Cluster-weighted stochastic subgrid-scale modelling

Frank Kwasniok

College of Engineering, Mathematics and Physical Sciences
University of Exeter
North Park Road, Exeter EX4 4QF, United Kingdom
F.Kwasniok@exeter.ac.uk

ABSTRACT

This paper presents the main results of a recent publication (Kwasniok, 2011a).

A new approach for data-based stochastic parametrisation of unresolved scales and processes in numerical weather and climate prediction models is introduced. The subgrid-scale model is conditional on the state of the resolved scales, consisting of a collection of local models. A clustering algorithm in the space of the resolved variables is combined with statistical modelling of the impact of the unresolved variables. The clusters and the parameters of the associated subgrid models are estimated simultaneously from data. The method is implemented and explored in the framework of the Lorenz '96 model using discrete Markov processes as local statistical models. Performance of the cluster-weighted Markov chain (CWMC) scheme is investigated for long-term simulations as well as ensemble prediction. It clearly outperforms simple parametrisation schemes and compares favourably with another recently proposed subgrid modelling scheme also based on conditional Markov chains.

1 Introduction

The dynamics of weather and climate encompass a wide range of spatial and temporal scales. Due to the nonlinear nature of the governing equations, which are the laws of fluid dynamics, thermodynamics, radiative energy transfer and chemistry, the different scales are dynamically coupled to each other. Finite computational resources limit the spatial resolution of weather and climate prediction models; small-scale processes such as convection, clouds or ocean eddies are not properly represented. The necessity arises to account for unresolved scales and processes through the use of some form of subgrid modelling. This is usually referred to as a closure in fluid dynamics and theoretical physics, and as a parametrisation in meteorology and climate science.

Traditionally, parametrisations of unresolved scales and processes in numerical weather and climate prediction models have been formulated deterministically. Such bulk formulae are expected to capture the mean effect of small-scale processes in terms of some larger-scale resolved variables. However, there is in general a strong non-uniqueness of the unresolved scales with respect to the resolved scales. Thus, no one-to-one correspondence between values of the resolved variables and subgrid-scale effects can be expected; rather, a particular realisation of the subgrid term can be imagined to be drawn from a probability distribution conditional on the resolved variables.

Adding stochastic terms to climate models, in an attempt to capture the impacts of unresolved scales has been suggested in a seminal paper by Hasselmann (1976). First implementations of this concept were in the context of sea-surface temperature anomalies (Frankignoul and Hasselmann, 1977) and a conceptual zonally averaged climate model (Lemke, 1977). Another early study looked at regime behaviour in a very simple atmospheric model under stochastic forcing (Egger, 1981).

Despite impressive improvements in the forecast skill of numerical weather and climate prediction in the

past decades, there are still limitations due to model uncertainty and error as well as problems in generating initial conditions for ensembles. Forecast ensembles tend to be underdispersive (e. g., Buizza, 1997), leading to overconfident uncertainty estimates and an underestimation of extreme weather events. Systematic biases are significant in subgrid-scale weather phenomena and state-of-the-art ensemble prediction systems occasionally miss extreme weather events in the ensemble distribution. One way of addressing these issues relating to model imperfection is to deliberately introduce an element of uncertainty into the model. This can be done by randomisation of existing parametrisation schemes; approaches include multi-model, multi-parametrisation and multi-parameter ensembles (Palmer et al., 2005). A more systematic and comprehensive representation of model uncertainty may be achieved by introducing stochastic terms into the equations of motion. This has been implemented in the form of stochastically perturbed tendencies (Buizza et al., 1999) and, most recently, stochastic-dynamic subgrid schemes (Palmer, 2001; Shutts, 2005; Berner et al., 2008). A general feature of stochastic parametrisations is that they enable the forecast ensemble to explore important regions of phase space better than more restricted deterministic parametrisations. See Palmer et al., 2005; Weisheimer et al., 2011 for an overview and comparison of different methods for representing model uncertainty and error in weather and climate prediction models.

There has been a lot of research activity on subgrid modelling in recent years in various contexts, from theoretical studies constructing deterministic equations for moments of coarse-grained variables using a constrained measure of the system (Chorin et al., 1998), to a systematic stochastic mode reduction strategy based on stochastic differential equations (Majda et al., 1999, 2003), to various approaches to stochastic convection parametrisation (Lin and Neelin, 2000; Majda and Khouider, 2002; Plant and Craig, 2008). A particular class of subgrid models are schemes which are derived purely from data (Wilks, 2005; Crommelin and Vanden-Eijnden, 2008). While being less transparent from a physics point of view, they are potentially more accurate as they are less restricted by a priori assumptions.

The purpose of the present paper is twofold: Firstly, it generally proposes a new approach to data-based stochastic subgrid parametrisation using the methodology of cluster-weighted modelling. Secondly and more specifically, a cluster-weighted Markov chain subgrid scheme is outlined, building on recent work on conditional Markov chains (Crommelin and Vanden-Eijnden, 2008).

The paper is organised as follows: Section 2 introduces the general framework of cluster-weighted modelling for subgrid parametrisation. In section 3, we describe the Lorenz '96 system which is here used as a testbed to explore the method. The detailed formulation of the subgrid parametrisation in the context of the Lorenz '96 system and how to estimate its parameters from data is discussed in section 4. Then the results are presented in section 5. The paper concludes with some general discussion and implications.

2 Subgrid-scale parametrisation using cluster-weighted modelling

Assume the climate system is described by a high-dimensional state vector $\mathbf{u}$ which is decomposed as $\mathbf{u} = (\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ is the part resolved in a given weather or climate prediction model of a particular spatial resolution and complexity, and $\mathbf{y}$ is the unresolved part. The true tendency of $\mathbf{x}$ is schematically given by

$$\dot{\mathbf{x}} = \mathbf{R}(\mathbf{x}) + \mathbf{U}(\mathbf{x}, \mathbf{y}) \tag{1}$$

with $\mathbf{R}(\mathbf{x})$ being the resolved tendency, arising from the interactions among the resolved variables $\mathbf{x}$, and $\mathbf{U}(\mathbf{x}, \mathbf{y})$ being the unresolved tendency, arising from interactions with the unresolved variables $\mathbf{y}$. In a simulation with the model resolving only $\mathbf{x}$, we need to parametrise $\mathbf{U}(\mathbf{x}, \mathbf{y})$. Such a parametrisation has the general form

$$\mathbf{U}(\mathbf{x}, \mathbf{y}) \sim \mathbf{f}(\mathbf{x}) + \eta(\mathbf{x}) \tag{2}$$

where $\mathbf{f}(\mathbf{x})$ is the deterministic part of the closure model and $\eta(\mathbf{x})$ is a stochastic process generally dependent on $\mathbf{x}$. A canonical choice for the deterministic part would be the conditional mean of the unresolved tendency:

$$\mathbf{f}(\mathbf{x}) = \langle \mathbf{U}(\mathbf{x},\mathbf{y}) | \mathbf{x} \rangle \tag{3}$$

