Stochastic Climate Theory

Georg A. Gottwald[1], Daan T. Crommelin[2,3], and Christian L. E. Franzke[4]

[1] School of Mathematics and Statistics, The University of Sydney, Sydney, Australia
[2] CWI, Amsterdam, The Netherlands
[3] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands
[4] Meteorological Institute and Center for Earth System Research and Sustainability, University of Hamburg, Hamburg, Germany

Abstract

In this chapter we review stochastic modelling methods in climate science. First we provide a conceptual framework for stochastic modelling of deterministic dynamical systems based on the Mori-Zwanzig formalism. The Mori-Zwanzig equations contain a Markov term, a memory term and a term suggestive of stochastic noise. Within this framework we express standard model reduction methods such as averaging and homogenization which eliminate the memory term. We further discuss ways to deal with the memory term and how the type of noise depends on the underlying deterministic chaotic system. Secondly, we review current approaches in stochastic data-driven models. We discuss how the drift and diffusion coefficients of models in the form of stochastic differential equations can be estimated from observational data. We pay attention to situations where the data stems from multi scale systems, a relevant topic in the context of data from the climate system. Furthermore, we discuss the use of discrete stochastic processes (Markov chains) for e.g. stochastic subgrid-scale modeling and other topics in climate science.

1 Introduction

The climate system is characterized by the mutual interaction of complex systems each involving entangled processes running on spatial scales from millimeters to thousands of kilometers, and temporal scales from seconds to millennia. Given current computer power it is impossible to capture the whole range of spatial and temporal scales and this will also not be possible in the foreseeable future. Depending on the question we pose to the climate system - be it forecasting regimes in the atmosphere or simulating past ice ages - we have to make a decision as to what components to include in the analysis and as to what scales to resolve. A corollary of this decision is that each numerical scheme inevitably fails to resolve so called *unresolved scales* or *subgrid-scales*. However, typically one is only interested at the slow processes active on large spatial scales. For example, for weather forecasts it is sufficient to resolve large scale high and low pressure fields rather than small scale fast oscillations of the stratification surfaces, whereas for climate predictions with a coupled ocean-atmosphere model we may want to distill the slow dynamics of the ocean ignoring weather systems interacting with the ocean on fast time-scales of days.

The dynamics of the unresolved scales, however, have a significant impact on the large scales and simply ignoring them has detrimental effects on reliably simulating the slow large scale variables of interest. For example, Dawson and Palmer (2014) showed that the ECMWF model produces unrealistic spatial patterns of atmospheric weather regimes due to not sufficiently resolving the small-scale processes. This has implications for the current global circulation models used for the intergovernmental Panel on Climate Change fifth assessment report (IPCC AR5) which typically use coarser resolutions.

The question we are concerned with in this chapter is whether it is possible to obtain reliable simulations of the slow large-scale characteristics of the climate system without having to resolve the small fast scales accurately by employing computationally very costly high-resolution simulations but rather parametrise them by judiciously chosen noise, and if so, under what conditions? Heuristically this should be possible in the following situations (Givon *et al.*, 2004): 1.) time scale separation and 2.) weak coupling to a large system.

In a time-scale separated system, during one slow-time unit the fast uninteresting variables y perform many "uncorrelated" events (provided the fast dynamics is sufficiently chaotic). The contribution of the uncorrelated events to the dynamics of the slow interesting variable x is as a sum of independent random variables. By the Central Limit Theorem this can be expressed by a normally distributed variable. Similarly, if a large number of uninteresting variables y are weakly coupled to the resolved interesting variables x, it takes many uncorrelated events of the unresolved variables to have a significant effect on the dynamics of the resolved variables. The resolved variables x experience a cumulative contribution of those events, which again by the Central Limit Theorem allows us to parameterise the unresolved "heat bath" y by a random process. Here the randomness is not mediated by chaotic dynamics and time-scale separation, but by a large number of weakly coupled variables with random initial conditions drawn typically from some thermodynamic equilibrium density.

Note that in both cases, the stochasticity arises only asymptotically, by either infinite time scale separation or by an infinitely large heat bath. Real life applications never satisfy these limits and care has to be taken. DelSole (2000) pointed out, for example, that on short time scales stochastic models are not able to capture deterministic dynamics. Short meaning here that the fast chaotic variables have not sufficiently decorrelated to allow for the central limit theorem to act.

In the climate science community it has been realized that stochasticity may be used to parametrize subgrid-scale phenomena. In climate modeling, the idea of modeling fast chaotic dynamics by stochastic processes and thereby reducing the effective dimension of the full system goes back to the seminal works by Hasselmann (1976) and Leith (1975). In their work Hasselmann (1976) and Leith (1975) have suggested studying climatic regime switches by introducing in an ad-hoc way a stochastic driver for the slow dynamics. Such an approximation describes the deviations from an averaged climatological system. Of course, it is natural to expect such behavior only if the fast variables (eg. weather in a coupled climatic ocean-atmosphere model) are sufficiently chaotic and approximately random.

These ideas have been used to simulate systems of increasing complexity including the barotropic vorticity equation (Duan and Nadiga, 2007; Franzke *et al.*, 2005), a 3-layer quasi-geostrophic prototype climate model (Franzke and Majda, 2006) and a primitive equation model (Zhang and Held, 1999).

The effective dimension reduction achieved if a large number of fast equations are replaced by only a few stochastic process, and the associated computational advantage of such a reduction is a huge driving force behind this research. Such reduction strategies also provide insight into the underlying dynamics of the climate system and pose new challenging mathematical problems.

The "Hasselmann program", as coined by Arnold (2001), of stochastic model reduction, which has received renewed attention in the past few years, has not been finished yet, and poses a fascinating challenge for mathematicians. In particular, how can we make the transition from a purely deterministic system to a stochastic system in a controllable way. In the following we will introduce a formalism which allows us to rewrite a deterministic system in a way that “looks like” a stochastic system in the form of generalized Langevin equations, and which may be a formal starting point for controlled stochastic model reduction.

2 Conceptual framework for stochastic modeling: The Mori-Zwanzig formalism

In a series of seminal papers Mori (1965b,a) and Zwanzig (1973) developed a formalism to rewrite a deterministic dynamical system in a form which resembles a general Langevin equation of the form

$$\frac{dz}{dt} = f(z(t)) + \int_0^t K(z(t-s),s)ds + \dot{W}_t \ . \tag{1}$$

The first term is Markovian, the second term describes possible memory of the process and W_t denotes a stochastic process. The Mori-Zwanzig formalism provides a conceptual framework to study dimension reduction and to parametrize uninteresting variables by a stochastic process.

We first briefly review the Mori-Zwanzig formalism before formulating standard deterministic and stochastic parameterization techniques such as averaging and homogenization within this framework.

2.1 The Mori-Zwanzig projection operator formalism

The main idea behind the reformulation of deterministic dynamics is simple and can be understood by the method of variation of constants. The following illustrative example is taken from the book by Zwanzig (2001). Consider the coupled linear system

$$\begin{aligned} \dot{x} &= L_{11}x + L_{12}y \\ \dot{y} &= L_{21}x + L_{22}y \ . \end{aligned}$$

Suppose we are only interested in the dynamics of x, and only have climatic knowledge of the initial condition of the variable y, i.e. its mean and variance. We can then solve for y to obtain

$$y(t) = e^{L_{22}t}y(0) + \int_0^t e^{L_{22}(t-s)}L_{21}x(s)\, ds \ ,$$

which we may use to express the dynamics of the “interesting” variable as

$$\dot{x} = L_{11}x + L_{12}\int_0^t e^{L_{22}(t-s)}L_{21}x(s)\, ds + L_{12}e^{L_{22}t}y(0) \ .$$

This is of the form of a generalized Langevin equation (1), where the first term is Markovian, the second term a memory term, and the third term is a noise term if we treat the initial conditions $y(0)$ as noise.

Let us now consider the general nonlinear case. Consider the deterministic dynamical system

$$\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}) \ , \tag{2}$$

with initial condition $\mathbf{z}(0) = \mathbf{z}_0$. Here $\boldsymbol{z}$ is either a finite-dimensional state vector $\mathbf{z} \in \mathbb{R}^d$ or an element of a Hilbert space. Associated with the typically nonlinear dynamical system (2) is the following linear partial differential equation for the temporal evolution of an observable $v(\boldsymbol{z}, t)$

$$\frac{\partial v}{\partial t} = \mathcal{L}v \quad \text{with} \quad v(\boldsymbol{z}, 0) = \Phi(\boldsymbol{z}) \ , \tag{3}$$

with the generator

$$\mathcal{L} = \mathbf{f}(\boldsymbol{z}) \cdot \nabla \ , \tag{4}$$

where ∇ denotes the gradient in phase space, i.e. $\mathbf{f}(\boldsymbol{z}) \cdot \nabla = f_i(\boldsymbol{z})\partial_{z_i}$. The solution of (3) is formally written as

$$v(\boldsymbol{z}, t) = e^{\mathcal{L}t}\Phi(\boldsymbol{z}) \ . \tag{5}$$

The equivalence of the nonlinear ordinary differential equation (2) and the linear partial differential equation (3) can be seen mathematically by employing the chain rule on $v(\mathbf{z}(t))$ or by the following heuristic consideration. To determine the value of an observable at time t one may either follow a trajectory and evaluate the observable along the trajectory or one may follow the evolution of the actual observable along the characteristic, i.e. $v(\boldsymbol{z}, t) = \Phi(\mathbf{z}(t))$ with $\mathbf{z}(0) = \boldsymbol{z}$. We assume here that the vector field $\mathbf{f}(\boldsymbol{z})$ is smooth enough to assure uniqueness and existence of solutions of (2) and of classical solutions of the associated partial differential equation (3). Note that $\mathcal{L}$ is the formal L^2-adjoint operator of the Liouville operator $\mathcal{L}^\star$ with $\mathcal{L}^\star\rho = -\nabla \cdot (f(\boldsymbol{z})\rho)$, controlling the evolution of densities of ensembles propagated according to (2).

Suppose one is not interested in resolving the full solution $\boldsymbol{z}(t)$ but rather is interested in only a few observables $\Phi(\boldsymbol{z}) = (\Phi_1(\boldsymbol{z}), \Phi_2(\boldsymbol{z}), \cdots, \Phi_n(\boldsymbol{z}))$. A particular relevant example occurs in the situation where the state space can be decomposed as $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ into "interesting variables" $\boldsymbol{x} = (z_1, \cdots, z_n) \in \mathbb{R}^n$ and the remainder of "uninteresting" variables $\boldsymbol{y} = (z_{n+1}, \cdots, z_d)$. The resolved observables would be $\Phi(\boldsymbol{z}) = (z_1, \cdots, z_n)$ in this case. In the infinite dimensional case where (2) denotes a partial differential equation, one may consider the dynamical system (2) as a Galerkin approximation and the resolved observables could, for example, be the low-order Fourier modes of a spectral representation of the solution, i.e. the relevant variables as those with small wavenumbers $k_< = \{k : k \leq k^\star\}$ and the irrelevant variables as those with high wavenumbers $k_> = \{k : k > k^\star\}$. The question we are concerned with in model reduction is how to distill the effective dynamics of these "interesting" observables?

