Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling

Frank Kwasniok

*Phil. Trans. R. Soc. A* 2012 **370**, 1061-1086

References

This article cites 28 articles, 4 of which can be accessed free
http://rsta.royalsocietypublishing.org/content/370/1962/1061.full.html#ref-list-1

Article cited in:
http://rsta.royalsocietypublishing.org/content/370/1962/1061.full.html#related-urls

Subject collections

Articles on similar topics can be found in the following collections

- applied mathematics (57 articles)
- atmospheric science (36 articles)
- statistics (29 articles)

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to *Phil. Trans. R. Soc. A* go to:
http://rsta.royalsocietypublishing.org/subscriptions
A new approach for data-based stochastic parametrization 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 scheme is investigated for long-term simulations as well as ensemble prediction. It clearly outperforms simple parametrization schemes and compares favourably with another recently proposed subgrid modelling scheme also based on conditional Markov chains.

Keywords: stochastic parametrization; model uncertainty; cluster-weighted modelling; ensemble prediction

1. Introduction

The dynamics of weather and climate encompass a wide range of spatial and temporal scales. Owing 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 with 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 parametrization in meteorology and climate science.

Traditionally, parametrizations 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 realization 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 [1]. First implementations of this concept were in the context of sea-surface temperature anomalies [2] and a conceptual zonally averaged climate model [3]. Another early study looked at regime behaviour in a very simple atmospheric model under stochastic forcing [4].

Despite impressive improvements in the forecast skill of numerical weather and climate prediction in the past decades, there are still limitations owing to model uncertainty and error as well as problems in generating initial conditions for ensembles. Forecast ensembles tend to be underdispersive [5], 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 randomization of existing parametrization schemes; approaches include multi-model, multi-parametrization and multi-parameter ensembles [6]. 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 [7] and, most recently, stochastic-dynamic subgrid schemes [8–10]. A general feature of stochastic parametrizations is that they enable the forecast ensemble to explore important regions of phase space better than more restricted deterministic parametrizations. See Palmer et al. [6] and Weisheimer et al. [11] 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 [12], to a systematic stochastic mode reduction strategy based on stochastic differential equations [13,14], to various approaches to stochastic convection parametrization [15–17]. A particular class of subgrid models are schemes that are derived purely from data [18,19]. 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 parametrization 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 [19].

The paper is organized as follows: §2 introduces the general framework of cluster-weighted modelling for subgrid parametrization. In §3, we describe the Lorenz ’96 (L96) system, which is here used as a testbed to explore the method.
The detailed formulation of the subgrid parametrization in the context of the L96 system and how to estimate its parameters from data is discussed in §4. Then the results are presented in §5. The paper concludes with some general discussion and implications.

2. Subgrid-scale parametrization 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

\[
\dot{\mathbf{x}} = \mathbf{R}(\mathbf{x}) + \mathbf{U}(\mathbf{x}, \mathbf{y})
\] (2.1)

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 parametrize \( \mathbf{U}(\mathbf{x}, \mathbf{y}) \). Such a parametrization has the general form

\[
\mathbf{U}(\mathbf{x}, \mathbf{y}) \sim f(\mathbf{x}) + \eta(\mathbf{x}),
\] (2.2)

where \( f(\mathbf{x}) \) 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:

\[
f(\mathbf{x}) = \langle \mathbf{U}(\mathbf{x}, \mathbf{y}) | \mathbf{x} \rangle.
\] (2.3)

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 [20,21], 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 [22]. Each cluster is associated with a local probabilistic subgrid model \( p(\eta | \mathbf{v}, c = m) \) that 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 \( \eta \). The conditional probability density of the stochastic subgrid term \( \eta \) is expanded into a sum over the clusters:

\[
p(\eta | \mathbf{z}, \mathbf{v}) = \sum_{m=1}^{M} g_m(\mathbf{z}) \, p(\eta | \mathbf{v}, c = m).
\] (2.4)

*Phil. Trans. R. Soc. A* (2012)
The state-dependent weights $g_m$ of the individual models are given by Bayes’ rule:

$$g_m(z) = p(c = m | z) = \frac{w_m p(z | c = m)}{\sum_{n=1}^{M} w_n p(z | c = n)}.$$  \hspace{1cm} (2.5)

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 $z$ and the explicit dependence of the subgrid models on $v$. The vectors $z$ and $v$ might overlap.