The stochastic component $\eta(\mathbf{x})$ is represented by a collection of local subgrid models, conditional on the state of the resolved variables. We build on the approach of cluster-weighted modelling (Gershenfeld et al., 1999; Kwasniok, 2011b) which is suitably adapted here. A finite number of clusters is introduced in a space of clustering variables $\mathbf{z}$. The number of clusters is M and m is the cluster index, running from 1 to M. The integer variable c takes values from 1 to M, according to which cluster has been chosen. Each cluster has an overall weight $w_m = p(c=m)$, satisfying the probabilistic constraints $w_m \geq 0$ and $\sum_m w_m = 1$, as well as a clustering probability density $p(\mathbf{z}|c=m)$, describing its domain of influence in the space of clustering variables $\mathbf{z}$. The vector $\mathbf{z}$ is a suitably chosen (low-dimensional) subset or projection of $\mathbf{x}$; it may also contain past values of $\mathbf{x}$, that is, a time-delay embedding (Sauer et al., 1991). Each cluster is associated with a local probabilistic subgrid model $p(\eta|\mathbf{v},c=m)$ which depends on a vector of variables $\mathbf{v}$. The vector $\mathbf{v}$ might encompass present and past values of components or projections of $\mathbf{x}$ as well as past values of η. The conditional probability density of the stochastic subgrid term η is expanded into a sum over the clusters:

$$p(\eta|\mathbf{z},\mathbf{v}) = \sum_{m=1}^{M} g_m(\mathbf{z})\, p(\eta|\mathbf{v},c=m) \tag{4}$$

The state-dependent weights g_m of the individual models are given by Bayes' rule:

$$g_m(\mathbf{z}) = p(c=m|\mathbf{z}) = \frac{w_m\, p(\mathbf{z}|c=m)}{\sum_{n=1}^{M} w_n\, p(\mathbf{z}|c=n)}. \tag{5}$$

The local model weights satisfy $g_m \geq 0$ and $\sum_m g_m = 1$. The cluster-weighted subgrid model has two types of conditioning on the resolved variables: the dependence of the model weights g_m on $\mathbf{z}$ and the explicit dependence of the subgrid models on $\mathbf{v}$. The vectors $\mathbf{z}$ and $\mathbf{v}$ might overlap.

The clustering densities $p(\mathbf{z}|c=m)$ and the local subgrid models $p(\eta|\mathbf{v},c=m)$ can take various forms. The canonical choice for the clustering densities $p(\mathbf{z}|c=m)$ in the continuous case is Gaussian. For non-negative or strongly skewed variables other choices may be more appropriate. One may also partition the space of $\mathbf{z}$ into a finite number of bins; the clustering probabilities are then discrete probability distributions over these bins. The subgrid models $p(\eta|\mathbf{v},c=m)$ may be regression models on $\mathbf{v}$ with Gaussian uncertainty. In the present study, they are actually Markov chains governing the switching between discrete values of η.

The parameters of the clusters and the subgrid models are estimated simultaneously from a learning data set by maximising a suitably defined likelihood function. The number of clusters M is a hyperparameter of the method controlling the overall complexity of the subgrid model. It may be determined within the fitting procedure of the subgrid model by minimising the Akaike or Bayesian information criterion in the learning data set, or by maximising the cross-validated likelihood function in a data set different from the learning data set. Alternatively, the number of clusters may be determined from the performance of the subgrid model in finite-time prediction or a long-term integration measured by a suitable metric of interest.

3 The Lorenz '96 model

The Lorenz '96 (L96) model (Lorenz, 1996) is used as a testbed to explore the new subgrid parametrisation scheme. It has become popular in the weather and climate science community as a toy model

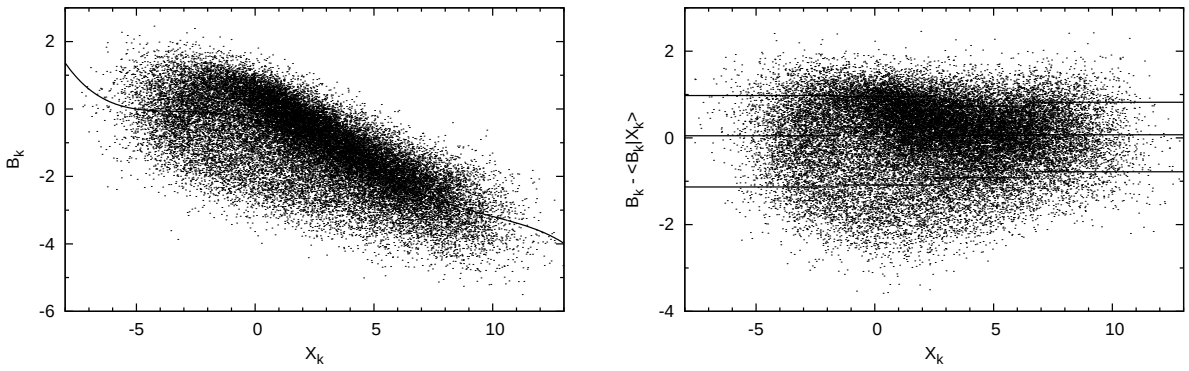

Figure 1: Left: Scatterplot of the subgrid term B_k versus the state X_k. The solid line indicates the conditional mean as estimated by a fifth-order polynomial least-squares fit. Right: Scatterplot of the deviation from the conditional mean, $\hat{B}_k = B_k - \langle B_k|X_k\rangle$, versus the state X_k. The solid horizontal lines indicate the values β_{il} used in the CWMC subgrid scheme (see text).