To distill the dynamics of the interesting variables $\Phi(\boldsymbol{z})$ we require a projection operator $\mathcal{P}$ which maps functions of $\boldsymbol{z}$ to functions of $\Phi(\boldsymbol{z})$. In the context of partial differential equations, the projection operator can then be defined, for example, as $(\mathcal{P}\omega(\mathbf{k}))(\mathbf{k}_<) = \omega(\mathbf{k}_<, 0)$ for functions $\omega(\mathbf{k})$. Let us restrict for simplicity of exposition to the case where $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ and $\Phi(\boldsymbol{z}) = \boldsymbol{x} \in \mathbb{R}^n$. A suitable projector for the situation when the initial conditions of the interesting variables $\boldsymbol{x}$ are known exactly but only statistical information is available for the unresolved variables $\boldsymbol{y}$, is the conditional expectation of a function $\omega(\boldsymbol{x}, \boldsymbol{y})$

$$(\mathcal{P}\omega)(\boldsymbol{x}) = \frac{1}{\Omega(\boldsymbol{x})} \int_{\mathbb{R}^{d-n}} \omega(\boldsymbol{x}, \boldsymbol{y})\, \mu_{\boldsymbol{x}}(d\boldsymbol{y}) \ , \tag{6}$$

where $\mu_{\boldsymbol{x}}(d\boldsymbol{y})$ denotes the conditional measure of the unresolved variables. The normalization

$$\Omega(\boldsymbol{x}) = \int_{\mathbb{R}^{d-n}} \mu_{\boldsymbol{x}}(d\boldsymbol{y})$$

measures in the language of statistical mechanics the number of *microstates* which give rise to the *macrostate* $\boldsymbol{x}$.

In the context of equilibrium statistical mechanics often a unique invariant measure exists which supports a density with respect to the Lebesgue measure with $\mu_{\boldsymbol{x}}(d\boldsymbol{y}) = \rho_{\rm eq}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}$. In the context of deterministic dynamical systems typically a multitude of ergodic measures exist and the value of $(\mathcal{P}\omega)(\boldsymbol{x})$ would depend on the choice of the initial conditions of $\boldsymbol{y}$. To complicate things further, these measures may not depend continuously on $\boldsymbol{x}$ (see below). These measures are singular and their support is not on a set of full Lebesgue-measure but rather on an attractor or on a surface of constant energy in the case of dissipative and conservative deterministic dynamical systems, respectively. Nevertheless, for a large class of dynamical systems one can introduce the notion of a physical measure which supports densities on the surfaces of constant energy or on the attractor. In the case of (dissipative) chaotic deterministic dynamical systems these are given by so called Sinai-Ruelle-Bowen (SRB) measures (Young, 1998, 1999, 2002). SRB measures $\mu^{\rm SRB}$ satisfy the property that for a set of non-zero Lebesgue measure initial conditions $\boldsymbol{z}(0)$ and for every continuous observable φ we have

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \varphi(\boldsymbol{z}(t))dt \to \int \varphi\,\mu^{\rm SRB}\;.$$

This property assures that meaningful averages can be calculated and the statistics of the dynamical system can be explored by the asymptotic distribution of orbits starting from Lebesgue almost every initial condition. The class of systems for which SRB are proven to exist includes for example, uniformly hyperbolic systems, logistic-map type systems, Hénon-like attractors, Lorenz attractors and many more. It has recently been conjectured by Gottwald and Melbourne (2014) that typical dynamical systems are either regular or belong to the above class which enjoys good statistical properties. In the following all measures are understood to be SRB measures. Furthermore we assume that all measures are normalized to $\int \mu = 1$.

Given a projection operator $\mathcal{P}$, we denote by $\mathcal{Q} = \mathbf{1} - \mathcal{P}$ the orthogonal projector. We then write the problem (3) as

$$\begin{aligned}\frac{\partial v}{\partial t}(\boldsymbol{z},t) &= \mathcal{L}e^{\mathcal{L}t}\Phi(\boldsymbol{z})\\ &= e^{\mathcal{L}t}\mathcal{P}\mathcal{L}\Phi(\boldsymbol{z}) + e^{\mathcal{L}t}\mathcal{Q}\mathcal{L}\Phi(\boldsymbol{z})\;,\end{aligned}$$

which upon using the Duhamel-Dyson formula (see, for example, Evans and Morriss (2008)) for operators $\mathcal{A}$ and $\mathcal{B}$

$$e^{t(\mathcal{A}+\mathcal{B})} = e^{t\mathcal{A}} + \int_0^t e^{(t-s)(\mathcal{A}+\mathcal{B})}\,\mathcal{B}\,e^{s\mathcal{A}}\,ds\;, \tag{7}$$

becomes the celebrated Mori-Zwanzig equation (Mori, 1965b; Zwanzig, 1973)

$$\frac{\partial v}{\partial t}(\boldsymbol{z},t) = e^{\mathcal{L}t}\mathcal{P}\mathcal{L}\Phi(\boldsymbol{z}) + \int_0^t e^{(t-s)\mathcal{L}}\,\mathcal{P}\mathcal{L}\,e^{s\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\Phi(\boldsymbol{z})\,ds + e^{t\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}\Phi(\boldsymbol{z})\;, \tag{8}$$

with $\mathcal{A} = \mathcal{Q}\mathcal{L}$, $\mathcal{B} = \mathcal{P}\mathcal{L}$ and $\mathcal{A}+\mathcal{B} = \mathcal{L}$.

The Mori-Zwanzig equation (8) is not an approximation but is exact and constitutes an equivalent formulation of the full dynamical system (2). The interested reader is referred to Zwanzig (2001); Chorin and Hald (2006); Evans and Morriss (2008); Chorin *et al.*

(2000); Givon *et al.* (2004) for more details.

The Mori-Zwanzig equation (8) is in the form of a generalized Langevin equation: the first term on the right-hand side is Markovian while the second term involves memory. The last term $n(\boldsymbol{z}, t) = e^{t\mathcal{QL}}\mathcal{QL}\Phi(\boldsymbol{z})$ is labeled the noise term. This is because its temporal evolution

$$\frac{\partial n}{\partial t}(\boldsymbol{z}, t) = \mathcal{QL}n(\boldsymbol{z}, t) \quad \text{with} \quad n(\boldsymbol{z}, 0) = \mathcal{QL}\phi(\boldsymbol{z}) \; , \tag{9}$$

assures that the dynamics remains orthogonal to the range of the $\mathcal{P}$. In the case $\Phi(\boldsymbol{z}) = \boldsymbol{x}$ where we split the dynamical system (2) into the resolved and unresolved variables $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively, as

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{y}) \tag{10}$$

$$\dot{\boldsymbol{y}} = g(\boldsymbol{x}, \boldsymbol{y}) \; , \tag{11}$$

the orthogonal dynamics describes the dynamics of the fluctuating part of the vector field of the resolved variables since $n(\boldsymbol{z}, 0) = f(\boldsymbol{x}, \boldsymbol{y}) - (\mathcal{P}f)(\boldsymbol{x})$.

It is a formidable challenge to find effective approximations which render the Mori-Zwanzig equation as a closed equation for the resolved variables and finding approximations for the noise term and the infinite-dimensional memory kernel. Obviously this programme can only be exercised within approximations. In the following section we will describe a simple formal procedure to unravel the memory kernel.

2.2 Writing the memory term as an infinite chain of Markov terms

Several approximations have been employed to simplify the memory term. Mori (1965b) considered the case where the dynamics is given by a Hamiltonian system and the projection operator is the average over all variables. In this case the Mori-Zwanzig equation is a linear equation for the resolved variables and the Laplace transform of the memory kernel could be written as a continued fraction rendering the Mori-Zwanzig equation as a Markov chain. An extension to the non-Hermitian case was given by Grigolini (1982). In the nonlinear case the short-memory approximation was introduced (Chorin *et al.*, 2000) which allows for an analytical treatment. Loosely speaking this assumption states that the resolved and the unresolved subspaces do not couple, and one may use the full dynamics in order to propagate the elements of the orthogonal subspace. In the short-memory approximation the memory term is replaced by a damping term which is linear in the time variable t which renders these approximation unsatisfactory for long time integrations and for estimating the statistics of the slow variables.

We will now present a simple reformulation of the Mori-Zwanzig equation which allows for a computationally accessible criterion for a truncation of the memory term. We rewrite problem (3) as

$$\frac{\partial v}{\partial t}(\boldsymbol{z}, t) = e^{\mathcal{L}t}\mathcal{PL}\Phi(\boldsymbol{z}) + e^{\mathcal{L}t}\mathcal{QL}\Phi(\boldsymbol{z}) \; .$$

The second term, as we have seen in the previous section, can be written as the sum of a memory kernel and a noise-like term, here however, we express the second term as a time dependent forcing of the first Markovian term, and specify its time evolution

$$\begin{aligned}\frac{\partial v}{\partial t}(\boldsymbol{z},t) &= e^{\mathcal{L}t}\mathcal{P}\mathcal{L}\Phi(\boldsymbol{z}) + n_1(t)\\ \frac{\partial n_1}{\partial t}(\boldsymbol{z},t) &= \mathcal{L}n_1(t)\ ,\end{aligned}$$

with

$$n_1(t) = e^{\mathcal{L}t}\mathcal{Q}\mathcal{L}\Phi(\boldsymbol{z})\ .$$

Note that n_1 and v solve the same linear partial differential equation. Repeating this process for the equation for n_1 we arrive at the infinite Markov chain

$$\begin{aligned}\frac{\partial v}{\partial t}(\boldsymbol{z},t) &= \boldsymbol{\Lambda}_0\Phi(\boldsymbol{z}) + n_1(\boldsymbol{z},t)\\ \frac{\partial n_1}{\partial t}(\boldsymbol{z},t) &= \boldsymbol{\Lambda}_1 n_1(\boldsymbol{z}) + n_2(\boldsymbol{z},t)\\ &\vdots\\ \frac{\partial n_k}{\partial t}(\boldsymbol{z},t) &= \boldsymbol{\Lambda}_k n_k(\boldsymbol{z}) + n_{k+1}(\boldsymbol{z},t)\\ &\vdots\end{aligned}$$

with the operator of the Markov term

$$\boldsymbol{\Lambda}_k = e^{\mathcal{L}t}\mathcal{P}_k\mathcal{L} \tag{12}$$

and the forcing

$$n_k(\boldsymbol{z},t) = e^{\mathcal{L}t}\mathcal{Q}_k\mathcal{L}n_{k-1}(\boldsymbol{z})\ ,$$

with $n_0(\boldsymbol{z}) = \Phi(\boldsymbol{z})$. Note that we allow for different projectors $\mathcal{P}_k$ and $\mathcal{Q}_k = \mathbf{1} - \mathcal{P}_k$ at different levels. The advantage of this formulation is that we can now introduce a condition to truncate this infinite series, thereby approximating the infinite-dimensional memory term. The Markov chain can be truncated at level k provided the autocorrelation function of the remainder $\langle n_{k+1}(t)n_{k+1}(s)\rangle$ corresponds to some known noise process, for example, if

$$\langle n_{k+1}(t)n_{k+1}(s)\rangle = \sigma^2\delta(t-s)\ . \tag{13}$$

Using the Mori-Zwanzig formalism to unravel possible memory and long-time persistence of the dynamics by enlarging the state-space until closure can be obtained has been algorithmically realized in the multi-level regression ideas promoted in Kravtsov *et al.* (2005) (see also Chekroun *et al.* (2011)). It has been employed, for example, to model the El-Niño-Southern Oscillation system (Kondrashov *et al.*, 2005) and low-frequency variability in a three-level quasi-geostrophic model (Kondrashov *et al.*, 2006). The particular implementation of restricting the vector fields Λ_k in (12) to quadratic polynomials in those works was pointed out to lead to undesirable instabilities in energy-conserving systems by Majda and Yuan (2012). Attempts to remedy those short-comings whilst keeping the general idea of Kravtsov *et al.* (2005), and thereby of the framework advocated here, were proposed by Majda and Harlim (2013) and Kondrashov *et al.* (2015).