The clustering densities $p(z | c = m)$ and the local subgrid models $p(\eta | v, c = m)$ can take various forms. The canonical choice for the clustering densities $p(z | c = m)$ 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 $z$ into a finite number of bins; the clustering probabilities are then discrete probability distributions over these bins. The subgrid models $p(\eta | v, c = m)$ may be regression models on $v$ with Gaussian uncertainty. In the present study, they are actually Markov chains governing the switching between discrete values of $\eta$.

The parameters of the clusters and the subgrid models are estimated simultaneously from a learning dataset by maximizing 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 minimizing the Akaike or Bayesian information criterion in the learning dataset, or by maximizing the cross-validated likelihood function in a dataset different from the learning dataset. 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.

The present approach is similar to, yet different from the optimal prediction scheme [12]. The optimal prediction scheme derives deterministic evolution equations for moments of coarse-grained variables where the dynamics is given by averaging over the invariant measure of the system constrained by the state of the coarse-grained variables. The main idea of the cluster-weighted scheme is the synthesis of simple local models to build a complex global model. It results in a stochastic subgrid model as it involves sampling from the probability density of $\eta$ at each step.

### 3. The Lorenz ’96 model

The L96 model [23] is used as a testbed to explore the new subgrid parametrization scheme. It has become popular in the weather and climate science community as a toy model of the atmosphere to test concepts and algorithms relating to predictability, model error, ensemble post-processing and subgrid parametrization [8,18,19,23,24]. The model equations are

$$\dot{X}_k = X_{k-1}(X_{k+1} - X_{k-2}) - X_k + F + B_k \hspace{1cm} (3.1)$$

and

$$\dot{Y}_{j,k} = \frac{1}{\varepsilon} [Y_{j+1,k}(Y_{j-1,k} - Y_{j+2,k}) - Y_{j,k} + h_Y X_k] \hspace{1cm} (3.2)$$
with

\[ B_k = \frac{h_x}{J} \sum_j Y_{j,k} \]  

(3.3)

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 with 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 [24] is exactly equivalent to the original formulation by Lorenz [23]. With \( X^*_{k} \) and \( Y^*_{j,k} \) denoting the variables in the original system [23] with parameters \( F, h, c \) and \( b \), the corresponding system in the formulation of equations (3.1)–(3.3) is obtained by a linear scaling of the variables \( (X_k = X^*_k \) and \( Y_{j,k} = bY^*_{j,k} \) and the parameter setting \( \varepsilon = 1/c, \ h_x = -hcJ/b^2 \) 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 \( \varepsilon \). 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 & Vanden-Eijnden [19]. 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 parametrized somehow in a reduced model in order to account for the impact of the unresolved variables. This constitutes the subgrid-scale parametrization problem in the context of the L96 model. The present choice \( \varepsilon = 0.5 \) corresponds to only very weak time-scale separation between resolved and unresolved variables, which is more realistic for the real atmosphere and more challenging for parametrizations than the more common choice \( \varepsilon = 0.1 \) [18,23,24] corresponding to time-scale separation by an order of magnitude.

Figure 1a 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 per cent of the variance of the subgrid term \( B_k \). The properties of the conditional probability density function (PDF) \( p(B_k | X_k) \) depend strongly on \( X_k \). In particular, it is markedly non-Gaussian for a range of values of \( X_k \). Figure 1b 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 §2, a cluster-weighted subgrid scheme based on local Markov chains is developed and implemented for the L96 model.

(a) Model formulation

We here combine the framework of cluster-weighted modelling [20,21] with the use of conditional Markov chains [19] for stochastic subgrid-scale parametrization. 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 [18]. 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 $\delta t$ is the sampling interval of the data. This choice is motivated by the fact that the PDF of the subgrid term $B_k$ has been shown to depend also on the increment $\delta X_k$ [19]. 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 dataset 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 \(\alpha\) refers to time in an equally sampled time series with sampling interval \(\delta t\) 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 \(\{I^X_i\}_{i=1}^{N_X}\); we have \(s = i\) if \(X_k \in I^X_i\). The \(\delta X_k\)-space is divided into \(N_{\delta X}\) disjoint intervals \(\{I_{\delta X}^j\}_{j=1}^{N_{\delta X}}\); we have \(d = j\) if \(\delta X_k \in I_{\delta X}^j\). Given \(s = i\), the range of possible values of \(\hat{B}_k\) is divided into \(N_B\) disjoint, equally populated intervals \(\{I_{\hat{B}}^l\}_{l=1}^{N_B}\); we have \(b = l\) if \(\hat{B}_k \in I_{\hat{B}}^l\). 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:

\[
\beta_{il} = \frac{\sum_{\alpha} \hat{B}_k^\alpha 1(s_{\alpha} = i) 1(b_{\alpha} = l)}{\sum_{\alpha} 1(s_{\alpha} = i) 1(b_{\alpha} = l)}. \quad (4.1)
\]