of the atmosphere to test concepts and algorithms relating to predictability, model error, ensemble post-processing and subgrid parametrisation (e. g., Lorenz, 1996; Palmer, 2001; Fatkullin and Vanden-Eijnden, 2004; Wilks, 2005; Crommelin and Vanden-Eijnden, 2008). The model equations are

$$\dot{X}_k = X_{k-1}(X_{k+1} - X_{k-2}) - X_k + F + B_k \tag{6}$$

$$\dot{Y}_{j,k} = \frac{1}{\varepsilon}\left[Y_{j+1,k}(Y_{j-1,k} - Y_{j+2,k}) - Y_{j,k} + h_y X_k\right] \tag{7}$$

with

$$B_k = \frac{h_x}{J}\sum_j Y_{j,k} \tag{8}$$

and $k = 1,\ldots,K$; $j = 1,\ldots,J$. The variables X_k and $Y_{j,k}$ are arranged on a circle. They can be interpreted either as variables on a circle of constant latitude or as meridional averages, each representing a segment of longitude. As such, the model is a spatially extended system. The X_k are large-scale, slow variables, each coupled to a collection of small-scale, fast variables $Y_{j,k}$. The variables are subject to the periodic boundary conditions $X_k = X_{k+K}$, $Y_{j,k} = Y_{j,k+K}$ and $Y_{j+J,k} = Y_{j,k+1}$ reflecting the periodicity of the spatial domain. The system is invariant under spatial translations; therefore all statistical properties are identical for all X_k. The model formulation employed here (Fatkullin and Vanden-Eijnden, 2004) is exactly equivalent to the original formulation by Lorenz (1996). With X_k^* and $Y_{j,k}^*$ denoting the variables in the original system (Lorenz, 1996) with parameters F, h, c and b, the corresponding system in the formulation of eqs.(6)–(8) is obtained by a linear scaling of the variables ($X_k = X_k^*$ and $Y_{j,k} = bY_{j,k}^*$) and the parameter setting $\varepsilon = \frac{1}{c}$, $h_x = -\frac{hcJ}{b^2}$ and $h_y = h$, leaving the forcing F unchanged. The present formulation of the system makes the time scale separation between the slow and fast variables explicit in the positive parameter ε. If $\varepsilon \to 0$, we have infinite time scale separation; if $\varepsilon \approx 1$, there is no time scale separation. We here use the parameter setting $K = 18$, $J = 20$, $F = 10$, $\varepsilon = 0.5$, $h_x = -1$ and $h_y = 1$, which is the same as in Crommelin and Vanden-Eijnden (2008). The system has 18 large-scale and 360 small-scale variables, 378 variables in total.

In a reduced model of the L96 system, only the variables X_k are resolved explicitly. The impact of the unresolved variables $Y_{j,k}$ on the resolved variables X_k is described by the term B_k which is referred to as the subgrid term or unresolved tendency. It needs to be parametrised somehow in a reduced model in order to account for the impact of the unresolved variables. This constitutes the subgrid-scale parametrisation problem in the context of the L96 model.

Figure 1 displays a scatterplot of the subgrid term B_k versus the state X_k obtained from a long (post-transient) numerical integration of the L96 model. The mean of B_k conditional on X_k as estimated

by a fifth-order polynomial least-squares fit is also indicated. A higher order of the polynomial does not improve the fit significantly. In practice, all numerical values of the conditional mean $\langle B_k|X_k\rangle$ are calculated using the fifth-order polynomial. There is a strong non-uniqueness of the subgrid term with respect to the resolved state: For a fixed value of X_k, B_k can take on a range of values. The conditional mean explains only 52.4% of the variance of the subgrid term B_k. The properties of the conditional probability density function $p(B_k|X_k)$ depend strongly on X_k. In particular, it is markedly non-Gaussian for a range of values of X_k. Figure 1 also shows a scatterplot of the deviation of the subgrid term from its conditional mean, $\hat{B}_k = B_k - \langle B_k|X_k\rangle$, versus X_k.

4 Subgrid-scale modelling with cluster-weighted Markov chains

As an example for the methodology outlined in Section 2, a cluster-weighted subgrid scheme based on local Markov chains is developed and implemented for the L96 model.

4.1 Model formulation

We here combine the framework of cluster-weighted modelling (Gershenfeld et al., 1999; Kwasniok, 2011b) with the use of conditional Markov chains (Crommelin and Vanden-Eijnden, 2008) for stochastic subgrid-scale parametrisation. The subgrid term $\hat{B}_k$ is replaced by a collection of discrete Markov processes conditional on the state of the resolved variables. The closure model is formulated independently for each resolved variable X_k as there is only little spatial correlation in the subgrid term B_k in the L96 system (Wilks, 2005). We choose to condition the subgrid model at time t both on the current state $X_k(t)$ and the increment $\delta X_k(t) = X_k(t) - X_k(t-\delta t)$ where δt is the sampling interval of the data. This choice is motivated by the fact that the probability density function of the subgrid term B_k has been shown to depend also on the increment δX_k (Crommelin and Vanden-Eijnden, 2008). It seems conceivable that the probability density of the subgrid term could be further sharpened by conditioning on more past values of X_k but we restrict ourselves to just one past value for simplicity.

The subgrid model is derived from an equally sampled data set of length N, $\{X_k^\alpha, \delta X_k^\alpha, \hat{B}_k^\alpha\}_{\alpha=1}^N$. Here and in the following, a subscript or superscript α refers to time in an equally sampled time series with sampling interval δt and runs from 1 to N. A data point $(X_k, \delta X_k, \hat{B}_k)$ is mapped to a discrete state (s,d,b) by partitioning the $(X_k, \delta X_k, \hat{B}_k)$-space into bins. The X_k-space is divided into N_X disjoint intervals $\{\mathscr{I}_i^X\}_{i=1}^{N_X}$; we have $s=i$ if $X_k \in \mathscr{I}_i^X$. The δX_k-space is divided into $N_{\delta X}$ disjoint intervals $\{\mathscr{I}_j^{\delta X}\}_{j=1}^{N_{\delta X}}$; we have $d=j$ if $\delta X_k \in \mathscr{I}_j^{\delta X}$. Given $s=i$, the range of possible values of $\hat{B}_k$ is divided into N_B disjoint, equally populated intervals $\{\mathscr{I}_{il}^B\}_{l=1}^{N_B}$; we have $b=l$ if $\hat{B}_k \in \mathscr{I}_{il}^B$. The subgrid term $\hat{B}_k$ is then represented by a set of N_B discrete values $\{\beta_{il}\}_{l=1}^{N_B}$ given by the mean of $\hat{B}_k$ in each interval:

$$\beta_{il} = \frac{\sum_\alpha \hat{B}_k^\alpha \mathbf{1}(s_\alpha = i)\mathbf{1}(b_\alpha = l)}{\sum_\alpha \mathbf{1}(s_\alpha = i)\mathbf{1}(b_\alpha = l)} \tag{9}$$

We introduce M clusters in the discrete (s,d,b)-space. Each cluster has an overall weight or probability of that cluster being chosen, $w_m = p(c=m)$, and a clustering probability distribution $\psi_{mij} = p(s=i, d=j|c=m)$, describing its domain of influence in (s,d)-space. The parameters of the clusters satisfy a couple of probabilistic constraints. The overall weights form a probability distribution: $w_m \geq 0$, $\sum_m w_m = 1$. The clustering probability distributions satisfy $\psi_{mij} \geq 0$ and $\sum_{i,j} \psi_{mij} = 1$. The clusters are required to add up to the joint climatological distribution (invariant measure) of s and d, that is, $\sum_m w_m \psi_{mij} = p(s=i, d=j) = \rho_{ij}$ where ρ_{ij} is empirically given as the fraction of data points in these bins: $\rho_{ij} = \frac{1}{N}\sum_\alpha \mathbf{1}(s_\alpha = i)\mathbf{1}(d_\alpha = j)$. It follows that the clusters also sum up to the marginal climatological distributions: $\sum_{m,j} w_m \psi_{mij} = p(s=i) = \sum_j \rho_{ij}$ as well as $\sum_{m,i} w_m \psi_{mij} = p(d=j) = \sum_i \rho_{ij}$.