In the following we discuss two generic situations for which truncations of the Markov chain can be rigorously justified and for which the memory kernel vanishes all together. At the end of this section we briefly describe some promising new directions in going beyond the assumption of infinite-timescale separation underlying the rigorous theory.

2.3 Averaging in the Mori-Zwanzig framework

We will consider systems for slow variables $x \in \mathbb{R}^n$ and fast variables $y \in \mathbb{R}^m$ of the form

$$\dot{x} = f_0(x, y) \tag{14}$$

$$\dot{y} = \frac{1}{\varepsilon} g_0(x, y) \ , \tag{15}$$

where the fast y-dynamics is assumed to be ergodic with unique invariant measure $\mu_x(dy)$ conditioned on the slow variables x. The associated generator is

$$\mathcal{L} = \frac{1}{\varepsilon}\mathcal{L}_0 + \mathcal{L}_1 \ ,$$

with

$$\mathcal{L}_0 = g_0(x, y)\partial_y \quad \text{and} \quad \mathcal{L}_1 = f_0(x, y)\partial_x \ .$$

We consider the case $\Phi(x, y) = x$ and the projection operator

$$(\mathcal{P}\omega)(x) = \int \omega(x, y)\, \mu_x(dy) \ . \tag{16}$$

We obtain at the first level

$$\begin{aligned} \dot{x} &= e^{\mathcal{L}t}\mathcal{P}\mathcal{L}x + n_1 \\ &= \langle f_0 \rangle + n_1 \end{aligned}$$

with

$$n_1 = e^{\mathcal{L}t}\mathcal{Q}\mathcal{L}x \ ,$$

where we introduced $\langle \omega \rangle = (\mathcal{P}\omega)(x)$ for ease of notation. In the limit of infinite time scale separation $\varepsilon \to 0$ we obtain $n_1 = 0$ and hence the Mori-Zwanzig equation reduces to the effective deterministic averaged equation (17). This is seen by

$$\begin{aligned} n_1 &= e^{\mathcal{L}t}\left(f_0(x, y) - \langle f_0 \rangle\right) \\ &= e^{\mathcal{L}_0 \frac{t}{\varepsilon}}\left(f_0(x, y) - \langle f_0 \rangle\right) + \int_0^t e^{(t-s)\mathcal{L}}\, \mathcal{L}_1\, e^{\mathcal{L}_0 \frac{s}{\varepsilon}}\left(f_0(x, y) - \langle f_0 \rangle\right) ds \ , \end{aligned}$$

where we used the Duhamel-Dyson formula (7). Using the *large deviation principle* whereby deviations of time averages from the corresponding spatial averages are rare (Melbourne and Nicol, 2008), we argue

$$\lim_{\varepsilon \to 0} e^{\mathcal{L}_0 \frac{t}{\varepsilon}} f_0(x, y) = \langle f_0 \rangle \ ,$$

and hence obtain

$$\lim_{\varepsilon \to 0} n_1 = 0 \ .$$

The asymptotic slow dynamics is then summarized to be

$$dX = F(X)dt \ , \tag{17}$$

with

$$F(x) = \int f_0(x,y)\,\mu_x(dy)\ . \tag{18}$$

The slow dynamics (17) are the well-known deterministic *averaged* equations. It is well known that on bounded time scales $\mathcal{O}(1)$ the slow dynamics of the multi-scale system (14)-(15) is approximated by (17) (see for example Arnold *et al.* (1993); Sanders and Verhulst (1985); Givon *et al.* (2004); Pavliotis and Stuart (2008)).

The above exposition is entirely formal. For deterministic dynamical systems rigorous theory is established in the case when the chaotic fast dynamics is hyperbolic and the fast dynamics does not depend on the slow dynamics with $g_0 = g_0(y)$ by Anosov (1960); Kifer (1992, 1995, 2001, 2003, 2005). Therein also stochastic fluctuations around the mean behavior were treated on longer diffusive time scales (see the next Section2.4 on homogenization). An open problem is how to treat the general case where the slow variables couple back to the fast chaotic system, i.e. $g_0 = g_0(x,y)$, and the fast dynamics is not hyperbolic. Difficulties occur if the measure $\mu_x(dy)$ does not depend smoothly on the slow variable x – this is, for example, the case when the fast dynamics experiences bifurcations upon varying x. In this case the averaged vector field may not even be a continuous function of the slow variable and unique solutions of the averaged equations (17) are not guaranteed. The same type of problem appears when trying to apply linear response theory in climate science (see Baladi and Smania (2008)).

2.4 Homogenization in the Mori-Zwanzig framework

The slow averaged equations (17) are only valid on bounded time scales of order $\mathcal{O}(1)$ and solutions of the averaged equation (17) will not be close to solutions of the slow variable of the full multi-scale system (14)-(15) on long time scales. We consider here the case when the averaged drift is small with $\langle f_0\rangle = \mathcal{O}(\varepsilon)$. In this case fluctuations become important. To illustrate how stochasticity and diffusive behavior arises asymptotically in multi-scale systems on long time scales, we begin with a simplified version of the general dynamical system (14)-(15) in which the fast chaotic dynamics drives the slow dynamics non-multiplicatively and the slow dynamics does not couple back into the fast dynamics. We will now consider systems of the form

$$\begin{aligned}\dot{x} &= \frac{1}{\varepsilon}f_0(y) + f_1(x,y)\\ \dot{y} &= \frac{1}{\varepsilon^2}g_0(y)\ ,\end{aligned} \tag{19}$$

where the fast y-dynamics is again assumed to be ergodic with unique invariant measure $\mu(dy)$. The associated generator is

$$\mathcal{L} = \frac{1}{\varepsilon^2}\mathcal{L}_0 + \frac{1}{\varepsilon}\mathcal{L}_1 + \mathcal{L}_2\ ,$$

with

$$\mathcal{L}_0 = g_0(y)\partial_y\ , \quad \mathcal{L}_1 = f_0(y)\partial_x \quad \text{and} \quad \mathcal{L}_2 = f_1(x,y)\partial_x\ .$$

Upon neglecting $(\mathcal{P}f_0)(x) = \langle f_0\rangle = \mathcal{O}(\varepsilon)$ we obtain at the first level of the Mori-Zwanzig formalism

$$\begin{aligned}\dot{x} &= e^{\mathcal{L}t}\mathcal{P}\mathcal{L}x + n_1\\ &= \langle f_1\rangle + n_1\end{aligned} \tag{20}$$

with

$$\begin{aligned} n_1 &= e^{\mathcal{L}t}\mathcal{Q}\mathcal{L}x \\ &= e^{\mathcal{L}t}(\mathcal{I}-\mathcal{P})\mathcal{L}x \\ &= e^{\mathcal{L}t}\left(f_1(x,y)-\langle f_1\rangle(x)\right)+\frac{1}{\varepsilon}e^{\mathcal{L}_0\frac{t}{\varepsilon^2}}f_0(y)\ . \end{aligned}$$

The first term of n_1 vanishes in the limit $\varepsilon \to 0$ as part of the large deviation principle mentioned in the previous section. The second term gives rise to noise in (20) as can be motivated as follows. Integrating the second term we obtain

$$\frac{1}{\varepsilon}\int_0^t e^{\mathcal{L}_0\frac{t}{\varepsilon^2}}f_0(y)\,dt = \varepsilon\int_0^{\frac{t}{\varepsilon^2}} f_0(y(s))\,ds\ .$$

For sufficiently chaotic fast dynamics one may evoke the central limit theorem for $\varepsilon \to 0$ to justify

$$\frac{1}{\varepsilon}e^{\mathcal{L}_0\frac{t}{\varepsilon^2}}f_0(y) = \dot{W}_t\,,$$

where W_t is an n-dimensional Brownian motion with covariance matrix Σ given by the Green-Kubo type relation

$$\Sigma\Sigma^T = \int_0^\infty \mathcal{P}\left(f_0(y)\,e^{\frac{t}{\varepsilon^2}\mathcal{L}_0}f_0(y)\right)\,dt\ .$$

Hence, summarizing, on long time scales $\mathcal{O}(1)$ the slow dynamics (20) is given by the *homogenized* equation

$$dX = F(X)dt + \Sigma\,dW_t\ , \tag{21}$$

where the drift coefficient is given by

$$F(x) = \int f_1(x,y)\,\mu(dy)\ .$$

In the more general case

$$\dot{x} = \frac{1}{\varepsilon}f_0(x,y) + f_1(x) \tag{22}$$

$$\dot{y} = \frac{1}{\varepsilon^2}g_0(y)\ , \tag{23}$$

we expect

$$\frac{1}{\varepsilon}\int_0^{dt} e^{\mathcal{L}t}f_0(x,y)\,dt$$

to converge to Brownian motion W_t with variance $\Sigma(x)$. Now the question arises how to interpret the stochastic integral of $\Sigma(x)dW_t$. It is well known that stochastic integrals $\int\Sigma(x)dW_t$ are very sensitive with respect to the approximation of the Brownian motion with the Itô and the Stratonovich interpretations being two cases. A reader-friendly discussion on the Itô versus Stratonovich issue is contained in the book by Horsthemke (1984). The Wong-Zakai theorem and its extensions (Wong and Zakai, 1965; Ikeda and Watanabe, 1981) provide general conditions under which convergence holds with the Stratonovich interpretation for the stochastic integral. The rationale behind the Stratonovich interpretation of the noise in homogenized equations is that here rough noise arises as a limit involving only smooth functions of the smooth deterministic system. Hence in the limit classical calculus should prevail implying the Stratonovich interpretation. We remark that the conditions for the Wong-Zakai theorem are satisfied in the case of one-dimensional slow variables, but may fail in higher dimensions. The multi-dimensional homogenized equations associated with (22)-(23) are given by

$$dX = F(X)dt + \Sigma(X)dW_t \; , \tag{24}$$

where W_t denotes m-dimensional Brownian motion and the drift coefficient is given by

$$F(x) = \int f_1(x,y)\, \mu(dy) + \int_0^\infty ds \int f_0(x,y) \cdot \nabla f_0(x, y(s))\, \mu(dy) \; ,$$

and the diffusion coefficient is defined by

$$\Sigma(X)\Sigma^T(X) = \int_0^\infty ds \int \left(f_0(y) \otimes f_0(y(s)) + f_0(y(s)) \otimes f_0(y) \right) \mu(dy) \; ,$$

where the outer product between two vectors is defined as $(a \otimes b)_{ij} = a_i b_j$ (see Papanicolaou and Kohler (1974); Ikeda and Watanabe (1981); Kelly and Melbourne (2014)). It is pertinent to stress that mixing of the fast chaotic flow is not necessary for the stochastic limit systems (21) and (24) to exist.

Homogenisation is well understood in the context of multi-scale systems where the fast dynamics is stochastic with a unique invariant density (Khasminsky, 1966; Kurtz, 1973; Papanicolaou, 1976), see also Givon *et al.* (2004); Pavliotis and Stuart (2008). Rigorous results for diffusive limits of deterministic dynamical systems have only recently been obtained (Melbourne and Stuart, 2011; Gottwald and Melbourne, 2013b; Kelly and Melbourne, 2014). It is pertinent to mention that these rigorous results do not make any assumptions on the mixing properties of the fast chaotic dynamics as assumed in most heuristic homogenization approaches such as, for example, in Majda *et al.* (2006). These results are, however, at this stage restricted to the case where the slow dynamics does not couple back to the fast dynamics, i.e. $g = g(y)$. The general case $g = g(x, y)$ is still an interesting open question for the same reasons as discussed above for the case of averaging.

