{s-1}$
    analytically and identify $\Theta^s$
    step 2: solve (34) wrt. $\bar{x}$ for a fixed $\Theta^s$ numerically
    (via linear programming) and identify $\bar{x}^s$
    set s=s+1
```

Compared to the Tykhonov-regularized FEM-clustering problem as described in Horenko (2009b,c), the above problem (34) has two major numerical advantages: (i) while for the Tykhonov-regularized FEM-clustering problem the persistence could be influenced only indirectly through the choice of the regularization parameter ϵ^2, in context of (34) it is directly controllable via the persistency threshold C and (ii) sparse linear programming problem is solved in each iteration of the above algorithm (compared to the more numerically expensive quadratic programming in the case of the Tykhonov-regularized FEM-clustering problem).

In the next section it will be demonstrated how the above numerical procedure can be formulated for the parameter identification of nonstationary VARX models (11).

2. FEM-VARX Clustering: parameter estimation and determination of optimal values for number of states and persistency threshold

a. *Derivation of the estimator formulas*

In every iteration s of the *subspace iteration algorithm* described above, the unconstrained minimization problem

$$\Theta^s = \arg\min_{\Theta} \mathbf{L}_0\left(\Theta, \tilde{x}^{(s-1)}\right) \tag{35}$$

is solved for a fixed value of $\tilde{x}^{(s-1)}$ (step 2 of the above algorithm). If the dynamics of the analyzed time series is assumed to be governed by the VARX process (11) with memory m, the model distance functional in the clustering problem formulation will take the form (14). Let $\mathbf{AB}^i = (\mu^i, \mathbf{A}^i, \mathbf{B}^i) \in \mathbf{R}^{n\times(1+d+k)}$ (where d is the dimensionality of the output of function $\phi_1(\cdot)$) and $\Gamma^{s-1}(t)$ be the cluster affiliations reconstructed from the current values $\tilde{x}^{s-1}$ via the subsequent application of the formulas (32) and (30). After inserting the $\Gamma^{s-1}(t)$ into the function $g_i(\theta_i), i = 1, \ldots, K$ and taking the derivative of the functional $\mathbf{L}_0$ (34) wrt. the new variables $\mathbf{AB}^i$ we get the optimal estimators for $\mathbf{AB}^i_s$ in the iteration s as the solution of the following system of linear equations[4]

$$\begin{aligned}
\mathbf{X}_s\mathbf{AB}^i_s &= \mathbf{Y}_s,\\
\mathbf{X}_s &= \begin{bmatrix} 1 & \ldots & 1 \\ \sqrt{\gamma_i^{s-1}(m+1)}\phi_1^{m+2} & \ldots & \sqrt{\gamma_i^{s-1}(T)}\phi_1^{T-1} \\ \sqrt{\gamma_i^{s-1}(m+2)}\phi_2^{m+3} & \ldots & \sqrt{\gamma_i^{s-1}(T)}\phi_2^{T} \end{bmatrix},\\
\mathbf{Y}_s &= \begin{bmatrix} \sqrt{\gamma_i^{s-1}(m+2)}x_{m+2} & \sqrt{\gamma_i^{s-1}(m+3)}x_{m+3} & \ldots & \sqrt{\gamma_i^{s-1}(T)}x_T \end{bmatrix},
\end{aligned} \tag{36}$$

where $\phi_1^t = \phi_1(x_{t-\tau}, \ldots, x_{t-m\tau})$ and $\phi_2^t = \phi_2(u(t))$. If the matrix $(\mathbf{X}_s^*\mathbf{X}_s)$ is invertible then the solution of the above system exists, is unique and can be expressed as

$$\mathbf{AB}^i_s = (\mathbf{X}_s^*\mathbf{X}_s)^{-1}\mathbf{X}_s^*\mathbf{Y}_s, \tag{37}$$

[4] We get use of the convexity of the functional $g(x_t, \theta(t)) = \|x_t - \mu^i - \mathbf{A}(t)\phi_1(x_{t-\tau}, \ldots, x_{t-m\tau}) - \mathbf{B}^i\phi_2(u(t))\|_{\mathbf{P}^i}$ wrt. parameters AB^i and the necessary minimum condition for convex functionals.

where $*$ denotes the matrix conjugate (Golub and Loan 1989). The optimal estimates of the noise parameters $\mathbf{C}^i$ are straightforwardly calculated from the covariance matrices of the model residuals in the cluster state i

$$\begin{aligned}
\mathrm{res}_t^i &= \sqrt{\gamma_i^{s-1}(t)}\left(x_t - \mu^i - \mathbf{A}(t)\,\phi_1(x_{t-\tau},\ldots,x_{t-m\tau}) - B^i\phi_2(u(t))\right), \\
\mathrm{Cov}_s^i &= \frac{\sum_{t=(m+2)}^{T}\left(\mathrm{res}_t^i\right)^{\mathcal{T}}\mathrm{res}_t^i}{\sum_{t=(m+2)}^{T}\sqrt{\gamma_i^{s-1}(t)}}, \qquad (38)\\
\mathbf{C}_s^i &= \left[\mathrm{Cov}_s^i\right]^{0.5}.
\end{aligned}$$

Formulas (37) and (38) give *explicit* estimator expressions that are used in the step 2 of the *subspace iteration algorithm.*

As can be seen from the above estimator formulas, from the view point of the inverse numerical problem there is no difference between linear (12) and nonlinear (11) factor models: both result in solution of a linear least-squares problem (36). This is explained by the fact that in both cases the right-hand sides of models are *linear functions of model parameters*. However, as will be demonstrated in the following, linear autoregressive models can in some situations provide more insight allowing to use the available tools of linear data analysis and can be successfully applied to analyze the realistic (nonlinear) data.