Each cluster is associated with a Markov chain in the discrete space b described by an $(N_B \times N_B)$ transition matrix $\mathbf{A}_m$ with components $A_{ml_1l_2} = p(b_\alpha = l_2 | b_{\alpha-1} = l_1, c_\alpha = m)$. The matrices $\mathbf{A}_m$ are row-stochastic matrices, that is, $A_{ml_1l_2} \geq 0$ and $\sum_{l_2} A_{ml_1l_2} = 1$.

The conditional probability distribution for b_α is modelled as a sum over the clusters:

$$p(b_\alpha | b_{\alpha-1}, s_\alpha, d_\alpha) = \sum_{m=1}^{M} g_m(s_\alpha, d_\alpha) A_{mb_{\alpha-1}b_\alpha}. \tag{10}$$

The state-dependent model weights are given by Bayes' rule as

$$g_m(i,j) = p(c = m | s = i, d = j) = \frac{w_m \psi_{mij}}{\sum_{n=1}^{M} w_n \psi_{nij}} = \frac{w_m \psi_{mij}}{\rho_{ij}}. \tag{11}$$

The Markov chain is effectively governed by local transition matrices $\mathbf{A}^{\mathrm{loc}}(i,j) = \sum_m g_m(i,j)\mathbf{A}_m$ which as a convex combination of row-stochastic matrices are always row-stochastic matrices. The subgrid model jumps according to the local Markov process between the N_B possible values $\{\beta_{il}\}_{l=1}^{N_B}$ given by eq.(9) for $s = i$. The mean local model weights are found to be $\langle g_m \rangle = \frac{1}{N}\sum_\alpha g_m(s_\alpha, d_\alpha) = w_m$. Hence the overall weight w_m can be interpreted as the fraction of the data set (or the invariant measure of the system) accounted for by the cluster m.

The number of clusters M, the numbers of bins N_X and $N_{\delta X}$ as well as the number of states N_B of the Markov chain are hyperparameters of the method which have to be fixed beforehand; they control the overall complexity of the closure model. We call this subgrid model a cluster-weighted Markov chain (CWMC) model.

Given an equally sampled learning data set of length N, $\{b_0, s_1, d_1, b_1, \ldots, s_N, d_N, b_N\}$, the parameters of the CWMC subgrid model are estimated according to the maximum likelihood principle using the expectation-maximisation (EM) algorithm (Dempster et al., 1977; Kwasniok, 2011a, 2011b).

4.2 Model integration

The time integration of the reduced model with the CWMC subgrid scheme proceeds as follows: The subgrid scheme is constructed at time step δt; the deterministic equations for the resolved variables are integrated with time step h determined by the employed numerical scheme, stability and the desired accuracy. These two time steps may be different; typically, δt is larger than h. Assume for simplicity that δt is an integer multiple of h: $\delta t = N_{\mathrm{step}} h$. We then use a split-integration scheme (Crommelin and Vanden-Eijnden, 2008). The resolved dynamics are integrated with time step h; the subgrid model is propagated with time step δt, updated only every N_{step} time steps. At time $t_{\alpha-1}$, let the system state be $X_k^{\alpha-1}$ falling in bin $s_{\alpha-1}$ and let the state of the Markov chain of the subgrid model be $b_{\alpha-1}$. The state X_k^α at time t_α is calculated by propagating the resolved variables N_{step} times with step size h using the derivative given by eq.(6) with B_k set to $\langle B_k | X_k^{\alpha-1} \rangle + \beta_{s_{\alpha-1}b_{\alpha-1}}$. If X_k^α falls in bin s_α and $\delta X_k^\alpha = X_k^\alpha - X_k^{\alpha-1}$ falls in bin d_α the next state of the Markov chain b_α is determined by randomly drawing from the probability distribution given by eqs.(10) and (11). Then the subgrid term B_k is set to $\langle B_k | X_k^\alpha \rangle + \beta_{s_\alpha b_\alpha}$ for the next integration cycle. One could also choose to update the deterministic part of the closure model at time step h. In the present model setting there is virtually no difference between these two possibilities provided the sampling interval δt is not too large as the dependence of $\langle B_k | X_k \rangle$ on X_k is quite smooth. The rarer update is computationally more efficient.

5 Results

The full L96 model is integrated for 1000 time units of post-transient dynamics using the Runge-Kutta scheme of fourth order with step size 0.002. The state vector is archived at a sampling interval $\delta t = 0.01$,

resulting in a data set containing 100000 data points. The CWMC closure scheme is constructed from this data set. Such a large learning data set, virtually corresponding to the limit of infinite data, is used here to get rid of any sampling issues for the parameter estimates and study the ideal behaviour of the method. It should be noted that a very similar performance of the reduced model to that presented here can already be obtained with a much shorter learning data set (~ 5000 data points). We use $N_X = 4$ intervals in X_k-space. They are located between -5.5 and 10.5 and have equal size. We then extend the first and the last interval to minus and plus infinity, respectively. Thus, the intervals are $\mathscr{I}_1^X = (-\infty, -1.5]$, $\mathscr{I}_2^X = (-1.5, 2.5]$, $\mathscr{I}_3^X = (2.5, 6.5]$ and $\mathscr{I}_4^X = (6.5, \infty)$. In δX-space we use $N_{\delta X} = 2$ intervals given as $\mathscr{I}_1^{\delta X} = (-\infty, 0]$ and $\mathscr{I}_2^{\delta X} = (0, \infty)$, corresponding to downwards and upwards direction of the trajectory. The number of bins for the subgrid term, that is, the number of states of the Markov chain is set to $N_B = 3$. The values β_{il} used in the CWMC scheme given by eq.(9) are displayed in Fig. 1. The resolution of the binnings was determined from the performance of the resulting reduced model. We studied larger values for all of the parameters N_X, $N_{\delta X}$ and N_B but a higher resolution in the binning of any of the variables does not visibly improve the model. CWMC closure schemes were estimated from the data set with increasing number of clusters, starting from $M = 1$. Based on the performance of the reduced model, $M = 2$ is found to be the optimal number of clusters. There is no significant further improvement when using more than 2 clusters.