When simulating multi-scale systems such as the climate, one uses discretizations of the continuous-time dynamical systems. Homogenization results for the resulting multi-scale maps yield very different results compared to their associated continuous-time parents. It was shown in Gottwald and Melbourne (2013b) that for a one-dimensional slow deterministic dynamics the homogenized system is neither of the Itô nor of the Stratonovich type which yields widely different statistics than the limiting continuos multi-scale system which converges to a stochastic differential equation with Stratonovich noise.

The idea of homogenization was spearheaded in the climate community by the celebrated *MTV* approach. The acronym *MTV* stands for the surnames of the authors of the original paper Majda *et al.* (1999). The main message learned from homogenisation for developing reduced stochastic models is the inclusion of correlated additive and multiplicative noise (CAM) (Majda *et al.*, 2009) rather than simple additive noise. Homogenization relates this type of noise to the dependency of the term $f_0(x, y)/\varepsilon$ on both, y and x. To achieve stochastic consistency between the reduced stochastic system and the multi-scale parent system, the parameters of the stochastic process are estimated from a priori knowledge of the climatic behavior of the slow variables such as matching the autocorrelation function (Majda *et al.*, 2002). The MTV strategy has been successfully applied to an atmospheric barotropic model on the sphere (Franzke *et al.*, 2005) and a 3-layer quasi-geostrophic model (Franzke and Majda, 2006). Both models have realistic atmospheric circulation features and the MTV approach is able to derive reduced order models which reasonably capture these features with as little as 10 resolved modes.

In general, the stochastic reduction techniques as described above do not respect a possible underlying conservation law of the full multi-scale system. In particular, if the multi-scale dynamics were Hamiltonian, energy conservation would not be guaranteed. Dubinkina and Frank (2007, 2010) have illustrated how the overall statistical properties depend on the conservation properties of a numerical discretization. The stochastic reduced normal forms by Majda *et al.* (2009) which are inspired by homogenization theory ensure that the nonlinear drift terms respect energy conservation, but the CAM noise does not impose any constraint on energy conservation. In energy conserving systems the noise would need to be projected onto the surface of constant energy. Frank and Gottwald (2013) have carried out such a homogenization theory for a simplified Lagrangian particle description of the shallow-water equations. Energy-conserving fast systems have also been considered in Jain *et al.* (2015) where the slowly-varying energy is treated as an additional slow hidden variable. It is at this stage though still unclear whether conserving certain dynamical quantities (and not others) may lead to dynamically and statistically inconsistent states.

Although the rigorous theory requires an infinite time-scale separation, i.e. $\epsilon \to 0$, it has been observed in numerical simulations that homogenized reduced equations remain reliable reduced models for the slow dynamics even for moderate time-scale separation. This is a familiar situation in asymptotic methods here the range of validity often extends the notion of what is small. It is, nevertheless, important to devise methods which go beyond the assumption of infinite timescale separation. For systems that can be seen as weakly coupled dynamical systems a recent interesting direction was proposed by Wouters and co-workers (Wouters and Lucarini, 2012, 2013). Therein the Mori-Zwanzig formalism was combined with linear response theory to provide a closure of the relevant slow dynamics which does not rely on any time-scale separation and which retains some information form the memory kernel.

2.5 What type of noise?

A question which so far has not attracted much attention is what type of noise one can expect as a limit in multi-scale systems? It is tacitly assumed in the current body of work on reduced stochastic models that fast degrees of freedom are parameterized by Brownian motion. The current justification is the central limit theorem. The central limit theorem indeed holds for a large class of deterministic dynamical systems (see Melbourne and Nicol (2005, 2009) for mathematical details). In particular, the central limit theorem holds for *strongly chaotic* systems[1]. *Weakly chaotic* dynamics for which the central limit theorem does not hold are characterized by a large degree of intermittency whereby periods of chaotic dynamics are intermittently disturbed by long laminar periods of seemingly regular behavior. The central limit theorem, however, can be modified for weakly chaotic dynamics

[1] It is pertinent to mention that strong chaoticity is not related to an exponential decay of correlations. See Gottwald and Melbourne (2013a, 2014) for details and definitions of *strong* and *weak* chaoticity.

(Gouëzel, 2004). This has been used by Gottwald and Melbourne (2013b) to show that the limiting noise on the homogenized equations is an α-stable noise (or often called Lévy process). These non-Gaussian processes are characterized by the occurrence of jumps of all sizes and have a probability density function with algebraically decaying so called fat tails. The power-law decay of the distribution tails implies a non-vanishing probability of large jumps and causes an infinite variance (see for example Chechkin *et al.* (2008) for an introduction). This type of noise has been observed in planetary-scale atmospheric circulation (Viecelli, 1998) as well as in abrupt millennial scale climate changes during the last ice age in ice-core data (Ditlevsen, 1999).

In one dimension the stochastic integrals arising in the reduced homogenized dynamics are then to be interpreted in the sense of Marcus integrals (see Applebaum (2009); Chechkin and Pavlyukevich (2014)) which is the interpretation allowing for the validity of classical calculus akin to the Stratonovich integral in the case of Brownian motion. It is well known that the simple occurrence of fat tails may not necessarily imply an α-stable distribution but may as well be associated with multiplicative Brownian noise. Penland and Ewald (2008) suggested therefore to favor Brownian CAM noise over Lévy noise for practical purposes. However, these two processes are dynamically entirely different and, moreover, can be distinguished with relatively easy diagnostic tools such as the p-variation (Hein *et al.*, 2009; Burnecki and Weron, 2010; Burnecki *et al.*, 2012). We believe it will be an interesting avenue to study how intermittent dynamics, caused by for example persistent atmospheric pattern such as blocking, can lead to fat tails in the probability density function of slow processes such as ocean temperatures using homogenization techniques.

3 Data-driven models

The reduction techniques described in the previous section provide a systematic approach to analytically derive stochastic models for the dynamics of slow degrees of freedom in the climate system. However, there are situations where these techniques are not feasible, e.g. because of the complexity of the underlying model equations, or because of the absence of a clear scale separation. In such cases, a useful alternative approach can be to infer stochastic models from data. These data, usually in the form of time series, can come from observations of the real climate system, but it can also be useful to infer reduced stochastic models from data obtained from simulations with comprehensive numerical models.

The central task in this data-driven approach is one of statistical inference: one must fit stochastic processes from a suitable class to the observations at hand. The most commonly used class is formed by diffusion processes, i.e. models consisting of stochastic differential equations (SDEs) driven by standard Brownian motion. Other classes considered in this context include Lévy processes, discrete processes (finite-state Markov chains) and Hidden Markov Models (HMMs).

Let us consider a general d-dimensional diffusion process $X(t) \in \Omega \subseteq \mathbb{R}^d$ with corresponding SDE

$$dX(t) = b(X(t))\, dt + \sigma(X(t))\, dW(t)\,, \tag{25}$$

in which $W(t)$ is a d-dimensional vector of independent Wiener processes. We define the diffusion coefficient as

$$a(x) := \sigma(x)\sigma(x)^T\,. \tag{26}$$

Note that for $d > 1$, b is vector-valued and a is matrix-valued. Furthermore, we focus on situations where the drift $b(X(t))$ and diffusion $a(X(t))$ do not depend explicitly on time, but only implicitly through their dependence on $X(t)$.

Inferring the functions $b(X(t))$ and $a(X(t))$ from time series (observations) of $X(t)$ can be very challenging. A key difficulty is that the finite-time transition density of the process (25) is in general unknown, i.e. there is no closed-form expression, in terms of b and a, for the density at time $t + \Delta t$ given the density at time t. Since observations are usually discrete in time, this is a major problem for inference procedures that rely on the likelihood function.

In the simplest situation, b is a linear function of $X(t)$, σ is a constant, and the process is univariate ($d = 1$), so that (25) is a scalar Ornstein-Uhlenbeck process (in fact, the OU process is one of the few diffusion processes for which the transition density is known). Difficulties arise if b is nonlinear, σ is $X(t)$-dependent (multiplicative noise) or $d > 1$ (or a combination of these). A more detailed discussion of these difficulties can be found in Gobet *et al.* (2004); Sørensen (2004) and references therein.

3.1 Linear Inverse Modeling

The technique of so-called Linear Inverse Modeling (LIM) is used frequently in climate science to fit stochastic models with linear drift and additive noise to data. It has, amongst others, been used for modeling and predicting sea surface temperatures in the equatorial Pacific ocean (e.g. Penland and Magorian (1993); Penland and Sardeshmukh (1995)) and atmospheric low-frequency variability (e.g. Winkler *et al.* (2001)). Assuming zero mean, the SDE of such a linear model with additive noise is

$$dX(t) = B\,X(t)\,dt + L\,dW(t)\,, \tag{27}$$

where B and L are both $(d \times d)$ constant matrices. This diffusion process has a Gaussian invariant probability distribution (provided B is negative definite), and is not a suitable model for phenomena that are manifestly non-Gaussian. This restriction to the simple class of Gaussian processes with constant diffusion, however, has the major advantage that the inference of the matrices B and L from observational data can be done in a computationally efficient way even for high-dimensional systems with large d.

We define $C(\tau)$ as the lag-τ covariance matrix of $X(t)$, i.e. its matrix elements are the expectations

$$C_{ij}(\tau) = \mathbb{E}[X_i(t+\tau)X_j(t)]\,. \tag{28}$$

From (27) it follows that $C(\tau)$, with any $\tau > 0$, and $C(0)$ are related according to

$$C(\tau) = \exp(B\,\tau)\,C(0)\,. \tag{29}$$

If we have time series data available with sampling interval Δt, i.e. a set of observations $\{X^{\mathrm{obs}}(0), X^{\mathrm{obs}}(\Delta t), X^{\mathrm{obs}}(2\Delta t), ..., X^{\mathrm{obs}}(N\Delta t)\}$, we can estimate the elements of $C(0)$ and $C(\Delta t)$, assuming ergodicity of the underlying dynamical process, as

$$\hat{C}_{ij}(0) \;=\; \frac{1}{N}\sum_{n=0}^{N} X_i^{\mathrm{obs}}(n\Delta t)X_j^{\mathrm{obs}}(n\Delta t) \tag{30}$$

$$\hat{C}_{ij}(\Delta t) \;=\; \frac{1}{N}\sum_{n=0}^{N} X_i^{\mathrm{obs}}(n\Delta t)X_j^{\mathrm{obs}}((n-1)\Delta t)\;. \tag{31}$$

The estimate of B can then be obtained using (29) as

$$\hat{B} = (\Delta t)^{-1}\,\log[\hat{C}(\Delta t)\,(\hat{C}(0))^{-1}]\,. \tag{32}$$

We note that computing the logarithm of a matrix, as is done in (32), is not entirely trivial due to the non-uniqueness of the logarithm (see Higham (2008) for more details).