b. *FEM-VARX model oder selection for fixed K and C*

In order to select a proper model order m and the optimal functional form $\phi_2(u) = \left(\phi_2^1(u(t)), \dots, \phi_2^k(u(t))\right)$ for the external factors, standard tools of *information theory* like *Akaike information criterion (AIC)* or *Bayesian information criterion (BIC)* (McQuarrie and Tsai 1998) developed for the linear stationary VARX models can be applied *a posteriori* to the locally stationary VARX models identified via the FEM-VARX procedure. In terms of the information criteria, various models are compared wrt. the special functional consistent of the model *log-likelihood* with added regularization term penalizing the total number of parameters involved in the model (to avoid the *overfitting*). For example, the BIC functional (being in general more robust then AIC, cf. McQuarrie and Tsai (1998)) for the cluster state i will have a form

$$\text{BIC}(i) \quad = \quad -2\log\mathcal{L}_i + N_i \log\left(\sum_{t=0}^{T}\gamma_i(t)\right), \tag{39}$$

where $\mathcal{L}_i = \gamma_i g_i^\dagger(\theta_i)$ and N_i is the number of the model parameters in the cluster state i. Given any two estimated cluster models (e. g., models with different memory depth and/or different functions $\phi_2(u)$), the model with the lower value of BIC (i) is the one to be preferred. In the following it will be demonstrated how the criterion (39) can be applied in the praxis to make a decision about the optimal functional form of the local VARX models and to test the significance of different external factors u for the dynamical process explaining the analyzed time series data.

c. *Choosing the optimal number of local models K*

The *upper bound for the number of statistically distinguishable cluster states* for each value of the persistency threshold C can be algorithmically estimated in the following way: starting with some a priori chosen (big) K one solves the optimization problem (34) for different fixed value of C and calculates the confidence intervals of the resulting local parameter matrices $\Theta^i, i = 1, \ldots, K$ (this can be done applying the standard bootstrap sampling procedures, see (Chernik 1999)). If two of the estimated parameter sets for two of the identified cluster states 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 $K = K - 1$. If at a certain point all of the matrices are *statistically distinguishable* the procedure is stopped and $K_{max}(C) = K$.

Another possibility to estimate the optimal number of clusters can be used, if the identified transition process $\Gamma(t)$ is shown to be Markovian for given K, C.

Markovianity can be verified applying some standard tests, e. g., one can check the *generator structure* of the hidden process, see (Metzner et al. 2007)[5]. In such a case the hidden transition matrix can be calculated and its spectrum can be examined for a presence of the *spectral gap* (a gap separating the fast and the slow time scales in the Markov dynamics). If the *spectral gap* is present, then the number of the dominant eigenvalues (i.e., eigenvalues between the *spectral gap* and 1.0) gives the number of the metastable clusters in the system (Schütte and Huisinga 2003). Positive verification of the hidden process' Markovianity has an additional advantage: it allows to construct a *reduced dynamical model* of the analyzed process and to estimate some dynamical characteristics of the analyzed process, e.g., one can calculate *relative statistical weights*, *mean exit times* and *mean first passage times* for the identified clusters (Horenko et al. 2008a).

Verification of the Markov-assumption also allows to construct *predictive Markovian models* of the persistent dynamical process $\Gamma(t)$ switching between the cluster states (Horenko et al. 2008a; Horenko 2009c). As it will be demonstrated for the numerical examples in the following, the respective $K \times K$ Markovian transition matrix together with the parameter estimates (37) and (38) of the locally-

[5] Note that all of the numerical criteria for verification of the Markov assumption known in the literature imply the stationarity (or homogeneity) of the underlying transition matrix and are in general not applicable if the analyzed process is known to be strongly nonstationary.

stationary cluster states can help to construct the dynamical ensemble prediction models. It is important to mention that without this *a posteriori* Markov verification, the minimum of the functional (34) can not be directly used to generate the data-based predictions since the identified cluster affiliation function Γ is just an abstract BV function with some predefined persistency C and has no intrinsic representation in terms of the underlying dynamical process.

d. *Selection of the optimal persistency threshold C for a given number of states* K

Linearity of the functional (34) wrt. $\tilde{x}$ for *fixed* values of model distance parameters θ_i can help to apply the standard instruments from the theory of ill-posed linear problems, like L-curve approach (Calvetti et al. 2000) to identify the optimal value of the persistency threshold C for a fixed number of clusters K. For example, optimal value of C can be determined as the edge-point (or the point of maximal curvature) on the two-dimensional plot. This plot depicts a dependence of the residuum-norm of the solution from C (Horenko 2009c). Alternatively, as would be demonstrated in the numerical examples, decreasing C up to some point (meaning the increasing regularity of the optimal solutions of (34) in the BV sense) results in decreasing of the respective values $\mathbf{L}_0$ of the clustering functional in the solution point (meaning the increasing quality of the resulting clustering). The point with the minimal value of $\mathbf{L}_0$ in such a case indicates the optimal persistency threshold guaranteeing the best clustering quality.

3. Postprocessing of the FEM-VARX clustering results

Application of the FEM-VARX numerical scheme with fixed values of K and C results in the identification of K optimal locally-stationary VARX models and the persistent $BV\left(0,T\right)$ function $\Gamma\left(t\right)$ switching between them. As was explained above, the *a posteriori* verification of the Markovian hypothesis for $\Gamma\left(t\right)$ allows to construct the reduced representation of the overall dynamics in the multiscale sense, e. g., as some slow persistent Markovian process switching between K locally stationary VARX parameter sets. Postprocessing of the derived local VARX models can give some additional insight into the analyzed time series. For example, expectation values of *mean dynamical equilibrium positions* $\mathbb{E}^{(i)}\left(u(t)\right) = \mathbb{E}^{i}\left[x_t|u(t)\right]$ of the analyzed dynamical process in the cluster state i can be calculated as functions of the external forcing u_t. In the case of the linear autoregressive VARX models (12), this is done via the solution of the following system of linear equations[6]

$$\mathbb{E}^{(i)}\left(u(t)\right) \;=\; \mu^i + \sum_{q=1}^{m} \mathbf{A}_q^i \mathbb{E}^{(i)}\left(u(t)\right) + \mathbf{B}^i \phi_2\left(u(t)\right) \qquad (40)$$

Note that if $\mathbf{A}_q^i = 0, \forall q$ then the above result is equivalent to the multivariate trend estimate in context of the recently introduced FEM-K-Trends clustering algorithm (Horenko 2009b).

The local linearity of the identified VARX models also allows to apply various techniques of model reduction known in the literature, e. g., *proper orthogonal decomposition* or *balanced truncation* (Moore 1981) allowing for construction of the *energy preserving low-rank approximations* to the identified VARX models.

Another kind of insight into the underlying multidimensional dynamics can be gained via the analysis of the Fourier transforms of the identified models, e. g., via the analysis of the *transfer function matrices* and *directed transfer functions* (Pereda et al. 2005). This kind of analysis can help to quantify the causal influence of different data dimensions on each other.