We introduce \(M\) clusters in the discrete \((s, d, b)\)-space. Each cluster has an overall weight or probability of that cluster being chosen,

\[
w_m = p(c = m), \quad (4.2)
\]

and a clustering probability distribution

\[
\psi_{mij} = p(s = i, d = j | c = m), \quad (4.3)
\]

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 \(\sum_{ij} \psi_{mij} = 1\). 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 \(\rho_{ij}\) is empirically given as the fraction of data points in these bins: \(\rho_{ij} = (1/N) \sum_{\alpha} 1(s_{\alpha} = i) 1(d_{\alpha} = j)\). It follows that the clusters also sum up to the marginal climatological distributions:

\[
\sum_{m,j} w_m \psi_{mij} = p(s = i) = \sum_j \rho_{ij} \quad \text{as well as} \quad \sum_{m,i} w_m \psi_{mij} = p(d = j) = \sum_i \rho_{ij}.
\]

Each cluster is associated with a Markov chain in the discrete space \(b\) described by an \((N_B \times N_B)\) transition matrix \(A_m\) with components

\[
A_{mlb} = p(b_{\alpha} = l | b_{\alpha-1} = l_1, c_{\alpha} = m). \quad (4.4)
\]

The matrices \(A_m\) are row-stochastic matrices, that is, \(A_{mlb} \geq 0\) and \(\sum_l A_{mlb} = 1\).

The conditional probability distribution for \(b_{\alpha}\) is modelled as a sum over the clusters:

\[
p(b_{\alpha}|b_{\alpha-1}, s_{\alpha}, d_{\alpha}) = \sum_{m=1}^{M} g_m(s_{\alpha}, d_{\alpha}) A_{mb_{\alpha-1}b_{\alpha}}. \quad (4.5)
\]
The state-dependent model weights are given by Bayes’ rule as
\[ g_m(i, j) = p(c = m | s = i, d = j) = \frac{w_m \psi_{mij}}{\sum_{n=1}^{M} w_n \psi_{nij}} = \frac{w_m \psi_{mij}}{p_{ij}}. \quad (4.6) \]

The Markov chain is effectively governed by local transition matrices \( A^{\text{loc}}(i, j) = \sum_m g_m(i, j) A_m \), 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 \( \{ \beta_d \}_{i=1}^{N_B} \) given by equation (4.1) for \( s = i \). The mean local model weights are found to be \( \langle g_m \rangle = \frac{1}{N} \sum_a g_m(s_a, d_a) = w_m \). Hence, the overall weight \( w_m \) can be interpreted as the fraction of the dataset (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 \( N_d X \) 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.

Modifications in the specification of the subgrid model are possible. The discrete values \( \beta \) representing the subgrid term may be chosen to depend also on \( d \) rather than only on \( s \); or they may be chosen as a global set of values, depending neither on \( s \) nor on \( d \). One may also add a further layer of stochasticity by replacing the constant values \( \beta \) with, for example, Gaussian outputs whose means and variances are to be determined from the data together with the other parameters of the subgrid model (cf. [25]).

(b) Parameter estimation

Given an equally sampled learning dataset 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-maximization (EM) algorithm [26,27]. The likelihood function of the data in the CWMC model setting as specified in §4a is
\[ L = p(b_1, \ldots, b_N | b_0, s_1, d_1, \ldots, s_N, d_N) = \prod_{\alpha} p(b_{\alpha} | b_{\alpha-1}, s_{\alpha}, d_{\alpha}) \]
\[ = \prod_{\alpha} \sum_m g_m(s_{\alpha}, d_{\alpha}) A_{mb_{\alpha-1}b_{\alpha}} \quad (4.7) \]

taking into account the Markovian property of the subgrid model and the conditioning only on \( s_{\alpha} \) and \( d_{\alpha} \). The EM algorithm is iterative. Having arrived at parameter estimates \( w_m^{(r)}, \psi_{mij}^{(r)} \) and \( A_{mlb}^{(r)} \), the expectation step of the \( r \)th iteration consists of calculating the posterior probabilities
\[ \delta_{ma}^{(r)} = p(c_{\alpha} = m | b_0, s_1, d_1, b_1, \ldots, s_N, d_N, b_N) \]
\[ = p(c_{\alpha} = m | b_{\alpha-1}, s_{\alpha}, d_{\alpha}, b_{\alpha}) = \frac{w_m^{(r)} \psi_{ms_{\alpha}d_{\alpha}}^{(r)} A_{mb_{\alpha-1}b_{\alpha}}^{(r)}}{\sum_n w_n^{(r)} \psi_{ns_{\alpha}d_{\alpha}}^{(r)} A_{nb_{\alpha-1}b_{\alpha}}^{(r)}}. \quad (4.8) \]
In the maximization step, the parameters are updated according to the following re-estimation formulae [21]:

\[ w_m^{(r+1)} = \frac{1}{N} \sum_\alpha \delta_{ma}^{(r)}, \quad (4.9) \]

\[ \psi_{mij}^{(r+1)} = \frac{\sum_\alpha \delta_{ma}^{(r)} 1(s_\alpha = i) 1(d_\alpha = j)}{\sum_\alpha \delta_{ma}^{(r)}} \quad (4.10) \]

and

\[ A_{ml_1l_2}^{(r+1)} = \frac{\sum_\alpha \delta_{ma}^{(r)} 1(b_\alpha - 1 = l_1) 1(b_\alpha = l_2)}{\sum_\alpha \delta_{ma}^{(r)}}. \quad (4.11) \]

Initial guesses for the parameters \( w_m^{(0)}, \psi_{mij}^{(0)}, \text{ and } A_{ml_1l_2}^{(0)} \) need to be provided to start the iteration. The algorithm is monotonically non-decreasing in likelihood and converges to a maximum of the likelihood function \( L \). The algorithm ensures that the parameter estimates satisfy all the constraints outlined in §4a given the initial guesses to satisfy them. The likelihood landscape may be complicated and have multiple local maxima. It is therefore advisable to run the algorithm with an ensemble of different randomly drawn (subject to the constraints) initial guesses for the parameters.

(c) 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 \( \delta t \); 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, \( \delta t \) is larger than \( h \). Assume for simplicity that \( \delta t \) is an integer multiple of \( h \): \( \delta t = N_{\text{step}} h \). We then use a split-integration scheme [19]. The resolved dynamics are integrated with time step \( h \); the subgrid model is propagated with time step \( \delta t \), updated only every \( N_{\text{step}} \) 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 \( X_{k}^{\alpha} \) at time \( t_{\alpha} \) is calculated by propagating the resolved variables \( N_{\text{step}} \) times with step size \( h \) using the derivative given by equation (3.1) with \( B_k \) set to \( \langle B_k | X_{k}^{\alpha-1} \rangle + \beta_{s_{\alpha-1}b_{\alpha-1}} \). If \( X_{k}^{\alpha} \) falls in bin \( s_\alpha \) and \( \delta X_{k}^{\alpha} = X_{k}^{\alpha} - X_{k}^{\alpha-1} \) falls in bin \( d_\alpha \), the next state of the Markov chain \( b_\alpha \) is determined by randomly drawing from the probability distribution given by equations (2.4) and (2.5). 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 \( \delta t \) 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 dataset containing 100 000 data points. The CWMC closure scheme is constructed from this dataset. Such a large learning dataset, 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 dataset (approx. 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 $I_{X1} = (-\infty, -1.5]$, $I_{X2} = (-1.5, 2.5]$, $I_{X3} = (2.5, 6.5]$ and $I_{X4} = (6.5, \infty)$. In $dX$-space, we use $N_{dX} = 2$ intervals given as $I_{dX1} = (-\infty, 0]$ and $I_{dX2} = (0, \infty)$, 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 $\beta_{kl}$ used in the CWMC scheme given by equation (4.1) are displayed in figure 1b. The resolution of the binnings was determined from the performance of the resulting reduced model. We studied larger values for all of the parameters $N_X$, $N_{dX}$ and $N_B$, but a higher resolution in the binning of any of the variables does not visibly improve the model. The CWMC closure schemes were estimated from the dataset 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 two clusters.

In order to assess the complexity of the likelihood surface and check for the possible existence of multiple extrema, the EM algorithm was run 20 times from randomly chosen initial parameters. With $M = 2$, all initial conditions ended up with the same clusters and parameters, giving a strong indication that the global maximum of the likelihood function has been found.

The CWMC scheme is compared with two simple generic parametrization schemes: a deterministic closure scheme and the AR(1) scheme proposed in Wilks [18]. The deterministic scheme consists of parametrizing $B_k$ by the conditional mean as estimated by the fifth-order polynomial fit shown in figure 1a: $\hat{B}_k \sim \langle B_k|X_k \rangle$. The AR(1) scheme models $B_k$ by the conditional mean plus an AR(1) process:

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

where $\xi$ denotes Gaussian white noise with zero mean and unit variance and $\sigma$ 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 $\sigma/\sqrt{1-\phi^2} = 0.866$, 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 §4c for the CWMC scheme, updating the subgrid term at time step $\delta t$.