An estimator for the matrix L can be obtained from the equations for the second moments (covariances) of the process (27). Let $\rho(x,t)$ be the probability density function at time t associated with (27). The second moments are

$$\langle x_p\, x_q \rangle_\rho := \int_\Omega dx\, \rho(x,t) x_p\, x_q \,. \tag{33}$$

Clearly, the time evolution of the moments is determined by the time evolution of ρ, which is in turn governed by the Fokker-Planck equation. For the process (27) it reads

$$\frac{\partial}{\partial t}\, \rho(x,t) = -\sum_i \frac{\partial}{\partial x_i}\, (Bx)_i \rho(x,t) + \frac{1}{2} \sum_{i,j} A_{ij} \frac{\partial^2}{\partial x_i\, \partial x_j}\, \rho(x,t) \,, \tag{34}$$

where $A = LL^T$. If we assume that $\rho(x,t)$ and its first spatial derivatives are zero at the boundary of Ω for all t, it is straightforward to derive that

$$\partial_t \, \langle x_p\, x_q \rangle_\rho = \sum_j B_{pj} \langle x_j\, x_q \rangle_\rho + \sum_j B_{qj} \langle x_p\, x_j \rangle_\rho + A_{pq} \tag{35}$$

As $t \to \infty$, ρ tends to the invariant density, so that $\partial_t \, \langle x_p\, x_q \rangle_\rho \to 0$ and $\langle x_p\, x_q \rangle_\rho \to C_{pq}(0)$, cf. (28). Thus, we have

$$B\, C(0) + C(0)\, B^T + A = 0 \,. \tag{36}$$

Together with the estimates $\hat{B}$ and $\hat{C}(0)$ obtained before, this relation can be used to determine an estimate $\hat{A}$ of the matrix A. For the final step, arriving at an estimate of the matrix L that appears in (27), one can use a Cholesky decomposition of the estimate $\hat{A}$. Because $\hat{A}$ is symmetric positive-definite, the Cholesky decomposition is unique (Golub and Van Loan, 2013). However, other decompositions are possible and $\hat{L}$ is not uniquely defined by $\hat{A}$ since, if $\hat{A} = \hat{L}\, \hat{L}^T$ then also $\hat{A} = \tilde{L}\, \tilde{L}^T$ with $\tilde{L} = \hat{L}\, Q$ and Q any orthogonal matrix, i.e. $Q\, Q^T = 1$. This non-uniqueness reflects the fact that the same diffusive behavior as described by the Fokker-Planck equation (34), can be generated by different stochastic differential equations (27) if they have different L (but the same $L\, L^T$ and B).

To summarize, for the LIM procedure one needs to compute estimates for the two covariance matrices $C(0)$ and $C(\Delta t)$ from the observations $X(0)$, $X(\Delta t)$, ..., $X(N\Delta t)$. The estimates for B and A are then obtained using (32) and (36). The computations are quite straightforward and can easily be performed for processes in high-dimensional spaces, e.g. $d = O(10^2)$. A drawback is that it applies, as discussed above, to a rather restrictive class of Gaussian processes with constant diffusion described by (27). For further details the interested reader is referred to Penland and Magorian (1993); Penland and Sardeshmukh (1995); Winkler *et al.* (2001).

3.2 Inference for general diffusion processes

Statistical inference for diffusion processes with nonlinear drift and/or multiplicative noise is much more difficult than for (27), in particular in the case of multivariate processes. Several approaches have been developed and used in the context of atmosphere-ocean science. It must be mentioned that statistical inference for diffusions is a relevant tool for a wide range of applications in physics, chemistry, biology, econometrics and finance, going well beyond the context of climate science which is the focus of this chapter. Not surprisingly, there exists a large body of literature on this topic, both on the theory in mathematical statistics as well as on applications to specific problems. We do not attempt to give a broad overview here (see e.g. Rao (1999); Sørensen (2004); Kutoyants (2004); Bishwal (2008) for such overviews), rather we focus on a few methodologies that are used in atmosphere-ocean science.

In one approach, the drift and diffusion functions are inferred using their statistical definitions as conditional first and second moments of the process increments. Starting again from the general diffusion process (25), the drift $b(x)$ and diffusion $a(x)$ as defined in (26) are related to the increments $X(t+\Delta t) - X(t)$ as follows:

$$\begin{aligned} b(x) &= \lim_{\Delta t \downarrow 0} (\Delta t)^{-1} \mathbb{E}[X(t+\Delta t) - X(t) \,|\, X(t) = x] \,, && (37) \\ a(x) &= \lim_{\Delta t \downarrow 0} (\Delta t)^{-1} \mathbb{E}[(X(t+\Delta t) - X(t)) \otimes (X(t+\Delta t) - X(t)) \,|\, X(t) = x] \,. && (38) \end{aligned}$$

By binning the state space, i.e. subdividing Ω into non-overlapping sets Ω_k (bins), one can use these definitions to compute estimates for b and a in each bin from the observations $X^{\text{obs}}(t)$. This approach was proposed in the physics community in Siegert *et al.* (1998); Friedrich *et al.* (2000), and was used to analyze atmospheric datasets in e.g. Sura (2003); Berner (2005). This approach is very general and does not assume a specific functional form for the drift or diffusion, so it can be applied to processes involving nonlinear drift and non-constant diffusion. However, its practical use is limited to low-dimensional processes because the number of bins grows exponentially in d. Therefore the amount of data needed to obtain statistically meaningful estimates in each bins also grows exponentially with d. Another difficulty is that the estimators based on (38) rely on Δt being small. This becomes a problem if the observations are not generated by a d-dimensional diffusion process but rather if the underlying dynamics of the observed system is of a deterministic chaotic nature or if the observations are given as a projection of a higher-dimensional process. For example, the limit of the diffusion $a(x)$ becomes zero for $\Delta t \to 0$ in deterministic systems. As for a projected process, this will generally be non-Markov, so the result of fitting a Markov process to it will depend on the choice of Δt. The issue of the choice of the sampling time and possible biases of the estimation of drift $b(x)$ and diffusion $a(x)$ for too small or too large observation intervals will be discussed in more detail in Section3.3.

The methodology proposed in Crommelin and Vanden-Eijnden (2006, 2011) (see also Gobet *et al.* (2004)) is partly motivated by the need to overcome the small Δt limit without introducing time discretization errors as in the estimators based on (38). At the core of this method lies the relationship between the conditional expectation operator denoted $P_{\Delta t}$ and the diffusion operator (or backward Fokker-Planck operator) denoted $\mathcal{L}$. For suitable functions $h(x)$, we define the former as

$$(P_{\Delta t}\, h)(x) = \mathbb{E}[h(X(\Delta t)) \,|\, X(0) = x] \,. \tag{39}$$

The diffusion generator is

$$\mathcal{L} = \sum_{i=1}^{d} b_i(x) \partial_i + \tfrac{1}{2} \sum_{i,j=1}^{d} a_{ij}(x) \partial_i \partial_j \,, \tag{40}$$

where ∂_i is shorthand notation for $\partial / \partial x_i$ (cf (4)). For diffusion processes, $\mathcal{L}$ is the generator of the semigroup of operators $P_{\Delta t}$ with $\Delta t \geq 0$

$$(\mathcal{L}\, h)(x) = \lim_{\Delta t \downarrow 0} (\Delta t)^{-1} [(P_{\Delta t}\, h)(x) - h(x)] \,, \tag{41}$$

and thus

$$P_{\Delta t} = \exp(\Delta t\, \mathcal{L}) \,. \tag{42}$$

This implies that the eigenfunction-eigenvalue pairs of $P_{\Delta t}$ and $\mathcal{L}$ are closely related, and we identify

$$P_{\Delta t}\,\phi = \Lambda\phi \quad \Leftrightarrow \quad \mathcal{L}\phi = \lambda\phi \;\text{ with }\; \Lambda = \exp(\lambda\,\Delta t)\;. \tag{43}$$

Note that this relation is exact and holds for all $\Delta t \geq 0$. A similar relation holds for the adjoints in $L^2(\Omega, dx)$ of $P_{\Delta t}$ and $\mathcal{L}$ and their eigenpairs, see Crommelin and Vanden-Eijnden (2011).

These relations can be used to estimate the drift and diffusion functions b and a that determine $\mathcal{L}$ in the following way: from discrete-in-time observations with sampling interval Δt we can infer (a Galerkin approximation to) the operator $P_{\Delta t}$. Denoting the eigenpairs of this estimated operator by $(\hat{\phi}_k, \hat{\Lambda}_k)$ with index k (ordered by decreasing $|\hat{\Lambda}_k|$), we compute $\hat{\lambda}_k = (\Delta t)^{-1}\log\hat{\Lambda}_k$, cf. (43). From the (leading) estimated eigenpairs $(\hat{\phi}_k, \hat{\lambda}_k)$ we can compute estimates of b and a by minimizing the residuals $r_k := \mathcal{L}\hat{\phi}_k - \hat{\phi}_k\lambda_k$, in a suitable way, under variation of b and a.

The minimization problem can be formulated with various cost functions. One example is a sum of squared norms $\sum_k \alpha_k \|r_k\|^2$ with weights α_k. Another possibility is $\sum_{k,l} |\langle r_k, \omega_l\rangle|^2$, where $\langle r_k, \omega_l\rangle$ denotes r_k integrated against suitable test functions $\omega_l(x)$. We refer to Crommelin and Vanden-Eijnden (2011) for more details. Here we only mention that this method can be used for parametric as well as for non-parametric estimation of b and a, and that the minimization problem is often of convex quadratic form.

This approach has the advantage that it can be used for general diffusions, and is not dependent on any small Δt approximation. It was used in Thompson *et al.* (2014) for estimating parameters in a 2-dimensional stochastic model for sea surface winds. Notwithstanding, this method is also limited to low-dimensional processes and since the estimation of the operator $P_{\Delta t}$ from observations becomes impractical for higher dimensional systems.

In Sitz *et al.* (2002), it is proposed to use a nonlinear extension of the well-known Kalman filter, the so-called unscented Kalman filter, for parameter estimation of a diffusion process with observation noise. This method is used in Kwasniok and Lohmann (2009) to estimate the parameters of a nonlinear drift function in a 1-dimensional model for glacial-interglacial transitions. The data are provided by an ice-core record from Greenland. This method can handle nonlinear drift, however it does not provide a way to estimate the diffusion coefficient. The diffusion must be estimated by a different method, and is used as input for the unscented Kalman filter approach.

Finally, we mention here the recent work presented in Peavoy *et al.* (2015), where a Bayesian framework is developed for parameter estimation using Markov Chain Monte Carlo (MCMC) methods, which is, in principle, applicable to high-dimensional systems. In this work, the structural form of the diffusion process is motivated by the stochastic mode reduction (MTV) methods for climate models discussed in the previous section. It builds on recent advances (e.g. Chib *et al.* (2004); Golightly and Wilkinson (2008)) in Bayesian inference and MCMC methods for diffusion processes, by imposing physical constraints such as global stability. This methodology can handle nonlinear drifts and non-constant diffusions, and is demonstrated on examples with dimensions of the state vector with $d = 1$ and $d = 2$ in Peavoy *et al.* (2015), but can be extended to higher dimensional problems.

3.3 Inference from multi scale data

Stochastic models are often used as coarse-grained models for phenomena that are generated by a complex dynamical system with many scales. An example is the model considered in Kwasniok and Lohmann (2009), where a diffusion process is inferred from Greenland ice-core data ($\delta^{18}\,O$ values) going back 120 000 years. These $\delta^{18}\,O$ data are a proxy for northern hemisphere (NH) temperatures in the past, whose dynamics are generated by the full climate system with its many temporal scales. Thus, a scalar stochastic model for the dynamics of NH temperature (or $\delta^{18}\,O$ values) over timescales of centuries and longer is inevitably a coarse-grained model. Such a model is aimed at capturing the correct long timescale behavior but not necessarily the dynamics on short timescales.