The CWMC scheme is compared to two simple generic parametrisation schemes: a deterministic closure scheme and the AR(1) scheme proposed by Wilks (2005). The deterministic scheme consists in parametrising B_k by the conditional mean as estimated by the fifth-order polynomial fit shown in Fig. 1: $B_k \sim \langle B_k | X_k \rangle$. The AR(1) scheme models B_k by the conditional mean plus an AR(1) process:

$$\hat{B}_k^\alpha = B_k^\alpha - \langle B_k | X_k^\alpha \rangle = \phi \hat{B}_k^{\alpha-1} + \sigma \xi \tag{12}$$

ξ denotes Gaussian white noise with zero mean and unit variance, σ is the standard deviation of the driving noise. For B_k, this amounts to an AR(1) process with state-dependent mean $\langle B_k | X_k \rangle$ but constant autoregressive parameter and standard deviation of the noise. A least-squares fit to the time series of $\hat{B}_k$ at time step $\delta t = 0.01$ yields $\phi = 0.9977$ (corresponding to an e-folding time of 4.25 time units) and $\sigma = 0.059$. The standard deviation of the AR(1) process is $\frac{\sigma}{\sqrt{1-\phi^2}} = 0.866$, equal to the standard deviation of $\hat{B}_k$. The reduced models with the deterministic and the AR(1) subgrid schemes are integrated in time in a manner analogous to that described in subsection 4.2 for the CWMC scheme, updating the subgrid term at time step δt.

The CWMC scheme is also compared to the subgrid modelling study by Crommelin and Vanden-Eijnden (2008) based on conditional Markov chains using the L96 system with exactly the same parameter setting as an example. They condition the Markov chain on X_k^α and $X_k^{\alpha-1}$, both partitioned into 16 bins. Taking into account that due to the autocorrelation of the system at lag δt only transitions within the same bin and into neighbouring bins actually occur this roughly (not exactly) corresponds to $N_X = 16$ and $N_{\delta X} = 3$ in the present setting. Then a separate transition matrix (with $N_B = 4$) is determined for each pair of bins, amounting to about 45 active transition matrices. We occasionally refer to this subgrid model for comparison as the full Markov chain scheme. The present CWMC scheme offers a much more condensed description of the subgrid term. It uses only $M = 2$ independent transition matrices. Moreover, a much coarser binning ($N_X = 4$, $N_{\delta X} = 2$) and only $N_B = 3$ states in the Markov chain are used. The number of parameters to be estimated from data is actually about 40 times larger in the full Markov chain scheme than in the CWMC scheme. Consequently, a longer learning data set is necessary to estimate the full Markov chain model.

5.1 Long-term dynamics of the reduced model

We investigate to what extent the reduced models with the various subgrid schemes are able to reproduce the statistical properties of the long-term behaviour of the large-scale variables X_k in the full L96 model.

	Mean	Std. dev.	D
Full L96 model	2.39	3.52	
Deterministic scheme	2.53	3.56	0.017
AR(1) scheme	2.51	3.57	0.015
CWMC scheme	2.40	3.51	0.004

Table 1: Mean and standard deviation of X_k in the L96 model and the reduced models with the various subgrid schemes. The last column gives the Kolmogorov-Smirnov distance between the probability distribution of X_k in the reduced model and that in the full L96 model.

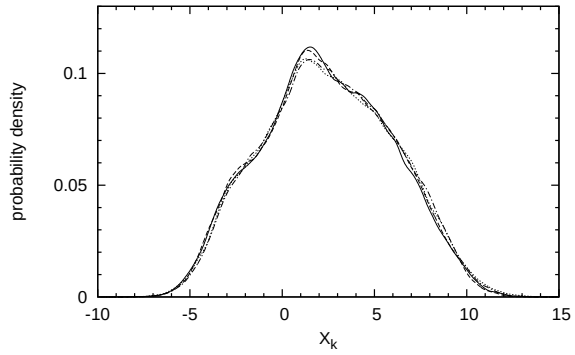

Figure 2: Probability density function of X_k in the full L96 model (solid) and in reduced models with deterministic subgrid scheme (dot-dashed), AR(1) scheme (dotted) and CWMC scheme (dashed).

The reduced models are integrated in time as described in subsection 4.2 using a fourth-order Runge-Kutta scheme with step size $h = 0.002$. The closure model is updated every fifth time step. The reduced model with CWMC subgrid scheme runs more than 30 times faster than the full L96 model. Starting from random initial conditions, after discarding the first 50 time units of the integration to eliminate transient behaviour 2500 time units worth of data are archived at a sampling interval of $\delta t = 0.01$, resulting in time series of length 250000. All the results reported below are calculated from these time series.

Figure 2 shows the probability density function of X_k in the reduced models with the three different closure schemes and that of the full L96 model for comparison. Additionally, table 1 gives the mean and the standard deviation of the PDFs. It also lists the deviation of the probability distributions of the reduced models from that of the full L96 model as measured by the Kolmogorov-Smirnov statistic $D = \max_{X_k} |\Phi(X_k) - \Phi_\mathrm{r}(X_k)|$ where Φ is the (cumulative) probability distribution of the L96 model and Φ_r is the probability distribution of the reduced model. All models reproduce the PDF quite well. The deterministic and the AR(1) schemes are very close to each other. The CWMC scheme offers an improvement on the two other schemes; it is about as good as the full Markov chain scheme (Crommelin and Vanden-Eijnden, 2008). The deterministic and the AR(1) schemes exhibit a considerable shift in the mean state and slightly too much variance. The CWMC scheme has almost exactly the correct mean and variance.