The CWMC scheme is also compared with the subgrid modelling study by Crommelin & Vanden-Eijnden [19] based on conditional Markov chains using the L96 system with exactly the same parameter setting as an example. They
condition the Markov chain on $X_k^a$ and $X_k^{a-1}$, both partitioned into 16 bins. Taking into account that owing to the autocorrelation of the system at lag $\delta t$ 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 dataset is necessary to estimate the full Markov chain model.

(a) Clusters

In the L96 model with the present parameter settings, a state variable $X_k$ and the corresponding subgrid term $\hat{B}_k$ exhibit a particular dynamical structure. The time evolution of a point in the $(X_k, \hat{B}_k)$-plane follows a clockwise noisy oscillation [19]. When $X_k$ increases, $\hat{B}_k$ is likely to take a noisy path through the upper part of the cloud of points in figure 1b. When $X_k$ decreases, $\hat{B}_k$ tends to evolve in the lower part of the cloud.

Figure 2 shows the probability distributions of the two clusters in $s$- and $d$-spaces together with the (marginal) climatological distributions of $s$ and $d$. Figure 3 displays the local model weights $g_m$ as a function of $s$ for $\delta X_k \leq 0$ ($d = 1$) and $\delta X_k > 0$ ($d = 2$). The transition matrices of the Markov chain associated with the clusters are

\[
A_1 = \begin{pmatrix}
0.998 & 0.002 & 0.000 \\
0.044 & 0.956 & 0.000 \\
0.000 & 0.074 & 0.926
\end{pmatrix},
\]

and

\[
A_2 = \begin{pmatrix}
0.878 & 0.122 & 0.000 \\
0.002 & 0.914 & 0.084 \\
0.000 & 0.000 & 1.000
\end{pmatrix}.
\]

The states are arranged in decreasing order of the values $\{\beta_d\}_{d=1}^{N_B}$. The two clusters can be interpreted as different phases of the oscillation in the $(X_k, \hat{B}_k)$-plane. The first cluster with an overall weight of $w_1 = 0.605$ is mainly localized at small and medium values of $X_k$ and at $\delta X_k > 0$. For small $X_k$ ($s = 1$), it has virtually all the model weight regardless of the sign of $\delta X_k$. The associated transition matrix of the Markov chain attracts the subgrid term to the upper part of the scatterplot, in particular the first state, and keeps it there. This cluster drives the trajectory into the clockwise direction towards large values of $\hat{B}_k$ when at small $X_k$ and keeps it in the upper part of the cloud when $X_k$ increases. The second cluster with a weight of $w_2 = 0.395$ concentrates on medium and large values of $X_k$ and on $\delta X_k \leq 0$, having a large local model weight there. The transition matrix strongly attracts and keeps the subgrid term in the third state (lowest value for $\hat{B}_k$). The cluster captures the
phase of the oscillation in the lower part of the scatterplot when $X_k$ is decreasing. The transition phase from large to small values of $\hat{B}_k$ at large $X_k$ ($s = 4$) when the trajectory reverses its direction is described by a superposition of both clusters.

(b) 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. The reduced models are integrated in time as described in §4c 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 the 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 250 000. All the results reported below are calculated from these time series.

Figure 4 shows the PDF 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 $D = \max_{X_k} |\Phi(X_k) - \Phi_r(X_k)|$, where $\Phi$ is the (cumulative) probability distribution of the L96 model and $\Phi_r$ 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 [19]. 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 5. 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) wavevector $\mathbf{G} = (G_0, \ldots, G_{K-1})$ with $G_0 = G_{K-1}^*$ for $\nu = 1, \ldots, K - 1$. The wave variance of wavenumber $\nu$ is then $\langle |G_\nu - \langle G_\nu \rangle|^2 \rangle$.
Figure 4. Probability density function of $X_k$ in the full L96 model (solid line) and in reduced models with deterministic subgrid scheme (dotted-dashed line), AR(1) scheme (dotted line) and the CWMC scheme (dashed line).

Figure 5. Wave variances in the full L96 model (solid line) and in reduced models with deterministic subgrid scheme (dotted-dashed line), AR(1) scheme (dotted line) and the CWMC scheme (dashed line).