[6] Formula (40) is derived under the assumptions that the analyzed data x are locally weakly stationary (e. g., in this case, that the expectation value of x is (locally) time-independent) and the external noise process ϵ_t is i.i.d. and has a zero expectation.

4. Numerical examples

In the following we will illustrate the proposed FEM-VARX clustering strategy on two practical examples: (a.) on data from Lorenz'96 model (Lorenz 1996) with external periodical forcing switching between the deterministic and the chaotic regime behavior, (c.) a set of averaged daily ERA40 500 hPa geopotential data between 1958 and 2003 on a 16×9 spatial grid covering Europe and part of the north Atlantic.

a. *Subgrid scale modeling for Lorenz'96 system with forcing*

The Lorenz'96 model of type II (Lorenz 1996; Orrell 2003) is a two-scale simplified ODE describing advection, damping and forcing of some (slow) resolved atmospheric variable $\tilde{x}_i$ being coupled to some (fast) subscale variables $y_{i,j}$

$$\begin{aligned}
\dot{\tilde{x}}_i &= \tilde{x}_{i-1}\left(\tilde{x}_{i+1}-\tilde{x}_{i-2}\right)-\tilde{x}_i+F\left(t\right)-\frac{hc}{b}\mathbf{F}_i^y\left(t\right), \quad \mathbf{F}_i^y\left(t\right)=\sum_{j=1}^{\mathcal{M}}y_{i,j} \\
\tilde{x}_{i-\mathcal{N}} &= \tilde{x}_{i+\mathcal{N}}=\tilde{x}_i, \\
\dot{y}_{i,j} &= cby_{i,j+1}\left(y_{i,j-1}-y_{i,j+2}\right)-cy_{i,j}+\frac{hc}{b}\tilde{x}_i, \\
y_{i+\mathcal{N},j} &= y_{i,j}, \quad y_{i,j-\mathcal{M}}=y_{i-1,j}, \qquad (41)
\end{aligned}$$

with all of the model parameters ($\mathcal{N} = 8, \mathcal{M} = 4, h = 1, b = c = 10$) set to be the same as in the paper of Orrell (2003). In contrast to the paper of Orrell (2003), we choose the external forcing to be explicitly time dependent function of the form $F(t) = 10 \sin\left(\frac{2\pi}{80}t\right)$ resulting in the switching between the "deterministic" and the "chaotic" regime behavior (Orrell and Smith 2003) on the time scale of forcing (e. g., producing a system with three different time scales).

We generate one realization of (41) in the time interval $[0, T]$ with an adaptive Runge-Kutta method of the fourth order (MATLAB command ode45) resulting in the 40 dimensional time series with 4000 instances (time step $\tau = 0.1$). We further aim at parameterization of the subgrid scale influences $x_t^i = \mathbf{F}_i^y(t)$ in the nonstationary VARX form (11) with external factors defined by the resolved variables ($\phi_2(u(t)) = (\tilde{x}_1(t), \ldots, \tilde{x}_8(t))$). Fig. 1 demonstrates the application of the FEM-VARX to the time series of x_t and $\phi_2(u(t))$ defined in such a way for $K = 2$ (number of clusters), $q = 1$ (memory depth) and with $N = 200$ (number of finite elements) for various numbers of persistency threshold parameter. The minimum is achieved for $C = 6$, that corresponds to the number of transitions through the bifurcation point of (41). As can be seen from Fig. 2, the FEM-VARX clustering procedure for $K = 2$ and $C = 6$ identifies both dynamical regimes of the model (41), whereas the minimization of (34) without the persistency constraint (or, equivalently, $C = 2N = 400$) results in a much worse clustering result (in terms of the optimal value of $\mathbf{L}_0$) and is due to the trapping in the local minimum of the clustering functional. Inspection of the confidence intervals of the estimated local stationary VARX model parameters in Fig. 3 reveals that whereas the matrices A are statistically indistinguishable (e. g., the "own dynamics" of the subgrid scale process is basically the same in both states), there are statistically significant differences in B (in this case it governs the coupling from the resolved degrees of freedom $\tilde{x}$ to subgrid scale d.o.f.). Also the noise intensities C are different in both cases (in the identified chaotic regime the noise intensity is significantly higher).

Finally, we verify the Markov assumption for the identified affiliation function $\Gamma(t)$, generate the ensemble predictions (with 10000 ensemble members) based on the estimated 8-dimensional FEM-VARX parameterization (based on the first 90% of the data), calculate the expectation values of the relative prediction errors (in 2-norm, based on the last 10% of the data, black solid line in Fig. 4) and compare them with predictions obtained by other data-driven models (see Fig. 4): (i) stationary constant (grey dotted line) and regressive (black dash-dotted line) subgrid scale models (Orrell 2003); (ii) stationary linear stochastic model (black dashed line) (Wilks 2005); and (iii) one-dimensional FEM-VARX model, estimated under the assumption that different subgrid scale processes in different dimensions does not interact, e.g., the matrices A are diagonal (grey solid line). Fig. 4 shows that fully-interactive 8-dimensional persistent FEM-VARX model has a much better predictive skill for the considered nonstationary model series.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

b. *Analysis of ERA40 geopotential data in Europe (1958-2003)*

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

Using the FEM-VARX method introduced above, we analyze daily mean values of the 500 hPa air temperature field from the ERA 40 reanalysis data (Simmons and Gibson 2000)[7]. We consider a region with the coordinates: $27.5°$ W – $47.5°$ E and $32.5°$ N – $75.0°$ N , which includes Europe and a part of the Eastern North Atlantic. The resolution of the data is $5°$ which implies a grid with 16 points in the zonal and 9 in the meridional direction. For the analysis we have considered temperature values only for the period 01-12-1958 till 31-07-2003, thus we end with a 144-dimensional time series of 16314 days. In order to remove the seasonal trend we apply a standard procedure, where from each value in the time series we subtract a mean build over all values corresponding to the same day and month e.g., from the data on 01.01.1959 we subtract the mean value over all days which are first of January and so on. The resulting deviations $\Delta H(t)$ of the geopotential heights from their seasonal mean values are the subject of the data analysis in the following.

Since the overall dimensionality 144 of the analyzed data series will induce too high uncertainties for the given time span of 16314 days, the scope of the data analysis in the following is reduced to the projections on 20 dominant EOFs (describing approx. 99% of the total data variance).

[7] The author thanks the *Potsdam Institute for Climate Impact Research (PIK)* for the possibility to use the ERA 40 data.