An important question in this context is whether statistical inference from data of a multi scale system will yield an accurate representation of the long timescale dynamics, or whether it will result in a model that mainly reflects the short timescale behavior. In Pavliotis and Stuart (2007), multi scale diffusion processes are considered which possess a well-defined coarse-grained diffusion process in the limit of large scale separation, obtained by averaging or homogenization techniques as described in Sections 2.3 and 2.4. Inferring the coarse-grained process from data of the full multi scale process, however, was shown to yield results that are very different from the analytically derived coarse-grained process, if the sampling interval Δt is too short. For estimation methods that rely on small Δt (e.g. the estimators in (38)), this can be particularly bothersome. These methods can be caught between Δt being too small (giving the "multi scale bias" discussed above) and Δt being too large (giving bias due to the time discretization error of the estimator). Mitchell and Gottwald (2012) show that with the estimators (38), one typically obtains a linear drift term if Δt is too large, even if it should be nonlinear. We refer to Pavliotis and Stuart (2007); Papavasiliou *et al.* (2009); Crommelin and Vanden-Eijnden (2011); Mitchell and Gottwald (2012); Azencott *et al.* (2013) for a more detailed discussion of these issues.

As a final remark, we mention that the dependency of estimation results on the sampling interval has been noted in atmosphere-ocean applications as well. In Penland and Sardeshmukh (1995), the so-called "tau-test" is introduced to test if the results from the LIM method are independent of the sampling interval (denoted τ in that paper, rather than Δt). Penland and Sardeshmukh (1995) argue that a dependency on the sampling interval points towards a possible inadequacy of the functional form of equation (27), and to, for example, possible nonlinearity of the underlying dynamics. In Berner (2005), the sampling interval dependence of results obtained with the estimators (38) is demonstrated numerically. Berner reports that sampling intervals between 1 and 6 days are suitable for inferring a low-order stochastic model for planetary wave behavior from GCM time series data. However, for choosing the sampling interval "there seems to be a trade-off between reproducing the non-Gaussianities in the PDF versus capturing the temporal aspects of planetary wave behavior" (quoting Berner (2005)). A shorter Δt gives a better reproduction of the PDF, whereas with a longer Δt the temporal decay of correlations is better captured.

Also relevant in this context is the study by DelSole (2000), who investigates to what extent stochastic models are able to capture the statistics of deterministic dynamical systems. He notes, and analyzes in considerable detail, the sampling interval dependency when fitting stochastic models to deterministic dynamical systems. A key observation by DelSole is that the shape of the normalized autocorrelation functions (ACFs) of deterministic and stochastic (Markov) models differ at short time lags (Δt). For deterministic models, the ACF must be of the form $1-\gamma\,(\Delta t)^2$ with γ a positive constant if Δt is small, and thus the derivative of the ACF with respect to Δt vanishes as $\Delta t \downarrow 0$. By contrast, the ACF of frequently used Markov stochastic processes is of the form $\exp(-\beta|\Delta t|)$, with decay rate $\beta > 0$, for small Δt. Its derivative does not vanish but tends to $-\beta$ as $\Delta t \downarrow 0$. One can fit a Markov stochastic process so that its ACF intersects the ACF of the deterministic system at a chosen lag Δt^*, but the two ACFs will not coincide at other lags. Thus, the fitted stochastic process will depend on the chosen lag.

3.4 Beyond diffusion processes

Diffusion processes driven by standard Wiener processes, such as (25), are not the only type of stochastic process used for modeling (aspects of) the climate system. A generalization is the class of diffusions driven by Lévy processes, as already discussed in section 2.5. Inference for Lévy processes is an active area of research in mathematical statistics (e.g. Jongbloed *et al.* (2005)), however it has not been used much for climate applications yet. Some exceptions, already mentioned in 2.5, are by Viecelli (1998); Ditlevsen (1999).

A class of stochastic processes that has been used more widely in atmosphere-ocean science is that of finite-state Markov chains. These have been employed, for example, to study the regime behavior in the atmosphere (Spekat *et al.*, 1983; Mo and Ghil, 1987; Crommelin, 2004). Therein the transitions between a finite number of preferred states of the large-scale atmospheric flow (so-called regimes) are modeled as a Markov chain. If a time series of such finite-state dynamics is given, it is straightforward to estimate the elements of the transition probability matrix (or stochastic matrix) that defines the Markov chain.

Another application where finite-state Markov chains have been used is stochastic parameterization of atmospheric convection, see e.g. Khouider *et al.* (2003, 2010); Dorrestijn *et al.* (2013, 2015b,a); Gottwald *et al.* (2015)). In this approach, the range of convective activity of any atmospheric model column is discretized into a few distinct states. The transitions between these states as time evolves can then be modeled as a Markov chain, with transition probabilities that depend on the large-scale state of the atmosphere. The discretization has been carried out in several ways, e.g. Dorrestijn *et al.* (2013) discretize the vertical turbulent fluxes of heat and moisture using a clustering method, whereas Khouider *et al.* (2010) and Dorrestijn *et al.* (2015b) employ a small number of cloud states for discretization. In Gottwald *et al.* (2015), it is the convective area fraction that is discretized. Furthermore, the transition probabilities are obtained in different ways, using either physical intuition (e.g Khouider *et al.* (2003, 2010)) or statistical inference (e.g. Dorrestijn *et al.* (2013, 2015b,a); Gottwald *et al.* (2015)).

In Pasmanter and Timmermann (2003), Markov chains are used for studying ENSO predictability. The influence of the seasonal cycle is accounted for by employing so-called cyclic Markov chains: twelve different stochastic matrices $\mathbf{m}_i$ $(i = 1, ..., 12)$ are constructed (or in fact estimated), each specifying the transition probabilities of the various states from month i to month $i + 1$. The transition probabilities from month i to the same month i one year later is then given by the product $\mathbf{M}_i = \mathbf{m}_{i-1}\mathbf{m}_{i-2}\cdots\mathbf{m}_1\mathbf{m}_{12}\mathbf{m}_{11}\cdots\mathbf{m}_i$. The $\mathbf{M}_i$ are again stochastic matrices. Furthermore, Pasmanter and Timmermann (2003) construct their Markov chains by equipartition of the data, resulting in transition probability matrices that are in fact doubly stochastic matrices, satisfying both $\sum_k \mathbf{m}_i(k,l) = 1\ \forall l$ and $\sum_l \mathbf{m}_i(k,l) = 1\ \forall k$. This facilitates the analysis of their model in terms of Floquet theory and information loss properties. Some of these concepts are also used in Crommelin (2004).

The Markov chains as described above have finite state spaces. They are frequently used to model the dynamics of continuous quantities, requiring the discretization of these quantities. In the framework of Hidden Markov Models (HMMs), this discretization is no longer needed. HMMs still employ a finite state Markov chain, however the observed quantities or time series are assumed to be generated by another process whose properties depend on the Markov chain state. That other process may be continuous or discrete in time. An example is the case where the observations are independent draws from a normal distribution. In this case, if we denote the observations by Y_t, we have $Y_t \sim \mathcal{N}(\mu_k, \sigma_k^2)$. The values of the mean μ_k and variance σ_k^2 change over time, they can take on a finite number of values $\{\mu_1, ..., \mu_K\}$ and $\{\sigma_1^2, ..., \sigma_K^2\}$. The index k changes randomly in time in accordance with a Markov chain with $(K \times K)$ stochastic matrix. The value of k is unobserved (or hidden).

In the context of atmosphere-ocean science, HMMs have been used to model precipitation (e.g. Zucchini and Guttorp (1991)) as well as dynamics of large-scale atmospheric flow (e.g. Franzke *et al.* (2008)). The statistical inference for HMMs with normal distributions as output is tractable through the expectation-maximization algorithm, see Franzke *et al.* (2008) for more details and references.

We conclude by pointing out two more lines of research relevant in the context of this section. Egger (e.g. Egger (2001)) has employed master equations inferred from time series to study atmospheric phenomena, an approach closely related to Markov chain modeling. Furthermore, Horenko (e.g. Horenko (2010)) has developed techniques to deal with non-stationarity in time series data, an issue left out of consideration in most studies on data-driven approaches. It must be emphasized that the overview of data-driven methods presented here is by no means exhaustive. We have mainly focused on diffusion processes and Markov chains here, leaving out e.g. time series methods (ARMA models etc) for the sake of brevity.

4 Outlook

In this chapter we have described current approaches to either systematically derive reduced order stochastic climate models or to extract the stochastic dynamics from observed data. The two approaches of data-driven models and analytic physics-based models are complementary. In practice, in the future both approaches should be combined where the Mori-Zwanzig formalism provides the functional form for model fitting to observed data. This will enable us to fit more complex models with the currently available amount of data. Majda and Harlim (2013), Peavoy *et al.* (2015) and Kondrashov *et al.* (2015) put forward such physics constrained approaches which are based on energy conservation and global stability. Such reduced order models are used in many practical applications like long-range climate forecasts (e.g. El Nino-Southern Oscillation (ENSO)) or weather and climate catastrophe modeling (Born and William, 2006).

Another area where stochastic approaches are actively investigated is that of parameterizations, i.e. simplified representations of spatially localized, small-scale physical processes such as atmospheric convection. The need for stochastic parameterizations in complex numerical weather and climate prediction models becomes ever more clearer. A recent study by Dawson and Palmer (2014) showed that the European Centre for Medium Range Weather Forecasts (ECMWF) model with a stochastic physics scheme performs as well as a purely deterministic model version at a much higher horizontal resolution. Hence, stochastic weather and climate models offer the potential of achieving more accurate simulations at a lower computational expense. However, most of the current stochastic parameterization approaches are mainly ad hoc schemes (Shutts, 2004, 2005; Berner *et al.*, 2009; Franzke *et al.*, 2015b). There is a pressing need to base these stochastic parameterizations on a more sound mathematical and physical footing. Current systematic approaches for doing this include Frederiksen et al. (chapter in this book and references therein), Khouider *et al.* (2003); Crommelin and Vanden-Eijnden (2008); Plant and Craig (2008); Khouider *et al.* (2010); Wouters and Lucarini (2012, 2013); Dolaptchiev *et al.* (2013); Grooms and Majda (2013), but more fundamental work in this area is clearly needed. See also the review by Franzke *et al.* (2015b). Moreover, the Mori-Zwanzig formalism shows that such parameterization schemes might have to take memory effects into account. Most schemes currently do not account for this, and the issue of memory effects has hardly been explored yet in the context of parameterizations. Some exceptions are Crommelin and Vanden-Eijnden (2008); Verheul and Crommelin (pear); Gottwald *et al.* (2015); Chorin and Lu (2015).

Weather and climate prediction models use high-performance computers. These computing systems are expected to reach soon limits regarding energy use and heat production. These issues led to attempts to use imprecise computational techniques or stochastic processors (Palmer, 2014). It is thought, that these techniques can be carried out on less energy consuming computer systems. However, the use of stochastic processors also needs a firm mathematical underpinning since the implementation of the stochastic noise produced by the processors need to be appropriate. This is another area for future research.