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 models and that in the full L96 model.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>s.d.</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>full L96 model</td>
<td>2.39</td>
<td>3.52</td>
<td></td>
</tr>
<tr>
<td>deterministic scheme</td>
<td>2.53</td>
<td>3.56</td>
<td>0.017</td>
</tr>
<tr>
<td>AR(1) scheme</td>
<td>2.51</td>
<td>3.57</td>
<td>0.015</td>
</tr>
<tr>
<td>CWMC scheme</td>
<td>2.40</td>
<td>3.51</td>
<td>0.004</td>
</tr>
</tbody>
</table>

for $v = 0, \ldots, K - 1$, where $\langle \cdot \rangle$ denotes the time average. Figure 6 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 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 [19]. 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 owing 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 7 gives the autocorrelation function of $X_k$ and figure 8 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 with 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 [19].

(c) Ensemble prediction

We investigate the predictive skill of the reduced models with the different parametrizations. Given the stochastic nature of the models, an ensemble prediction framework appears to be most appropriate. Two different ensemble settings are considered.

— Firstly, we construct ensembles which sample only model uncertainty. All ensemble members have the same (unperturbed) initial condition and differ only in the realizations of the stochastic subgrid model. In this setting, the deterministic subgrid scheme provides only a single best-guess forecast.

— Secondly, we construct ensembles that 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 & Vanden-Eijnden [19]. 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 in Wilks [18]) or identify fastest growing perturbations using singular vectors.

In principle, the first ensemble setting, sampling only model uncertainty and not perturbing the initial condition, seems to reflect more accurately the situation here as the initial state of the resolved variables $X_k$ is exactly known and there is no genuine initial condition uncertainty. However, it seems worth also looking at perturbed initial conditions as these are routinely used in operational ensemble
weather prediction. Moreover, model and initial condition uncertainty may not be orthogonal, but overlap and be entangled. Including initial condition uncertainty may help in accounting for model uncertainty and error even if there is no genuine initial condition uncertainty. For both ensemble settings, 10,000 initial conditions are used taken from a long integration of the full L96 model, separated by five time units.

For prediction experiments, the question how to initialize the stochastic part of the subgrid model becomes potentially important for the prediction skill. Let $t_0$ be the point in time at which the forecast is started. For the AR(1) subgrid scheme, $\hat{B}_k(t_0)$ is initialized by randomly drawing, independently for each ensemble member, from the stationary probability density of the AR(1) process, that is, from a Gaussian with zero mean and variance $\sigma^2/(1 - \phi^2)$. For the CWMC scheme, the somehow canonical initialization of the Markov chain would be to randomly draw, independently for each ensemble member, from an equidistribution over the $N_B$ possible states $\{\hat{\beta}_i\}_{i=1}^{N_B}$ corresponding to $s(t_0)$. We here use a refined procedure which is still easy to implement; it is actually similar to common practice in data assimilation for numerical weather prediction. An assimilation time window of length $\theta$ is chosen prior to initial time $t_0$. The state of the Markov chain at time $t_0 - \theta$ is randomly drawn from an equidistribution over the $N_B$ possible values corresponding to $s(t_0 - \theta)$. Then the Markov chain is simulated, independently for each ensemble member, in the assimilation interval $[t_0 - \theta, t_0]$ using the known trajectory of the system in $(s, d)$-space. This builds up a probability distribution of the Markov chain at initial time $t_0$, dynamically constrained by the past values of $s$ and $d$ in the assimilation window, from which the ensemble is drawn. The length of the assimilation window is chosen to be $\theta = 0.1 = 10\delta t$ corresponding to 10 steps of the Markov chain. This state-dependent initialization turns out to visibly improve the prediction skill on the canonical initialization for short and medium lead times.

In order to assess the ensemble spread, we use rank histograms [28]. Rank histograms give the relative frequency of the ranks of the true state in the $N_{\text{ens}} + 1$ 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. Figures 9 and 10 show rank histograms for the different parametrization schemes in the two ensemble settings. The ensemble size is $N_{\text{ens}} = 20$; 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, the rank histograms are nearly uniform in both ensemble settings apart from some bias the origin of which is unclear. The rank histograms for the AR(1) scheme are in accordance with the findings in Wilks [18]. Crommelin & Vanden-Eijnden [19] report a substantial underdispersion for the AR(1) scheme. It is not clear where this discrepancy comes from; it may be due to a different initialization of the AR(1) process at initial time $t_0$. For the CWMC model, there is a slight underdispersion of the ensembles without perturbation of the initial condition and an almost ideal ensemble spread when perturbing the initial condition.
Figure 9. Rank histograms for ensembles accounting only for model uncertainty with (a) the AR(1) scheme and (b) the CWMC scheme. Prediction lead time is $\tau = 2$; ensemble size is $N_{\text{ens}} = 20$. The dashed horizontal lines indicate the expected relative frequency under rank uniformity.