Three different external influence factors u_t are tested on their significance for the change of the EOF projection data x_t in the setting of the VARX models (11): (i) atmospheric CO_2 values u_t^1 from the Mauna Loa observatory (data is available online at *http://cdiac.ornl.gov/ftp/trends/co2/maunaloa.co2*), (ii) seasonal trend factor in the form $u_t^2 = \sin\left(\frac{2\pi}{365.4}t\right)$ and (iii) the sunspot cycle data u_t^3 (data is available online at *http://solarscience.msfc.nasa.gov/SunspotCycle.html*).

We start the data analysis parameterizing the globally stationary VARX model for 20 dominant EOFs with different sets of external factors described above for various values of the memory parameter q, applying the BIC test to verify the statistical significance of the factors and testing the i.i.d. assumptions for the model residuals with the AR(1)-test (Brockwell and Davis 2002) (to check the validity of the underlying estimator assumptions). It shows up that the optimal globally stationary VARX model for the analyzed data is the one with $q = 2$ and *no external factors*, e. g., in the globally stationary framework all of the external factors are found to be statistically nonsignificant.

Application of the nonstationary FEM-VARX framework according to the K-selection procedure described above results in determining $K = 4$ as the maximal number of statistically distinguishable local VARX states (valid for a wide range of parameters C, N, q). For a fixed value of $K = 4$ we further determine the optimal value of the persistency threshold C (in the same manner as was demonstrated in Fig. 1), determine the optimal value of memory $q = 2$ via BIC-test (e. g., the same as for the globally stationary VARX model) and test the i.i.d. assumptions for the model residuals with the AR(1)-test (Brockwell and Davis 2002). Next, we repeat the verification of external factors in the same way as it was described above for the globally stationary VARX model. As shown in Fig. 5, the BIC-test confirms the hypothesis that the optimal FEM-VARX model is the one with the two external factors u_t^1 (atmospheric CO_2 values) and u_t^2 (seasonal factor), whereas the influence of the the sunspot cycle data u_t^3 is found to be statistically negligible. Therefore, all of the further tests are conducted only with factors u_t^1 and u_t^2.

Inspection of the respective cluster affiliation functions $\gamma_i\left(t\right)$ (see Fig. 6) together with the mean equilibrium positions in the cluster states (40) reveals (see Fig. 7 and Fig. 8) that two of the identified cluster states, namely the states 3 and 4, are describing the blocking situation in the upper troposphere. Fig. 6 shows the comparison of the sum of cluster affiliations $\gamma_3\left(t\right) + \gamma_4\left(t\right)$ for the part of the time series with the scaled zonally averaged *Lejenas-Okland blocking index.* It indicates the appearance of a blocking anticyclone and the duration of the event. We have a blocking if the geopotential height difference at 500 hP between $40°$ N and 60° N is negative over a region with 20° zonal extent. The exact formula is given in (Lupo et al. 1997), for the purpose of representation we have computed a zonally averaged value of the index, rescaled it and reversed its sign.

Next we aim at demonstrating how the postprocessing methods described in section 3 can help to gain the additional insight into the data to understand the impact of the external factors. Figs. 7 and 8 show the EOF-backprojection to the original 144-dim. space of the *mean dynamical equilibrium positions* in cluster 3 and 4 calculated at different times with the respective values of the factors u_t^1 and u_t^2. As can be seen from the graphics, impact of the seasonal factor u_t^2 results in significant weakening of the blocking states in summer, this finding is consistent to the observations reported in the literature. The impact of the increasing CO_2 concentration also results in the weakening of the blocking situation in both states 3 and 4, less pronounced but still statistically significant (see the respective confidence intervals in Figs. 7 and 8)

Finally, in the same manner as in the previous nostationary Lorenz'96 example, we construct an ensemble prediction model (with 10000 ensemble members) based on the FEM-VARX clustering results and compare the resulting predictions with the ones obtained by other methods. At first, we verify the Markov assumption in the time-homogenous approximation for the identified affiliation function

$\Gamma(t)$, generate the ensemble predictions based on the estimated 20-dimensional FEM-VARX parameterization (based on the first 90% of the data) and calculate the expectation values of the relative prediction errors (in 2-norm, based on the last 10% of the data). As can be seen from Fig. 9, FEM-VARX ensemble predictions produces much better predictions then the *constant model* (where the prediction is always the expectation value over the whole history) and the *"same as today model"* (where the prediction for the next day is just the state of the system now). Compared with the globally stationary VARX model (where the ensemble prediction is calculated from 10000 ensemble members of the global VARX model), FEM-VARX produces only slightly better predictions (approx. 2% better). This observation could be explained by the low overall persistence of the process $\Gamma(t)$ compared, for example, to the previous Lorenz'96 example where the hidden process was very persistent. This issue needs a deeper understanding and is a matter of future research.

5. Conclusion and discussion

A numerical FEM-VARX scheme for a data-driven parameterization of nonstationary multiscale dynamical processes was introduced. As demonstrated in the present paper, the application of the FEM-VARX method allows for the good description of the analyzed nonlinear and nonstationary data (in terms of recovering the nonlinear effects associated with the regime-switching and in terms of the good prediction quality of the resulting reduced representation). Besides that, a wide range of data-analysis techniques, e. g., from *information theory* (like *Bayesian information criterion*, cf. McQuarrie and Tsai (1998)), *model reduction approaches* (like *balanced truncation*, cf. Moore (1981)) and *adaptive Finite Element Methods* become available and can be applied in the FEM-VARX context to postprocess the obtained results and to get an additional insight into the analyzed data. In contrast to other multiscale approaches known from the literature (Majda et al. 2003; Fatkullin and Vanden-Eijnden 2004), presented numerical scheme is not a systematic strategy based on the knowledge of some ”first principles” but is purely data-driven and results in an approximation of the analyzed process by means of a sequence of ”simple” linear factor models. From this perspective, it is important to put further efforts into development of mixed numerical schemes, combining some features of systematic ”first principles” approaches (and with this some *a priory* knowledge about the analyzed dynamics) with some aspects of the purely data driven approaches (like the one presented in this paper).