Stochastic approaches are also important for the analysis of observed data and the understanding of their characteristics. In particular, the detection and attribution of forced trends is an important current research topic. For instance, there is currently a debate going on in the climate science community whether climate variability is long-range dependent (LRD) or whether it is better better described as short-range dependent (SRD). LRD systems are able to produce more persistent stochastic trends than SRD systems. So, if the climate system is LRD but is investigated with SRD methods then one is likely to mistake a stochastic LRD trend for a significant externally forced trend. Hence, the detection and attribution of external trends is hampered in LRD systems. See chapters by Bunde et al. and Watkins for more details. This topic is critically discussed in the contemporary climate literature and many climate scientists are sceptical about whether the climate system is LRD because there is a lack of physical mechanisms explaining the LRD characteristic (Franzke *et al.*, 2015a). The Mori-Zwanzig formalism provides an explanation how memory potentially arises in the climate system. However, whether this can explain LRD needs further research.

Acknowledgments
We would like to thank Jeroen Wouters for a careful reading of the manuscript. GAG acknowledges funding from the Australian Research Council. The research of DTC is supported by the Netherlands Organisation of Scientific Research (NWO) through a Vidi grant. CF was supported by the German Research Foundation (DFG) through the cluster of excellence CliSAP (EXC177).

References

Anosov, D. V. (1960). Averaging in systems of ordinary differential equations with rapidly oscillating solutions. *Izv. Akad. Nauk SSSR Ser. Mat.*, **24**, 721–742.

Applebaum, D. (2009). *Lévy processes and stochastic calculus*, volume 116 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition.

Arnold, L. (2001). Hasselmann's program revisited: The analysis of stochasticity in deterministic climate models. In J.-S. v. S. P. Imkeller, editor, *Stochastic Climate Models*, volume 49 of *Progress in Probability*, pages 141–155. Birkhäuser, Boston.

Arnold, V., Kozlov, V., , and Neishtadt, A. (1993). *Mathematical Aspects of Classical and Celestial Mechanics*. Springer-Verlag, New York.

Azencott, R., Beri, A., Jain, A., and Timofeyev, I. (2013). Sub-sampling and parametric estimation for multiscale dynamics. *Communications in Mathematical Sciences*, **11**(4).

Baladi, V. and Smania, D. (2008). Linear response formula for piecewise expanding unimodal maps. *Nonlinearity*, **21**(4), 677.

Berner, J. (2005). Linking nonlinearity and non-gaussianity of planetary wave behavior by the fokker-planck equation. *Journal of the atmospheric sciences*, **62**(7), 2098–2117.

Berner, J., Shutts, G. J., Leutbecher, M., and Palmer, T. N. (2009). A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, **66**(3), 603–626.

Bishwal, J. P. (2008). *Parameter estimation in stochastic differential equations*. Springer Science & Business Media.

Born, P. and William, M. (2006). Catastrophe modeling in the classroom. *Risk Manage. Insur. Rev.*, **9**, 219–229.

Burnecki, K. and Weron, A. (2010). Fractional Lévy stable motion can model subdiffusive dynamics. *Phys. Rev. E*, **82**, 021130.

Burnecki, K., Magdziarz, M., and Weron, A. (2012). Identification and validation of fractional subdiffusion dynamics. In J. Klafter, S. C. Lim, and R. Metzler, editors, *Fractional Kinetics*, volume 27 of *Springer Series in Synergetics*, pages 331–352. World Scientific, Singapore.

Chechkin, A. and Pavlyukevich, I. (2014). Marcus versus Stratonovich for systems with jump noise. *Journal of Physics A: Mathematical and Theoretical*, **47**(34), 342001.

Chechkin, A. V., Metzler, R., Klafter, J., and Gonchar, V. Y. (2008). Introduction to the Theory of Lévy Flights. In R. Klages, G. Radons, and I. M. Sokolov, editors, *Anomalous Transport*, pages 129–162. Wiley-VCH Verlag GmbH & Co. KGaA.

Chekroun, Mickaël, D., Kondrashov, D., and Ghil, M. (2011). Predicting stochastic systems by noise sampling, and application to the El Niño-Southern Oscillation. *Proceedings of the National Academy of Sciences*, **108**(29), 11766–11771.

Chib, S., Pitt, M., and Shephard, N. (2004). Likelihood inference for diffusion driven state space models. Technical report, Working Paper. Department of Economics, Nuffield College, University of Oxford.

Chorin, A. and Hald, O. (2006). *Stochastic Tools in Mathematics and Science.* STAMS. Springer-Verlag, New York.

Chorin, A., Hald, O., and Kupferman, R. (2000). Optimal prediction and the Mori-Zwanzig representation of irreversible processes. *PNAS*, **97**, 2968–2973.

Chorin, A. J. and Lu, F. (2015). Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proceedings of the National Academy of Sciences*, **112**(32), 9804–9809.

Crommelin, D. and Vanden-Eijnden, E. (2008). Subgrid-scale parameterization with conditional markov chains. *Journal of the Atmospheric Sciences*, **65**(8), 2661–2675.

Crommelin, D. and Vanden-Eijnden, E. (2011). Diffusion estimation from multiscale data by operator eigenpairs. *Multiscale Modeling & Simulation*, **9**(4), 1588–1623.

Crommelin, D. T. (2004). Observed nondiffusive dynamics in large-scale atmospheric flow. *Journal of the atmospheric sciences*, **61**(19), 2384–2396.

Crommelin, D. T. and Vanden-Eijnden, E. (2006). Reconstruction of diffusions using spectral data from timeseries. *Communications in Mathematical Sciences*, **4**, 651 – 668.

Dawson, A. and Palmer, T. (2014). Simulating weather regimes: impact of model resolution and stochastic parameterization. *Climate Dynamics*, pages 1–17.

DelSole, T. (2000). A fundamental limitation of Markov models. *Journal of the Atmospheric Sciences*, **57**(13), 2158–2168.

Ditlevsen, P. D. (1999). Observation of α-stable noise induced millennial climate changes from an ice-core record. *Geophysical Research Letters*, **26**(10), 1441–1444.

Dolaptchiev, S., Achatz, U., and Timofeyev, I. (2013). Stochastic closure for local averages in the finite-difference discretization of the forced burgers equation. *Theoretical and Computational Fluid Dynamics*, **27**(3-4), 297–317.

Dorrestijn, J., Crommelin, D. T., Siebesma, A. P., and Jonker, H. J. (2013). Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data. *Theoretical and Computational Fluid Dynamics*, **27**(1-2), 133–148.

Dorrestijn, J., Crommelin, D. T., Siebesma, A. P., Jonker, H. J., and Selten, F. (2015a). Stochastic convection parameterization with markov chains in an intermediate complexity gcm. *Journal of the Atmospheric Sciences*, page under review.

Dorrestijn, J., Crommelin, D. T., Siebesma, A. P., Jonker, H. J., and Jakob, C. (2015b). Stochastic parameterization of convective area fractions with a multicloud model inferred from observational data. *Journal of the Atmospheric Sciences*, **72**(2), 854–869.

Duan, J. and Nadiga, B. T. (2007). Stochastic parameterization for large eddy simulation of geophysical flows. *Proc. Amer. Math. Soc.*, **135**(4), 1187–1196 (electronic).

Dubinkina, S. and Frank, J. (2007). Statistical mechanics of Arakawa's discretizations. *Journal of Computational Physics*, **227**(2), 1286 – 1305.

Dubinkina, S. and Frank, J. (2010). Statistical relevance of vorticity conservation in the Hamiltonian particle-mesh method. *Journal of Computational Physics*, **229**(7), 2634 – 2648.

Egger, J. (2001). Master equations for climatic parameter sets. *Climate dynamics*, **18**(1-2), 169–177.

Evans, D. J. and Morriss, G. P. (2008). *Statistical Mechanics of Nonequilibrium Liquids*. Cambridge University Press.

Frank, J. E. and Gottwald, G. A. (2013). Stochastic homogenization for an energy conserving multi-scale toy model of the atmosphere. *Physica D: Nonlinear Phenomena*, **254**, 46 – 56.

Franzke, C. and Majda, A. J. (2006). Low-order stochastic mode reduction for a prototype atmospheric GCM. *Journal of the Atmospheric Sciences*, **63**(2), 457–479.

Franzke, C., Majda, A. J., and Vanden-Eijnden, E. (2005). Low-order stochastic mode reduction for a realistic barotropic model climate. *Journal of the Atmospheric Sciences*, **62**(6), 1722–1745.

Franzke, C., Crommelin, D. T., Fischer, A., and Majda, A. J. (2008). A hidden Markov model perspective on regimes and metastability in atmospheric flows. *Journal of Climate*, **21**(8), 1740–1757.

Franzke, C., Osprey, S., Davini, P., and Watkins, N. (2015a). A dynamical systems explanation of the hurst phenomenom and atmospheric low-frequency variability. *Sci. Rep.*, **5**, 9068.

Franzke, C., O'Kane, T., Berner, J., Williams, P., and Lucarini, V. (2015b). Stochastic climate theory and modelling. *WIREs Climate Change*, **6**(1), 63–78.

Friedrich, R., Siegert, S., Peinke, J., Siefert, M., Lindemann, M., Raethjen, J., Deuschl, G., Pfister, G., *et al.* (2000). Extracting model equations from experimental data. *Physics Letters A*, **271**(3), 217–222.

Givon, D., Kupferman, R., and Stuart, A. (2004). Extracting macroscopic dynamics: Model problems and algorithms. *Nonlinearity*, **17**(6), R55–127.

Gobet, E., Hoffmann, M., Reiß, M., *et al.* (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics*, **32**(5), 2223–2253.

Golightly, A. and Wilkinson, D. J. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, **52**(3), 1674–1693.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix computations, 4th edition*. JHU Press.

Gottwald, G. A. and Melbourne, I. (2013a). A Huygens principle for diffusion and anomalous diffusion in spatially extended systems. *Proc. Natl. Acad. Sci. USA*, **110**, 8411–8416.

Gottwald, G. A. and Melbourne, I. (2013b). Homogenization for deterministic maps and multiplicative noise. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, **469**(2156).

Gottwald, G. A. and Melbourne, I. (2014). A test for a conjecture on the nature of attractors for smooth dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **24**(2), 024403.

Gottwald, G. A., Peters, K., and Davies, L. (2015). A data-driven method for the stochastic parametrisation of subgrid-scale tropical convective area fraction. *Quarterly Journal of the Royal Meteorological Society*.

Gouëzel, S. (2004). Central limit theorem and stable laws for intermittent maps. *Probability Theory and Related Fields*, **128**, 82–122.

Grigolini, P. (1982). A generalized Langevin equation for dealing with nonadditive fluctuations. *J. Stat. Phys.*, **27**, 283–316.

Grooms, I. and Majda, A. J. (2013). Efficient stochastic superparameterization for geophysical turbulence. *Proceedings of the National Academy of Sciences*, **110**(12), 4464–4469.

Hasselmann, K. (1976). Stochastic climate models. Part 1: Theory. *Tellus*, **28**(6), 473–485.

Hein, C., Imkeller, P., and Pavlyukevich, I. (2009). Limit theorems for p-variations of solutions of sdes driven by additive stable Lévy noise and model selection for paleo-climatic data. In J. Duan, S. Luo, and C. Wang, editors, *Recent Development in Stochastic Dynamics and Stochastic Analysis*, volume 8 of *Interdisciplinary Math. Sciences*, pages 137–150. World Scientific, Singapore.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. Siam.

Horenko, I. (2010). On the identification of nonstationary factor models and their application to atmospheric data analysis. *Journal of the Atmospheric Sciences*, **67**(5), 1559–1574.

Horsthemke, W. (1984). Noise induced transitions. In C. Vidal and A. Pacault, editors, *Non-Equilibrium Dynamics in Chemical Systems*, volume 27 of *Springer Series in Synergetics*, pages 150–160. Springer Berlin Heidelberg.