In order to monitor the spatial pattern of variability in the system, the Fourier wave spectrum of the system is displayed in Figure 3. At each instant in time, the discrete spatial Fourier transform of the state vector $\mathbf{X} = (X_1, \ldots, X_K)$ is calculated, giving the (complex) wave vector $\mathbf{G} = (G_0, \ldots, G_{K-1})$ with $G_\nu = G^*_{K-\nu}$ for $\nu = 1, \ldots, K-1$. The wave variance of wavenumber ν is then $\langle |G_\nu - \langle G_\nu \rangle|^2 \rangle$ for $\nu = 0, \ldots, K-1$ where $\langle \cdot \rangle$ denotes the time average. Figure 3 also shows the correlation of two variables X_k and X_{k+l}, separated by a lag l on the circle. The deterministic and the AR(1) closure schemes give

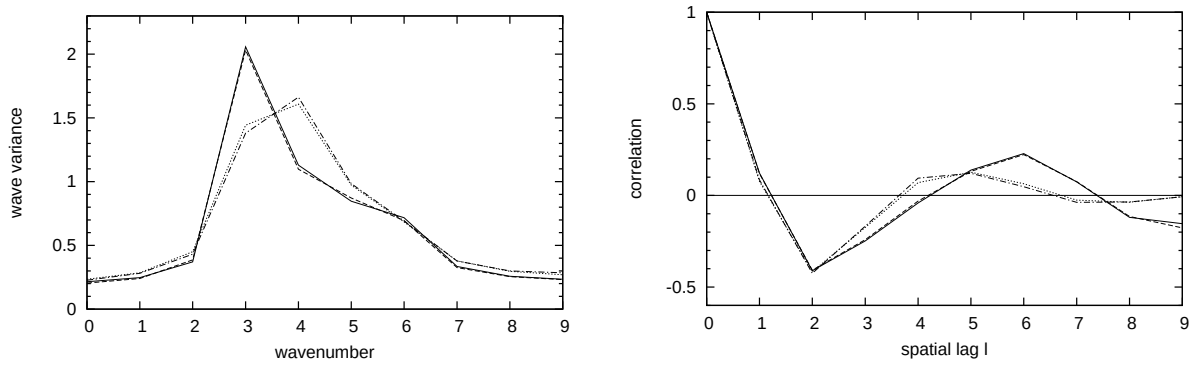

Figure 3: Left: Wave variances in the full L96 model (solid) and in reduced models with deterministic subgrid scheme (dot-dashed), AR(1) scheme (dotted) and CWMC scheme (dashed). Right: Correlation of X_k and X_{k+l} in the full L96 model (solid) and in reduced models with deterministic subgrid scheme (dot-dashed), AR(1) scheme (dotted) and CWMC scheme (dashed).

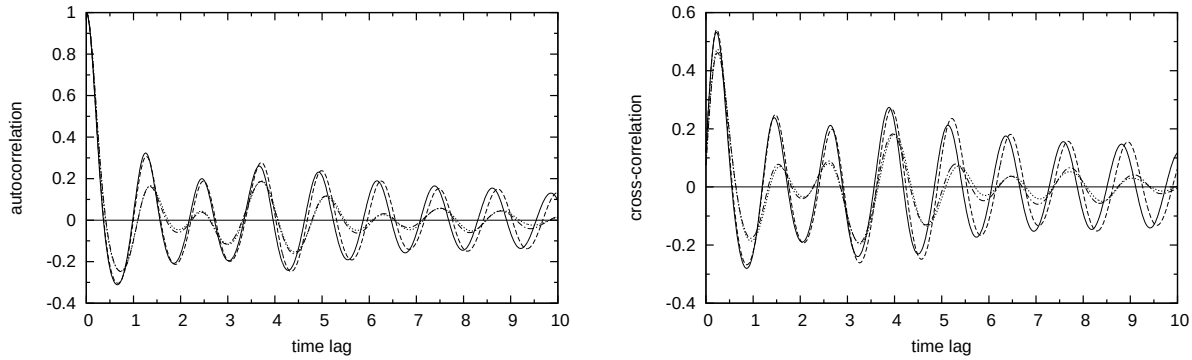

Figure 4: Left: Autocorrelation function of X_k in the full L96 model (solid) and in reduced models with deterministic subgrid scheme (dot-dashed), AR(1) scheme (dotted) and CWMC scheme (dashed). Right: Cross-correlation function of X_k and X_{k+1} in the full L96 model (solid) and in reduced models with deterministic subgrid scheme (dot-dashed), AR(1) scheme (dotted) and CWMC scheme (dashed).

very similar results. They exhibit large error; the peak in the wave spectrum is too small and broad, spread out over wavenumbers 3 and 4. The CWMC scheme captures the sharp peak at wavenumber 3 correctly and reproduces the whole spectrum almost perfectly; it performs as well as the full Markov chain model (Crommelin and Vanden-Eijnden, 2008). The same effect manifests itself in the spatial correlations. With the deterministic and the AR(1) schemes, the maximum positive correlation at lag 6 (associated with wavenumber 3) is shifted to lags 4 and 5 due to too much variance in the shorter waves with wavenumbers 4 and 5. The CWMC scheme reproduces the spatial correlations almost perfectly.

We now look at the behaviour of the models in the time domain. Figure 4 gives the autocorrelation function of X_k and the cross-correlation function of neighbouring variables X_k and X_{k+1}. They both have oscillatory behaviour over long time scales. The deterministic and the AR(1) scheme have the amplitude of the oscillations too small and there is a phase shift compared to the full L96 model. The CWMC scheme performs much better; the amplitude of the oscillations is correct and there is only a small phase shift visible at large time lags. The full Markov chain model is slightly better than the CWMC scheme; it reproduces the auto- and cross-correlation functions almost perfectly (Crommelin and Vanden-Eijnden, 2008).