One of the main accents in the present paper was on implicit mathematical assumptions being done on different stages of the derivation of the numerical method and postprocessing of the obtained results. In context of the meteorological application it was shown how big is the impact of implicit stationarity assumption on the analysis of climate factors influence: whereas in the case of the globally stationary VARX model with constant coefficients, the impacts of the seasonal trend and CO_2 concentration were found to be statistically insignificant, the nonstationary FEM-VARX clearly reveals the statistical significance of the two factors (see Fig. 5). It is important to emphasize that the applicability of the presented method as well as subsequent interpretation and postprocessing of the obtained results are dependent on the fulfillment of the mathematical assumptions involved. Conclusions that are drawn in every specific application case are reliable only modulo completion of those assumptions, it is one of the most important issues to be kept in mind when applying the methods like the one presented in this paper.

Acknowledgments

The work was partially supported by the DFG SPP 1276 METSTROEM ”Meteorology and Turbulence Mechanics”, DFG Research Center ”Matheon” (Project E8) and the Center of Scientific Simulations (CSS Berlin).

Appendix

Inserting (25) and (27) into (21) we get

$$f(v_i^+, v_i^-, \gamma_{00}^i) = \sum_{i=1}^{K} \left(g_i\left(\theta_i\right) \mathcal{D}^{-1}(v_i^+ - v_i^-) + \gamma_{00}^i g_i\left(\theta_i\right) \mathbb{1} \right) \rightarrow \min_{v_i^+, v_i^-, \gamma_{00}^i, \Theta} \tag{42}$$

subject to

$$\sum_t v_i^+(t) + v_i^-(t) \;\; \leq \;\; C \quad \forall i, \tag{43}$$

$$\sum_{i=1}^{K} \mathcal{D}^{-1}(v_i^+(t) - v_i^-(t)) \;\; = \;\; 1 - \sum_{i=1}^{K} \gamma_{00}^i \quad \forall t, \tag{44}$$

$$\mathcal{D}^{-1}(v_i^+(t) - v_i^-(t)) \;\; \geq \;\; -\gamma_{00}^i \quad \forall t, i, \tag{45}$$

$$v_i^+(t) \geq 0, v_i^-(t) \;\; \geq \;\; 0 \quad \forall i, t. \tag{46}$$

Note that the second above expression is the equivalent transformation of the original equality condition (6), therefore we can not just set $\sum_{i=1}^{K} \gamma_{00}^i = 1$ without losing the preservation of the equality constrain (6) in the transformed optimization problem.[8] By defining

$$A_{\text{eq}} = [A^1_{\text{eq}}, \ldots, A^K_{\text{eq}}], A^i_{\text{eq}} \quad = \quad [\mathcal{D}^{-1} \quad -\mathcal{D}^{-1} \quad \mathbb{1}], b_{\text{eq}} = \mathbb{1} \tag{47}$$

$$A^1_{\text{neq}} \quad = \quad \text{diag}\left(\begin{bmatrix} A^1_{\text{eq}} & A^2_{\text{eq}} & \ldots & A^K_{\text{eq}} \end{bmatrix}\right), b^1_{\text{neq}} = 0 \tag{48}$$

$$A^2_{\text{neq}} \quad = \quad \begin{bmatrix} \begin{bmatrix} \text{Id}_{n-1} & & \\ & \text{Id}_{n-1} & \\ & & 0 \end{bmatrix} & \\ & \ddots\,(\mathbb{K}\text{ times}) \end{bmatrix}, b^2_{\text{neq}} = 0, \tag{49}$$

$$A^3_{\text{neq}} \quad = \quad \text{diag}\left([-\mathbb{1} \quad -\mathbb{1} \quad 0]\right), b^3_{\text{neq}} = -C \tag{50}$$

and using the definition (28) we can re-write (21-24) in the matrix form (29).

[8] One could take the condition $\sum_{i=1}^{K} \gamma^i_{00} = 1$ as an additional explicit constrain ending up with two equality constrains instead of one. This will obviously increase the overall numerical cost of the method.

References

Bezdek, J., 1981: *Pattern recognition with fuzzy objective function algorithms.*. Plenum Press, New York.

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

Brockwell, P. and R. Davis, 2002: *Introduction to Time Series and Forecasting*. Springer, Berlin.

Calvetti, D., S. Morigi, L. Reichel, and F. Sgallari, 2000: Tikhonov regularization and the l-curve for large discrete ill-posed problems. *J. Comput. Appl. Math.*, **123**, 423–446, doi:http://dx.doi.org/10.1016/S0377-0427(00)00414-3.

Chernik, M., 1999: *Bootstrap methods and their application, a practitioner's guide*. Wiley Series in Probability and Statistics.

Crommelin, D. and E. Vanden-Eijnden, 2008: Subgrid-scale parameterization with conditional Markov chains. *J. Atmos. Sci.*, **65**, 2661–2675.

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

Fatkullin, I. and E. Vanden-Eijnden, 2004: *A computational strategy for multiscale systems with applications to Lorenz96 model*, volume 200 of *J. Comp. Phys.*. 605-638 pp.

Franzke, C., I. Horenko, A. Majda, and R. Klein, 2009: Systematic metastable atmospheric regime identification in an AGCM. *J. Atmos. Sci.*, **66(7)**, 1997–2012.

Golub, G. and C. V. Loan, 1989: *Matrix Computations*. Johns Hopkins University Press, Baltimore.

Höppner, F., F. Klawonn, R. Kruse, and T. Runkler, 1999: *Fuzzy cluster analysis.*. John Wiley and Sons, New York.

Horenko, I., 2009a: Finite element approach to clustering of multidimensional time series. *to appear in SIAM J. of Sci. Comp., (available via* page.mi.fu-berlin.de/horenko/*)*.

— 2009b: On clustering of non-stationary meteorological time series. *Dyn. of Atm. and Oc.*, doi:10.1016/j.dynatmoce.2009.04.003.

— 2009c: On robust estimation of low-frequency variability trends in discrete markovian sequences of atmospherical circulation patterns. *J. of Atmos. Sci.*, **66(7)**, 2059–2072.

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(10)**, 1–19.

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)**, 1125–1145.

Khouider, B., A. Majda, and M. Katsoulakis, 2003: Coarse-grained stochastic processes for tropical convection and climate. *PNAS*, **100**, 11941–11946.

Lorenz, E. N., 1996: Predictability - a problem partly solved. *Proc. of Sem. on Predictability*, **1**, 1–18.

Lupo, A. R., R. J. Oglesby, and I. I. Mokhov, 1997: Climatological features of blocking anticyclones: a study of northern hemisphere ccm1 model blocking events in present-day and double co_2 concentration atmospheres. *Climate Dynamics*, **13**, 181–195.