Ikeda, N. and Watanabe, S. (1981). *Stochastic Differential Equations and Diffusion Processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland, New York.

Jain, A., Timofeyev, I., and Vanden-Eijnden, E. (2015). Stochastic mode-reduction in models with conservative fast sub-systems. *Commun. Math. Sci.*, **13**(2), 297–314.

Jongbloed, G., Van Der Meulen, F. H., Van Der Vaart, A. W., *et al.* (2005). Nonparametric inference for lévy-driven ornstein-uhlenbeck processes. *Bernoulli*, **11**(5), 759–791.

Kelly, D. and Melbourne, I. (2014). Deterministic homogenization for fast-slow systems with chaotic noises. *preprint*.

Khasminsky, R. Z. (1966). On stochastic processes defined by differential equations with a small parameter. *Theory of Probability and its Applications*, **11**, 211–228.

Khouider, B., Majda, A. J., and Katsoulakis, M. A. (2003). Coarse-grained stochastic models for tropical convection and climate. *Proceedings of the National Academy of Sciences*, **100**(21), 11941–11946.

Khouider, B., Biello, J., Majda, A. J., *et al.* (2010). A stochastic multicloud model for tropical convection. *Communications in Mathematical Sciences*, **8**(1), 187–216.

Kifer, Y. (1992). Averaging in dynamical systems and large deviations. *Invent. Math.*, **110**(2), 337–370.

Kifer, Y. (1995). Limit theorems in averaging for dynamical systems. *Ergodic Theory Dynam. Systems*, **15**(6), 1143–1172.

Kifer, Y. (2001). Averaging and climate models. In J.-S. v. S. P. Imkeller, editor, *Stochastic Climate Models*, volume 49 of *Progress in Probability*, pages 171–188. Birkhäuser, Boston.

Kifer, Y. (2003). L^2 diffusion approximation for slow motion in averaging. *Stoch. Dyn.*, **3**(2), 213–246.

Kifer, Y. (2005). Another proof of the averaging principle for fully coupled dynamical systems with hyperbolic fast motions. *Discrete and Continuous Dynamical Systems*, **13**(5), 1187–1201.

Kondrashov, D., Kravtsov, S., Robertson, A. W., and Ghil, M. (2005). A hierarchy of data-based ENSO models. *Journal of Climate*, **18**(21), 4425–4444.

Kondrashov, D., Kravtsov, S., and Ghil, M. (2006). Empirical mode reduction in a model of extratropical low-frequency variability. *Journal of the Atmospheric Sciences*, **63**(7), 1859–1877.

Kondrashov, D., Chekroun, M. D., and Ghil, M. (2015). Data-driven non-Markovian closure models. *Physica D: Nonlinear Phenomena*, **297**, 33 – 55.

Kravtsov, S., Kondrashov, D., and Ghil, M. (2005). Multilevel regression modeling of nonlinear processes: derivation and applications to climatic variability. *J. Climate*, **18**, 4404–4424.

Kurtz, T. G. (1973). A limit theorem for perturbed operator semigroups with applications to random evolutions. *Journal of Functional Analysis*, **12**(1), 55–67.

Kutoyants, Y. A. (2004). *Statistical inference for ergodic diffusion processes*. Springer Science & Business Media.

Kwasniok, F. and Lohmann, G. (2009). Deriving dynamical models from paleoclimatic records: application to glacial millennial-scale climate variability. *Physical Review E*, **80**(6), 066104.

Leith, C. E. (1975). Climate response and fluctuation dissipation. *Journal of the Atmospheric Sciences*, **32**(10), 2022–2026.

Majda, A., Timofeyev, I., and Vanden-Eijnden, E. (2002). A priori tests of a stochastic mode reduction strategy. *Phys. D*, **170**(3-4), 206–252.

Majda, A., Timofeyev, I., and Vanden-Eijnden, E. (2006). Stochastic models for selected slow variables in large deterministic systems. *Nonlinearity*, **19**(4), 769.

Majda, A. J. and Harlim, J. (2013). Physics constrained nonlinear regression models for time series. *Nonlinearity*, **26**(1), 201.

Majda, A. J. and Yuan, Y. (2012). Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems. *Discrete and Continuous Dynamical Systems - Series B*, **17**(4), 1333–1363.

Majda, A. J., Timofeyev, I., and Vanden Eijnden, E. (1999). Models for stochastic climate prediction. *Proceedings of the National Academy of Sciences*, **96**(26), 14687–14691.

Majda, A. J., Franzke, C., and Crommelin, D. (2009). Normal forms for reduced stochastic climate models. *Proceedings of the National Academy of Sciences*, **106**(10), 3649–3653.

Melbourne, I. and Nicol, M. (2005). Almost sure invariance principle for nonuniformly hyperbolic systems. *Commun. Math. Phys.*, **260**, 131–146.

Melbourne, I. and Nicol, M. (2008). Large deviations for nonuniformly hyperbolic systems. *Trans. Amer. Math. Soc.*, **360**(12), 6661–6676.

Melbourne, I. and Nicol, M. (2009). A vector-valued almost sure invariance principle for hyperbolic dynamical systems. *Annals of Probability*, **37**, 478–505.

Melbourne, I. and Stuart, A. (2011). A note on diffusion limits of chaotic skew-product flows. *Nonlinearity*, **24**, 1361–1367.

Mitchell, L. and Gottwald, G. A. (2012). Data assimilation in slow-fast systems using homogenized climate models. *Journal of the Atmospheric Sciences*, **69**, 1359–1377.

Mo, K. C. and Ghil, M. (1987). Statistics and dynamics of persistent anomalies. *Journal of the Atmospheric Sciences*, **44**(5), 877–902.

Mori, H. (1965a). A continued-fraction representation of the time-correlation functions. *Prog. Theor. Phys.*, **34**, 399–416.

Mori, H. (1965b). Transport, collective motion, and Brownian motion. *Prog. Theor. Phys.*, **33**, 423–455.

Palmer, T. (2014). More reliable forecasts with less precise computations: a fast-track route to cloud-resolved weather and climate simulators? *Phil. Trans. R. Soc. A*, **372**, 20130391.

Papanicolaou, G. C. (1976). Some probabilistic problems and methods in singular perturbations. *Rocky Mountain Journal of Mathematics*, **6**(4), 653–674.

Papanicolaou, G. C. and Kohler, W. (1974). Asymptotic theory of mixing stochastic ordinary differential equations. *Comm. Pure Appl. Math.*, **27**, 641–668.

Papavasiliou, A., Pavliotis, G., and Stuart, A. (2009). Maximum likelihood drift estimation for multiscale diffusions. *Stochastic Processes and their Applications*, **119**(10), 3173–3210.

Pasmanter, R. and Timmermann, A. (2003). Cyclic markov chains with an application to an intermediate enso model. *Nonlinear Processes in Geophysics*, **10**, 197–210.

Pavliotis, G. and Stuart, A. (2007). Parameter estimation for multiscale diffusions. *Journal of Statistical Physics*, **127**(4), 741–781.

Pavliotis, G. and Stuart, A. (2008). *Multiscale Methods Averaging and Homogenization.* Texts in Applied Mathematics **53**, Springer.

Peavoy, D., Franzke, C. L., and Roberts, G. O. (2015). Systematic physics constrained parameter estimation of stochastic differential equations. *Computational Statistics & Data Analysis*, **83**, 182–199.

Penland, C. and Ewald, B. D. (2008). On modelling physical systems with stochastic models: diffusion versus Lévy processes. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **366**(1875), 2455–2474.

Penland, C. and Magorian, T. (1993). Prediction of nino 3 sea surface temperatures using linear inverse modeling. *Journal of Climate*, **6**(6), 1067–1076.

Penland, C. and Sardeshmukh, P. D. (1995). The optimal growth of tropical sea surface temperature anomalies. *Journal of climate*, **8**(8), 1999–2024.

Plant, R. and Craig, G. C. (2008). A stochastic parameterization for deep convection based on equilibrium statistics. *Journal of the Atmospheric Sciences*, **65**(1), 87–105.

Rao, B. P. (1999). *Statistical inference for diffusion type processes*. Arnold.

Sanders, J. and Verhulst, F. (1985). *Averaging methods in nonlinear dynamical systems*, volume 59 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

Shutts, G. J. (2004). A stochastic kinetic energy backscatter algorithm for use in ensemble prediction systems. *ECMWF Technical Memorandum*, **449**, available at http://www.ecmwf.int/publications.

Shutts, G. J. (2005). A stochastic kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, **131**(612), 3079–3102.

Siegert, S., Friedrich, R., and Peinke, J. (1998). Analysis of data sets of stochastic systems. *Physics Letters A*, **243**(5), 275–280.

Sitz, A., Schwarz, U., Kurths, J., and Voss, H. (2002). Estimation of parameters and unobserved components for nonlinear systems from noisy time series. *Physical review E*, **66**(1), 016210.

Sørensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, **72**(3), 337–354.

Spekat, A., Heller-Schulze, B., and Lutz, M. (1983). Über großwetter und markov-ketten. *Meteorologische Rundschau*, **36**(6), 243–248.

Sura, P. (2003). Stochastic analysis of southern and pacific ocean sea surface winds. *Journal of the atmospheric sciences*, **60**(4), 654–666.

Thompson, W. F., Monahan, A. H., and Crommelin, D. (2014). Parametric estimation of the stochastic dynamics of sea surface winds. *Journal of the Atmospheric Sciences*, **71**(9), 3465–3483.

Verheul, N. and Crommelin, D. (to appear). Data-driven stochastic representations of unresolved features in multiscale models. *Communications in Mathematical Sciences*.

Viecelli, J. A. (1998). On the possibility of singular low-frequency spectra and Lévy law persistence statistics in the planetary-scale turbulent circulation. *Journal of the Atmospheric Sciences*, **55**(5), 677–689.

Winkler, C. R., Newman, M., and Sardeshmukh, P. D. (2001). A linear model of wintertime low-frequency variability. part i: Formulation and forecast skill. *Journal of climate*, **14**(24), 4474–4494.

Wong, E. and Zakai, M. (1965). On the convergence of ordinary integrals to stochastic integrals. *Ann. Math. Statist.*, **36**, 1560–1564.

Wouters, J. and Lucarini, V. (2012). Disentangling multi-level systems: averaging, correlations and memory. *J. Stat. Mech.*, page 2012:PO3003.

Wouters, J. and Lucarini, V. (2013). Multi-level dynamical systems: connecting the ruelle response theory and the mori-zwanzig approach. *J. Stat. Mech.*, **151**(5), 850–860.

Young, L.-S. (1998). Statistical properties of dynamical systems with some hyperbolicity. *Annals of Mathematics*, **147**(3), 585–650.

Young, L.-S. (1999). Recurrence times and rates of mixing. *Israel Journal of Mathematics*, **110**(1), 153–188.

Young, L.-S. (2002). What are SRB measures, and which dynamical systems have them? *Journal of Statistical Physics*, **108**(5-6), 733–754.

Zhang, Y. and Held, I. M. (1999). A linear stochastic model of a gcm's midlatitude storm tracks. *Journal of the Atmospheric Sciences*, **56**(10), 3416–3435.

Zucchini, W. and Guttorp, P. (1991). A hidden markov model for space-time precipitation. *Water Resources Research*, **27**(8), 1917–1923.

Zwanzig, R. (1973). Nonlinear generalized Langevin equations. *J. Stat. Phys.*, **9**, 215–220.

Zwanzig, R. (2001). *Nonequilibrium Statistical Mechanics*. Oxford University Press, Oxford.