Figure 10. Rank histograms for ensembles accounting for both model and initial condition uncertainty with (a) the deterministic subgrid scheme, (b) the AR(1) scheme and (c) the CWMC scheme. Prediction lead time is $\tau = 2$; ensemble size is $N_{\text{ens}} = 20$. The dashed horizontal lines indicate the expected relative frequency under rank uniformity.

We now evaluate the actual predictive skill of the forecasts with the various subgrid schemes. Firstly, we consider the deterministic forecast given by the ensemble mean. Figures 11 and 12 provide the anomaly correlation and the root mean square error with the three closure schemes in the two ensemble settings for ensemble sizes $N_{\text{ens}} = 5$, $N_{\text{ens}} = 20$ and $N_{\text{ens}} = 50$. For all schemes, the prediction skill improves monotonically with the ensemble size for all lead times and is virtually converged at $N_{\text{ens}} = 50$. For the AR(1) and CWMC schemes, the predictive skill is virtually the same in both ensemble settings; the skill of the deterministic closure is improved by perturbing the initial condition. The CWMC scheme clearly outperforms the two other schemes at all lead times; with $N_{\text{ens}} = 5,$
Figure 11. Prediction skill of the ensemble mean for ensembles accounting only for model uncertainty. (a) Anomaly correlation with the AR(1) scheme (dashed line) and the CWMC scheme (solid line). Curves from bottom to top refer to ensemble sizes $N_{\text{ens}} = 5$, $N_{\text{ens}} = 20$ and $N_{\text{ens}} = 50$. The dotted curve gives the anomaly correlation with the deterministic scheme. (b) Root mean square error with the AR(1) scheme (dashed line) and the CWMC scheme (solid line). Curves from top to bottom refer to ensemble sizes $N_{\text{ens}} = 5$, $N_{\text{ens}} = 20$ and $N_{\text{ens}} = 50$. The dotted curve gives the root mean square error with the deterministic scheme. The solid horizontal line indicates the root mean square error of the climatology forecast.

Figure 12. Prediction skill of the ensemble mean for ensembles accounting for both model and initial condition uncertainty. (a) Anomaly correlation with the deterministic scheme (dotted line), the AR(1) scheme (dashed line) and the CWMC scheme (solid line). Curves from bottom to top refer to ensemble sizes $N_{\text{ens}} = 5$, $N_{\text{ens}} = 20$ and $N_{\text{ens}} = 50$. (b) Root mean square error with the deterministic scheme (dotted line), the AR(1) scheme (dashed line) and the CWMC scheme (solid line). Curves from top to bottom refer to ensemble sizes $N_{\text{ens}} = 5$, $N_{\text{ens}} = 20$ and $N_{\text{ens}} = 50$. The solid horizontal line indicates the root mean square error of the climatology forecast.

it is already better than the two other schemes with $N_{\text{ens}} = 50$. The AR(1) scheme cannot consistently outperform the deterministic scheme. The CWMC subgrid model is not worse than the full Markov chain scheme [19]; it is even better at small and medium lead times, probably owing to the state-dependent initialization of the Markov chain at initial time $t_0$.

The full probabilistic information in the ensembles propagated under the various models is assessed by making probabilistic predictions of $X_k$ being in each of five equiprobable classes defined by the quintiles of the climatological distribution. The forecast probabilities are given by the fraction of ensemble members falling in each class; no post-processing of the ensemble is applied.
The forecasts are evaluated using the ranked probability score and the ignorance score \cite{29,30}. The ignorance score becomes singular if zero probability is assigned to a class which is actually observed. A regularization of the score is applied to avoid this problem. If no ensemble member falls into a particular class, a probability of $10^{-4}$ is assigned to that class and the forecast distribution renormalized.

Figures 13 and 14 give the scores in the two ensemble settings for ensemble sizes $N_{\text{ens}} = 5$, $N_{\text{ens}} = 20$ and $N_{\text{ens}} = 50$. Climatology, defined as a uniform distribution over all the classes, is indicated as a reference forecast. For all closure schemes, the prediction skill improves monotonically with ensemble size at all lead times. The CWMC scheme clearly outperforms the deterministic and the AR(1) scheme in both ensemble settings at all ensemble sizes and all lead times. In the ignorance score, the deterministic scheme cannot compete at all due to its
heavy underdispersion, leading to systematically overconfident forecasts, which is strongly penalized by the ignorance score. All subgrid models have a rather poor ignorance score at small ensemble size. This is due to the insufficient resolution of the five classes with few ensemble members which leads to systematically overconfident forecasts.