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

Majda, A., I. Timofeev, and E. Vanden-Eijnden, 1999: Models for stochastic climate prediction. *PNAS*, **96**, 14687–14691.

— 2003: Systematic strategies for stochastic mode reduction in climate. *J. Atmos. Sci.*, **60**, 1705–1722.

McQuarrie, A. and C. Tsai, 1998: *Regression and time series model selection*. World Scientific.

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)**, 353–375.

Moore, B., 1981: Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE Trans. Automat. Contr.*, **26**, 17–32.

Moreau, J., 1988: *Bounded variation in time*. Birkhauser.

Orrell, D., 2003: Model error and predictability over different timescales in the Lorenz 96 systems. *J. Atmos. Sci.*, **60**, 2219–2228.

Orrell, D. and L. Smith, 2003: The spectral bifurcation diagram: Visualizing bifurcations in high-dimensional systems. *Int. J. Bifurcat. Chaos*, **13**, 3015–3027.

Pereda, E., R. Quiroga, and J. Bhattacharya, 2005: Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, **77(1-2)**, 1–37.

Reinsel, G., 1993: *Elements of multivariate time series analysis*. Springer, New York.

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, volume X, 699–744.

Simmons, A. and J. Gibson, 2000: The ERA 40 project plan. *ERA 40 Project Rep. Ser. 1*, european Center for Medium-Range Weather Forcasting, Reading.

Tsay, R., 2005: *Analysis of financial time series*. Wiley Series in Probability and Statistics, Wiley-Interscience.

Wilks, D. S., 2005: Effects of stochastic parametrization in the Lorenz96 model. *Quart. J. of Met. Soc.*, **131**, 389–407.

List of Figures

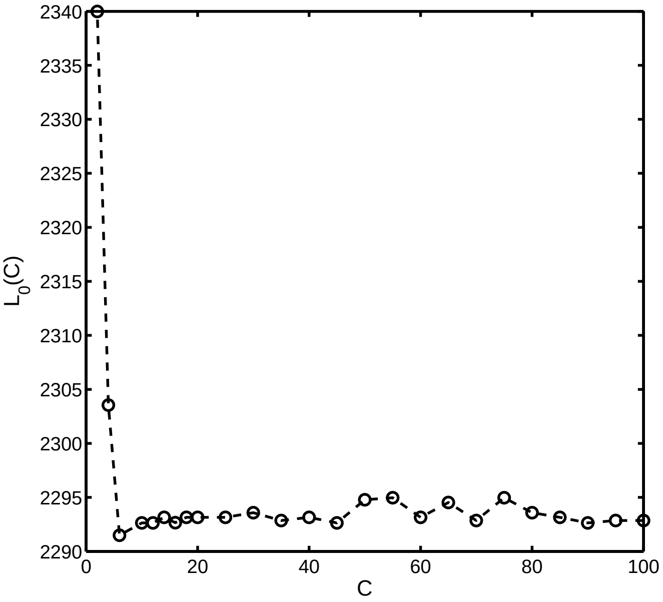

Figure 1: Optimal value of the functional (34) for different values of persistence threshold C as calculated for the time series of forced Lorenz'96 model ($K = 2, N = 200$, parameters of the Lorenz'96 model as specified in text, for each C optimization is repeated 100 times with randomly generated initial values and the result with minimal $\mathbf{L}_0(C)$ is kept).

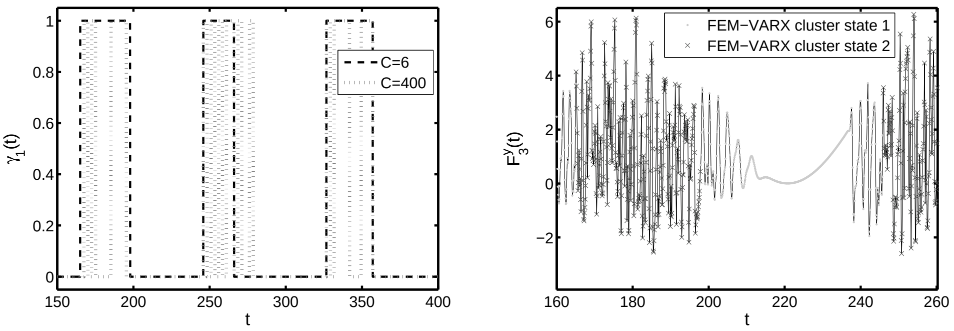

Figure 2: Left panel: Cluster affiliation function $\gamma_1(t)$ (the calculation is performed with the *FEM-VARX-algorithm* for $C = 6$ (dashed) and $C = 400$ (dotted). Right panel: fragment of the subscale data $F_3^y(t)$ together with the data affiliation correspondent to the cluster affiliation function from the left panel ($C = 6, N = 200, K = 2$).

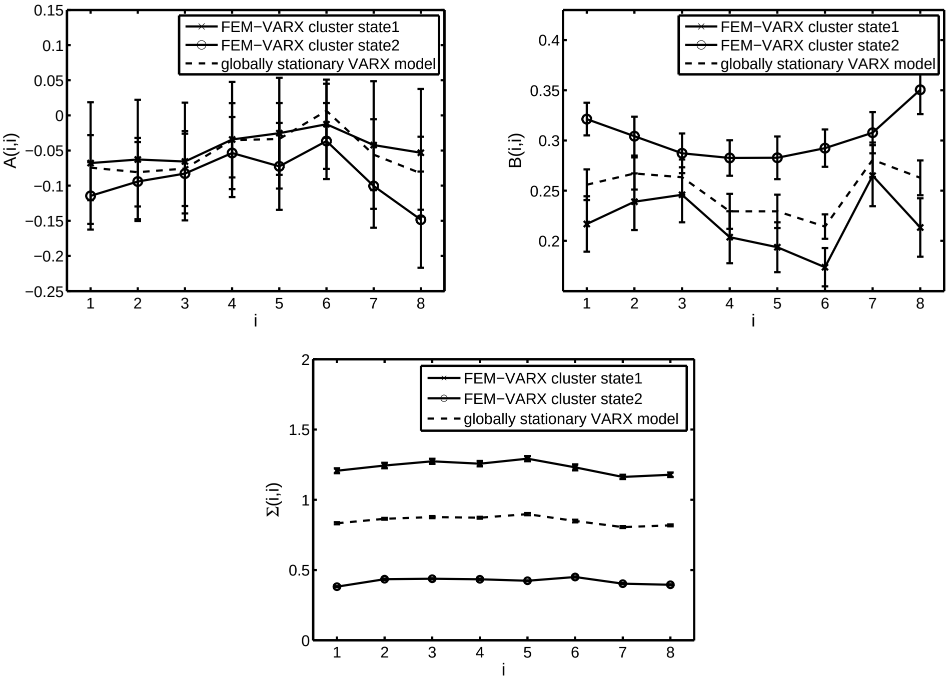

Figure 3: Diagonal elements of the local VARX parameters ($C = 6, N = 200, K = 2$).