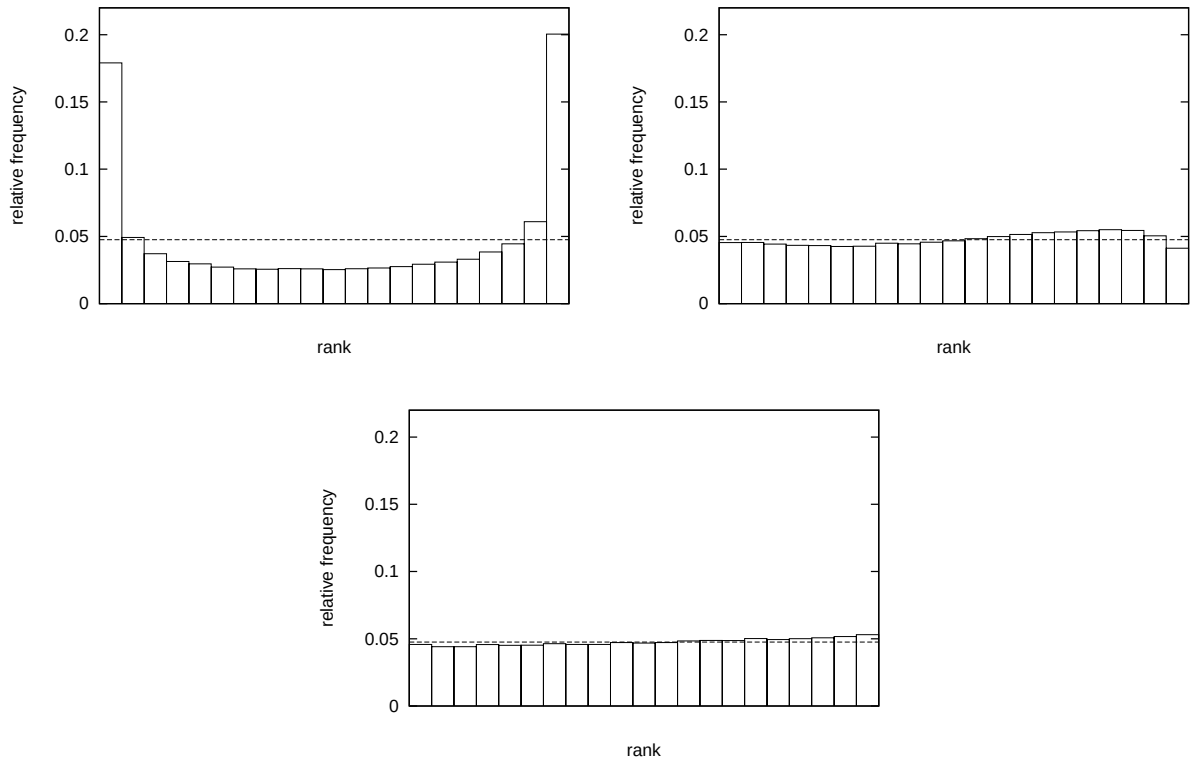

Figure 5: Rank histograms for ensembles with the deterministic subgrid scheme (top left), the AR(1) scheme (top right) and the CWMC scheme (bottom). Prediction lead time is $\tau = 2$; ensemble size is $N_{\mathrm{ens}} = 20$. The dashed horizontal lines indicate the expected relative frequency under rank uniformity.

5.2 Ensemble prediction

We investigate the predictive skill of the reduced models with the different parametrisations. Given the stochastic nature of the models an ensemble prediction framework appears to be most appropriate. We construct ensembles which account for both model and initial condition uncertainty. In these ensembles, each ensemble member starts from a randomly perturbed initial condition. We follow the procedure in Crommelin and Vanden-Eijnden (2008). The perturbations are drawn from a Gaussian distribution with zero mean and a standard deviation of 0.15 (about 4% of the climatological standard deviation of X_k), independently for each component X_k. The predictive skill turns out to be rather insensitive to the exact value of the amplitude of the perturbations. This simple generation of ensembles appears to be sufficient here for the purpose of comparing different subgrid-scale models. We do not sample unstable directions in phase space more heavily (as is done by Wilks (2005)) or identify fastest-growing perturbations using singular vectors. 10000 initial conditions are used taken from a long integration of the full L96 model, separated by 5 time units.

In order to assess the ensemble spread we use rank histograms (Hamill, 2001). Rank histograms give the relative frequency of the rank of the true state in the $N_{\mathrm{ens}} + 1$ member distribution formed by the ensemble members and the true state. Ideally, the rank histogram should be flat corresponding to a uniform distribution of the rank of the true state. For underdispersed ensembles the true state occupies the extreme ranks too often, showing up in a U-shaped rank histogram. Conversely, for overdispersed ensembles the extreme ranks are occupied too rarely, giving an inverse U-shape in the rank histogram. Figure 5 shows rank histograms for the different parametrisation schemes. The ensemble size is $N_{\mathrm{ens}} = 20$; the prediction lead time is $\tau = 2$. The rank histograms are representative also for other lead times. With the deterministic closure scheme the ensembles are strongly underdispersed. For the AR(1) scheme

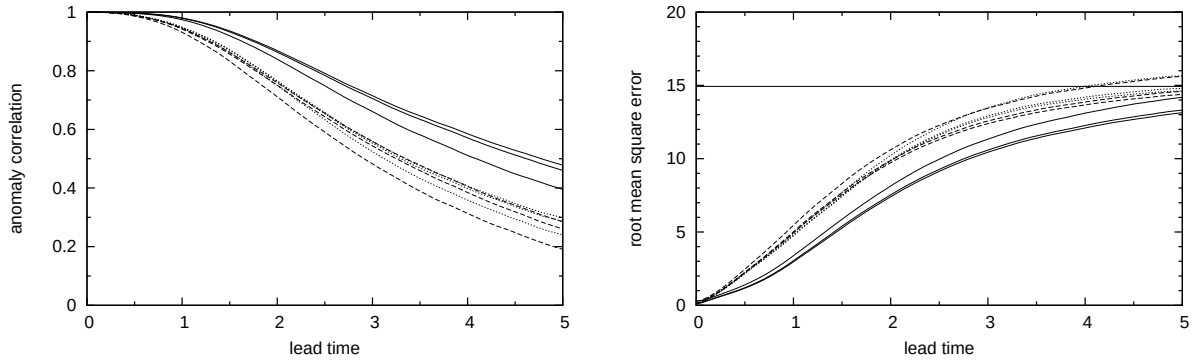

Figure 6: Prediction skill of the ensemble mean. Left: Anomaly correlation with the deterministic scheme (dotted), the AR(1) scheme (dashed) and the CWMC scheme (solid). Curves from bottom to top refer to ensemble sizes $N_{\mathrm{ens}} = 5$, $N_{\mathrm{ens}} = 20$ and $N_{\mathrm{ens}} = 50$. Right: Root mean square error with the deterministic scheme (dotted), the AR(1) scheme (dashed) and the CWMC scheme (solid). Curves from top to bottom refer to ensemble sizes $N_{\mathrm{ens}} = 5$, $N_{\mathrm{ens}} = 20$ and $N_{\mathrm{ens}} = 50$. The solid horizontal line indicates the root mean square error of the climatology forecast.

the rank histogram is nearly uniform apart from some bias the origin of which is unclear. The rank histogram for the AR(1) scheme is in accordance with the findings by Wilks (2005). Crommelin and Vanden-Eijnden (2008) report a substantial underdispersion for the AR(1) scheme. It is not clear where this discrepency comes from; it may be due to a different initialisation of the AR(1) process at initial time (Kwasniok, 2011a). For the CWMC model there is an almost ideal ensemble spread.

We now evaluate the actual predictive skill of the forecasts with the various subgrid schemes. We consider the deterministic forecast given by the ensemble mean. Figure 6 provides the anomaly correlation and the root mean square error with the three closure schemes for ensemble sizes $N_{\mathrm{ens}} = 5$, $N_{\mathrm{ens}} = 20$ and $N_{\mathrm{ens}} = 50$. For all schemes the prediction skill improves monotonically with the ensemble size for all lead times and is virtually converged at $N_{\mathrm{ens}} = 50$. The CWMC scheme clearly outperforms the two other schemes at all lead times; with $N_{\mathrm{ens}} = 5$ it is already better than the two other schemes with $N_{\mathrm{ens}} = 50$. The AR(1) scheme cannot consistently outperform the deterministic scheme. The CWMC subgrid model is not worse than the full Markov chain scheme (Crommelin and Vanden-Eijnden, 2008); it is even better at small and medium lead times, probably due to the state-dependent initialisation of the Markov chain at initial time (Kwasniok, 2011a).