**Prediction of extremes**

We also investigate the ability of the reduced models to predict extreme events in the L96 system. An extreme event is here defined as \( X_k \) exceeding the 95%-quantile of its climatological distribution. The forecast probability of an extreme event is given by the fraction of ensemble members falling in the extremal range; no post-processing of the ensemble is performed. Prediction lead time is \( \tau = 2 \). Figures 15–18 provide the reliability diagram and the relative operating characteristic (ROC) for the two ensemble settings. The reliability diagram displays the relative frequency of the extreme event conditional on the predicted probability. An ROC curve refers to deterministic predictions of the extreme event. An event is forecast if at least \( N_{thr} \) ensemble members fall into the extremal range. Then the threshold parameter \( N_{thr} \) is varied from 0 to \( N_{ens} + 1 \). The reliability diagrams are calculated for an ensemble size \( N_{ens} = 20 \); the ROC curves are based on \( N_{ens} = 50 \).

The predictions with the deterministic closure scheme are overconfident for most of the range of forecast probabilities and the extreme event is only poorly resolved. For example, in cases where almost all of the ensemble members fall into the extremal range, the actual relative frequency of an extreme event is only about 0.5–0.6. The AR(1) scheme provides a substantial improvement. These results are in accordance with those reported in Wilks [18]. The CWMC scheme enables clearly more reliable predictions of the extreme event. With unperturbed initial
Figure 16. Relative operating characteristic (ROC) for ensemble predictions of $X_k$ exceeding the 95%-quantile with the AR(1) scheme (dashed line) and the CWMC scheme (solid line). Prediction lead time is $\tau = 2$; ensemble size is $N_{\text{ens}} = 50$. Ensembles account only for model uncertainty.

Figure 17. Reliability diagram for ensemble predictions of $X_k$ exceeding the 95%-quantile with the deterministic scheme (dotted line), the AR(1) scheme (dashed line) and the CWMC scheme (solid line). Prediction lead time is $\tau = 2$; ensemble size is $N_{\text{ens}} = 20$. Ensembles account for both model and initial condition uncertainty.

conditions it has almost perfect behaviour over the whole range of the reliability diagram, showing only slight overforecasting at high forecast probabilities; with perturbed initial conditions it is perfectly reliable over the whole range of forecast probabilities. In forecast skill as measured by the ROC curves, the CWMC scheme substantially outperforms the deterministic and the AR(1) schemes.
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 L96 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 CWMC scheme performs about as well as the conditional Markov chain scheme proposed by Crommelin & Vanden-Eijnden [19] 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 discretization. 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 \( z \) on which one would like to condition the subgrid scheme is likely to be of higher dimension than in the L96 model. A conditioning based on binning into disjoint intervals as in Crommelin & Vanden-Eijnden [19] 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 & Vanden-Eijnden [19]. 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) [25].

It needs to be investigated how to best choose the clustering variables $z$ in a more realistic system. Two approaches are conceivable. Firstly, thinking of a hierarchy of scales in the climate system, a closure model at the truncation level could be conditional on the next higher properly resolved local scale. On a grid, the parametrization at a particular grid point would then be conditional on the neighbouring grid points. Secondly, the parametrization at all grid points could be conditional on global patterns occurring at the largest spatial scale of the system. These could be identified by EOF analysis. In an atmospheric model, they might be the well-known teleconnection patterns such as the Pacific North America pattern or the North Atlantic Oscillation.

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 criticized 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, as shown in §5a, the clusters represent phases of an oscillation in $(X_k, \hat{B}_k)$-space. 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 parametrization schemes based more on physical reasoning or stochastic dynamical systems theory. Approaches similar to that of Majda and co-workers [13,14] 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.

Another issue with data-based models like the present one is their limitation to stationary systems. Only features which have occurred in the learning dataset can be modelled reliably. It is conceivable that in a climate model under climate change the relationships between resolved and unresolved variables change. The extent to which the subgrid model would be able to extrapolate correctly is then unclear. Possible remedies for this potential drawback are the application of the technique in a sliding time window of appropriate size or the online update of the estimated closure model with new data at a certain learning rate. Both approaches have the effect of gradually discarding data from the distant past in favour of new data and should be able to track at least slow changes in the system.

References

Data-based stochastic subgrid modelling


*Phil. Trans. R. Soc. A* (2012)