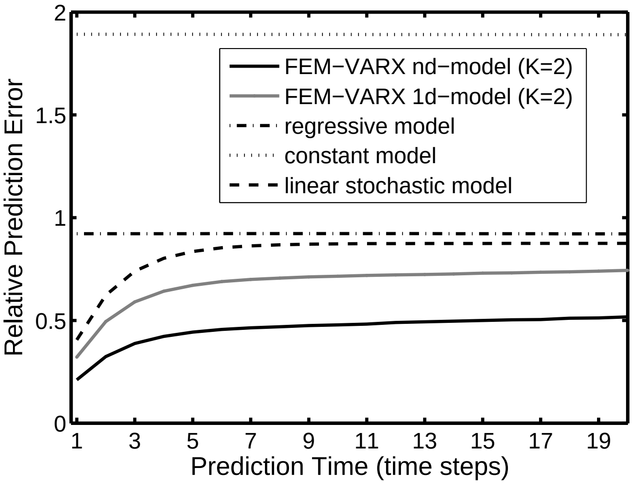

Figure 4: Comparison of the mean relative prediction errors calculated for the last 10% of the time series based on the different models trained on the first 90% of the data (see the description in text).

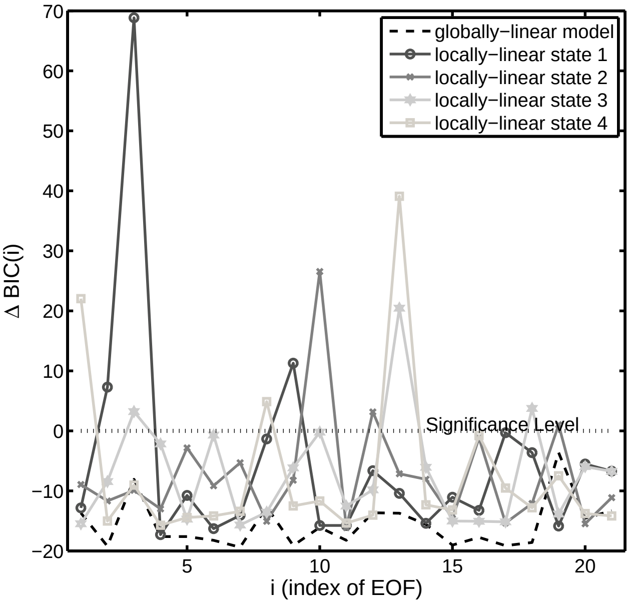

Figure 5: Differences between Bayesian Information Criterion (BIC) (39) as calculated for different EOF dimensions of the VARX models with and without the factors u_t^1 and u_t^2 (negative values indicate the EOF dimensions where the influence of both factors is statistically nonsignificant): for global stationary linear VARX model (dashed) and for local stationary linear FEM-VARX factor models (calculated with $K = 4, N = 4000, C = 3000, q = 2$) (solid lines). Dotted zero-line marks the statistical significance level (components above this line are significant in a sense of the BIC-criterion).

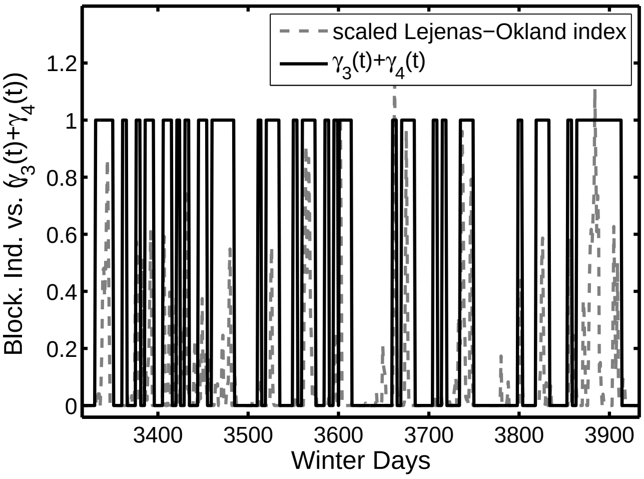

Figure 6: Comparison of the negative Lejenas-Oakland blocking index (dashed line) and the sum of cluster affiliations of locally-linear states 3 and 4 (solid line, calculated with FEM-VARX for $K = 4, N = 4000, C = 3000, q = 2$).

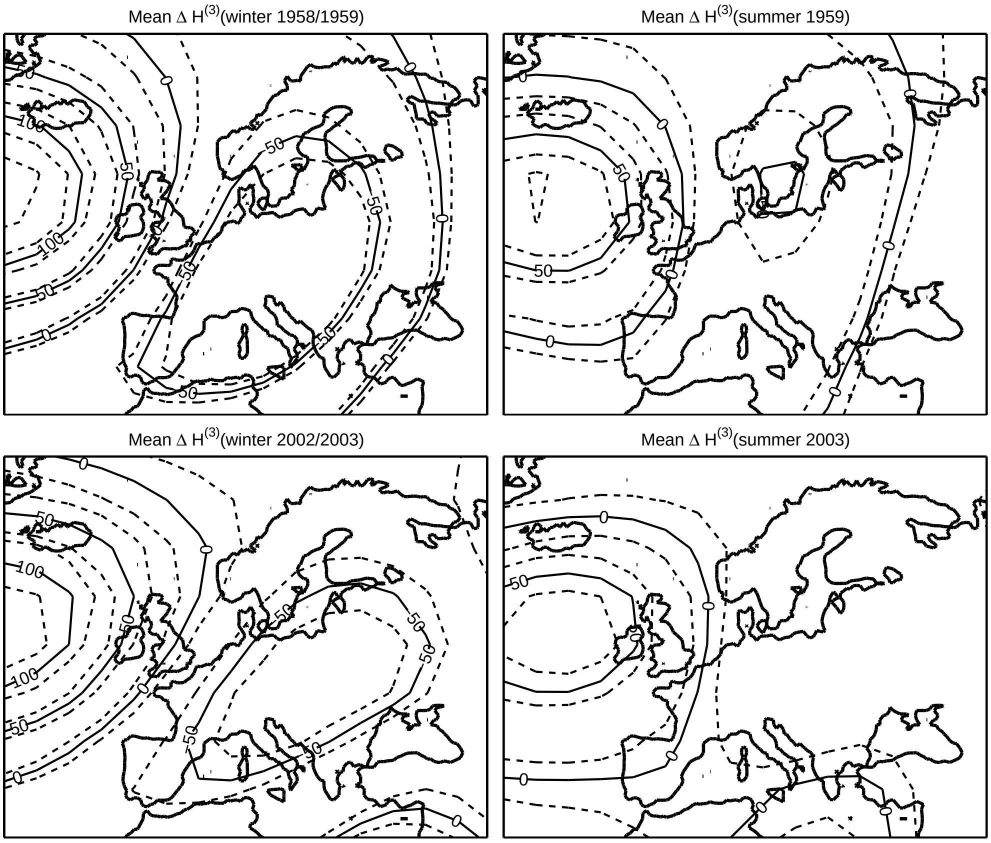

Figure 7: FEM-VARX cluster state 3 (blocking state): *Mean dynamical equilibrium positions* $\mathbb{E}^{(3)}\left(u_t^1, u_t^2\right)$ (see formula (40)) for the deviations ΔH of the geopotential heights (solid) and their confidence intervals (dashed) at different times (revealing the influence of both external factors separately, see explanation in text).