6 Discussion

A new approach to data-driven stochastic subgrid modelling has been proposed. The closure consists of a collection of local statistical models associated with clusters in the space of resolved variables. As an example, the method has been implemented and tested in the framework of the Lorenz '96 model using discrete Markov chains as local models. The present scheme substantially outperforms two simple generic closure schemes, a deterministic one given by the conditional mean of the subgrid term and a stochastic one given by the conditional mean plus an AR(1) process. The cluster-weighted Markov chain (CWMC) scheme performes about as well as the conditional Markov chain scheme proposed by Crommelin and Vanden-Eijnden (2008) but the number of parameters is smaller by a factor of about 40.

The present method has the potential to be used in atmospheric and oceanic models based on grid point discretisation. In some sense, the L96 model is a spatially extended system on a one-dimensional grid. In a more realistic model two or three dimensions are present. The scheme could be run independently at each grid point which is still computationally feasible even for a very large number of grid points.

In a more realistic model setting the vector of variables **z** one would like to condition the model on is likely to be of higher dimension than in the L96 model. A conditioning based on binning into disjoint intervals as in Crommelin and Vanden-Eijnden (2008) then becomes rapidly impractical and some form of clustering may be crucial to construct any feasible subgrid scheme.

The method of cluster-weighted subgrid modelling is more general than just a refinement or improvement of the conditional Markov chain scheme of Crommelin and Vanden-Eijnden (2008). Different clustering algorithms can be combined with various local statistical models. The method has also been used to construct a closure for a low-order model of atmospheric low-frequency variability based on empirical orthogonal functions (EOFs) (Kwasniok, 2011c).

The present approach is purely data-driven and not based on physical considerations. This may be a strength as well as a weakness. Empirical schemes are potentially more accurate as they are free from constraining a priori assumptions. On the other hand, data-based models are sometimes criticised as not helping with our understanding of the physics of the system. This drawback is here mitigated by the transparent architecture of cluster-weighted modelling. The local models have meaningful and interpretable parameters. Indeed, the clusters here represent phases of an oscillation in $(X_k, \hat{B}_k)$-space (Kwasniok, 2011a). This gives some hope that clusters could potentially be linked to physical processes when the technique was applied to a more realistic system.

There might be potential for improvement in combining predictive, purely data-driven subgrid schemes like the present approach with parametrisation schemes based more on physical reasoning or stochastic dynamical systems theory. Approaches like Majda et al. (1999, 2003) are able to derive the structural form of the closure model for a given system. This information might be used to guide the choice of statistical model or place a priori constraints on the parameters.

References

Berner, J., F. J. Doblas-Reyes, T. N. Palmer, G. Shutts, and A. Weisheimer (2008). Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model. *Phil. Trans. R. Soc. A*, **366**, 2559–2577.

Buizza, R. (1997). Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99–119.

Buizza, R., M. J. Miller, and T. N. Palmer (1999). Stochastic simulation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **125**, 2887–2908.

Chorin, A. J., A. P. Kast, and R. Kupferman (1998). Optimal prediction of underresolved dynamics. *Proc. Natl. Acad. Sci. USA*, **95**, 4094–4098.

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

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

Egger, J. (1981). Stochastically driven large-scale circulations with multiple equilibria. *J. Atmos. Sci.*, **38**, 2606–2618.

Fatkullin, I. and E. Vanden-Eijnden (2004). A computational strategy for multiscale systems with applications to Lorenz 96 model. *J. Comput. Phys.*, **200**, 605–638.

Frankignoul, C. and K. Hasselmann (1977). Stochastic climate models. Part II. Application to sea-surface temperature anomalies and thermocline variability. *Tellus*, **29**, 289–305.

Gershenfeld, N., B. Schoner, and E. Metois (1999). Cluster-weighted modelling for time series analysis. *Nature*, **397**, 329–332.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Hasselmann, K. (1976). Stochastic climate models. Part I. Theory. *Tellus*, **28**, 473–485.

Kwasniok, F. (2011a). Data-based stochastic subgrid-scale parametrisation: an approach using cluster-weighted modelling. *Phil. Trans. R. Soc. A*, in press.

Kwasniok, F. (2011b). Cluster-weighted time series modelling based on minimising predictive ignorance, submitted.

Kwasniok, F. (2011c). Nonlinear stochastic low-order modelling of atmospheric low-frequency variability using a regime-dependent closure scheme, submitted.

Lemke, P. (1977). Stochastic climate models. Part III. Application to zonally averaged energy models. *Tellus*, **29**, 385–392.

Lin, J. W.-B. and J. D. Neelin (2000). Influence of a stochastic moist convective parametrization on tropical climate variability. *Geophys. Res. Lett.*, **27**, 3691–3694.

Lorenz, E. N. (1996). Predictability – a problem partly solved. In *Proc. ECMWF Seminar on Predictability*, vol. 1, pp. 1–18, Reading, UK, ECMWF.

Majda, A. J. and B. Khouider (2002). Stochastic and mesoscopic models for tropical convection. *Proc. Natl. Acad. Sci. USA*, **99**, 1123–1128.

Majda, A. J., I. Timofeyev, and E. Vanden-Eijnden (1999). Models for stochastic climate prediction. *Proc. Natl. Acad. Sci. USA*, **96**, 14687–14691.

Majda, A. J., I. Timofeyev, and E. Vanden-Eijnden (2003). Systematic strategies for stochastic mode reduction in climate. *J. Atmos. Sci.*, **60**, 1705–1722.

Palmer, T. N. (2001). A nonlinear dynamical perspective on model error: a proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Q. J. Roy. Meteor. Soc.*, **127**, 279–304.

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung, and M. Leutbecher (2005). Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, **33**, 163–193.

Plant, R. S. and G. C. Craig (2008). A stochastic parameterization for deep convection based on equilibrium statistics. *J. Atmos. Sci.*, **65**, 87–105.

Sauer, T., J. A. Yorke, and M. Casdagli (1991). Embedology. *J. Stat. Phys.*, **65**, 579.

Shutts, G. (2005). A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. Roy. Meteor. Soc.*, **131**, 3079–3102.

Weisheimer, A., T. N. Palmer, and F. J. Doblas-Reyes (2011). Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles. *Geophys. Res. Lett.*, **38**, L16703.

Wilks, D. S. (2005). Effects of stochastic parametrizations in the Lorenz '96 system. *Q. J. Roy. Meteor. Soc.*, **131**, 389–407.