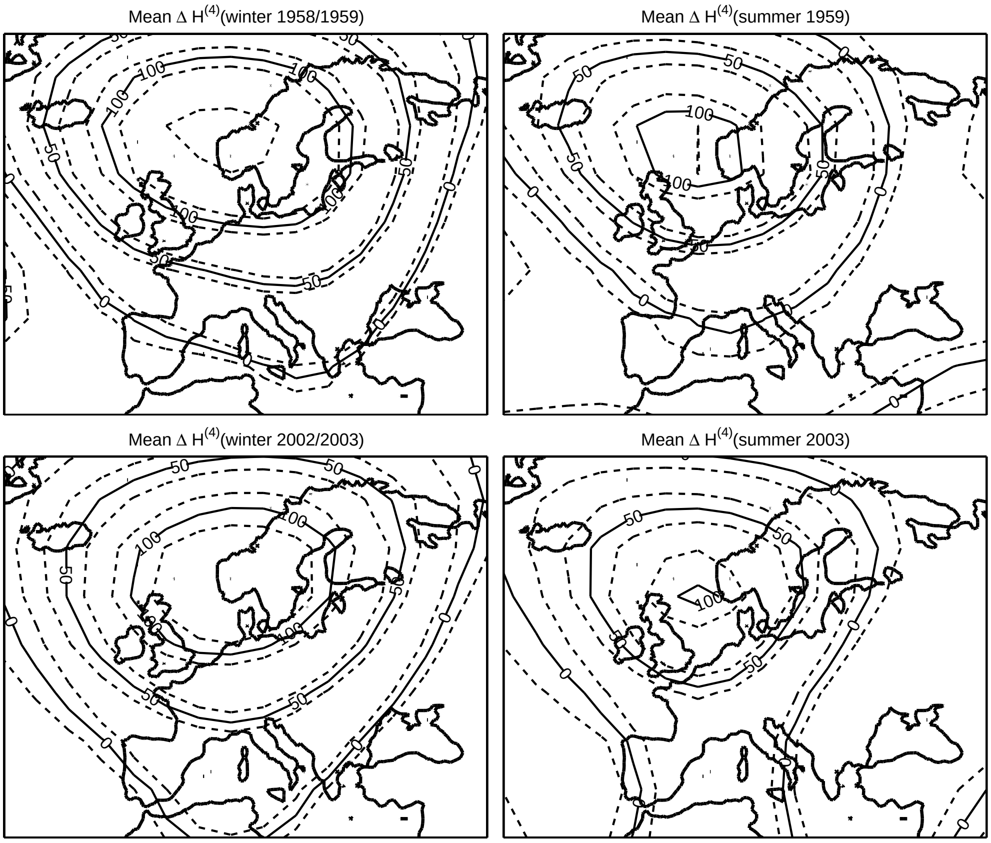

Figure 8: FEM-VARX cluster state 4 (blocking state): *Mean dynamical equilibrium positions* $\mathbb{E}^{(3)}\left(u_t^1, u_t^2\right)$ (see formula (40)) for the deviations ΔH of the geopotential heights (solid) and their confidence intervals (dashed) at different times (revealing the influence of both external factors separately, see explanation in text).

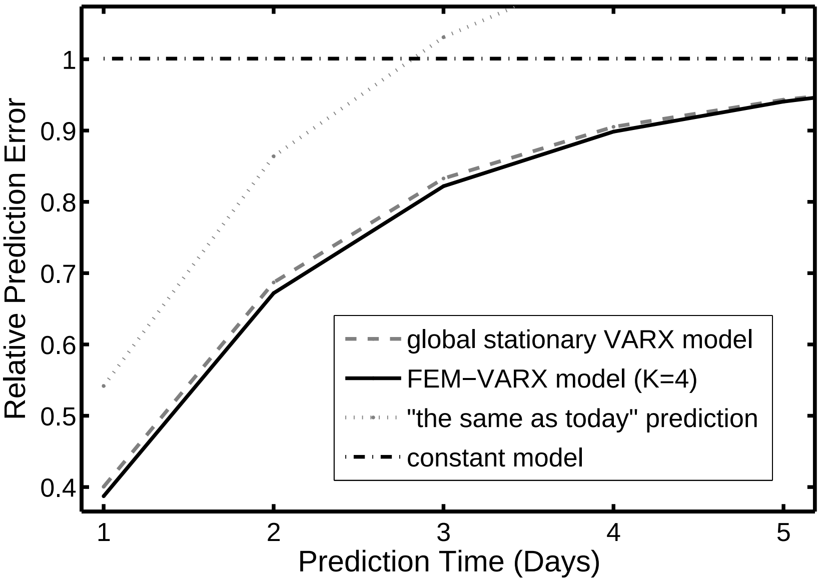

Figure 9: Comparison of the mean relative prediction errors calculated for the last 10% of the time series based on the different models trained on the first 90% of the data (see detailed description and discussion in text).