



# Variational Approach to Closure of Nonlinear Dynamical Systems: Autonomous Case

Mickaël D. Chekroun<sup>1,2</sup>  · Honghu Liu<sup>3</sup> · James C. McWilliams<sup>2</sup>

Received: 30 April 2019 / Accepted: 29 November 2019 / Published online: 14 December 2019  
© Springer Science+Business Media, LLC, part of Springer Nature 2019

## Abstract

A general approach for the derivation of nonlinear parameterizations of neglected scales is presented for nonlinear systems subject to an autonomous forcing. In that respect, dynamically-based formulas are derived subject to a free scalar parameter to be determined per mode to parameterize. For each high mode, this free parameter is obtained by minimizing a cost functional—a parameterization defect—depending on solutions from direct numerical simulation (DNS) but over short training periods of length comparable to a characteristic recurrence or decorrelation time of the dynamics. An important class of dynamically-based formulas, for our parameterizations to optimize, are obtained as parametric variations of manifolds approximating the invariant ones. To better appreciate the origins of the modified manifolds thus obtained, the standard approximation theory of invariant manifolds is revisited in Part I of this article. A special emphasis is put on backward–forward (BF) systems naturally associated with the original system, whose asymptotic integration provides the leading-order approximation of invariant manifolds. Part II presents then (i) the modifications of these approximating manifolds based also on integration of the same BF systems but this time over a finite time  $\tau$ , and (ii) the variational approach aimed at making an efficient selection of  $\tau$  per mode to parameterize. The parametric class of leading interaction approximation (LIA) of the high modes obtained this way, is completed by another parametric class built from the quasi-stationary approximation (QSA); close to the first criticality, the QSA is an approximation to the LIA, but it differs as one moves away from criticality. Rigorous results are derived that show that—given a cutoff dimension—the best manifolds that can be obtained through our variational approach, are manifolds which are in general no longer invariant. The minimizers are objects, called the optimal parameterizing manifolds (PMs), that are intimately tied to the conditional expectation of the original system, i.e. the best vector field of the reduced state space resulting from averaging of the unresolved variables with respect to a probability measure conditioned on the resolved variables. Applications to the closure of low-order models of Atmospheric Primitive Equations and Rayleigh–Bénard convection are then discussed. The approach is finally illustrated—in the context of the Kuramoto–Sivashinsky turbulence—as providing efficient closures without slaving for a cutoff scale  $k_c$  placed within the inertial

---

Communicated by Valerio Lucarini.

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s10955-019-02458-2>) contains supplementary material, which is available to authorized users.

---

Extended author information available on the last page of the article

range and the reduced state space is just spanned by the unstable modes, without inclusion of any stable modes whatsoever. The underlying optimal PMs obtained by our variational approach are far from slaving and allow for remedying the excessive backscatter transfer of energy to the low modes encountered by the LIA or the QSA parameterizations in their standard forms, when they are used at this cutoff wavelength.

**Keywords** Approximate invariance formulas · Backward–forward systems · Dynamical closure · Optimization · Parameterizing manifold

## 1 Introduction

A number of theories have been proposed to explain the phenomenon of turbulence in fluid dynamics, but none has been universally accepted. Landau [117] and Hopf [93] suggested that turbulence is the result of an infinite sequence of bifurcations, each adding another independent period to a quasi-periodic motion of increasingly greater complexity. More recently, it has been shown numerically that the original quasiperiodic Landau's view of turbulence, with the amendment of the inclusion of stochasticity, may be well suited to describe certain turbulent behavior [105], at least for the motion of large eddies. In the 1970s it has been theoretically argued and confirmed by many experiments that dynamical systems may exhibit strange attractors which result in chaotic but deterministic behavior after a (very) few bifurcations have taken place. Ruelle and Takens [151] and others have suggested this as a mechanism underlying turbulence. In realistic physical problems one is seldomly able to carry out the mathematics beyond the first or second bifurcation, in particular regarding the derivation of reduced equations that capture effectively the amplitude and frequency content of the bifurcated solutions [42,118]. Noteworthy is normal form reduction that have been carried for degenerate singularities with simultaneous onset of co-existing and possibly many instabilities, but still close to first criticality [4,40,59].

It is typical of many bifurcation problems that, as the condition for instability is exceeded, increasingly many modes become unstable. This circumstance considerably complicates an effective reduction because it often corresponds to going through higher-order bifurcations to reach possibly chaos, for which a failure of the *slaving principle* of the unresolved variables onto the resolved ones—mandatory for the success of standard reduction techniques—is typically observed.

Center manifold techniques [42,81,172] require such a slaving principle to provide an efficient reduction of the dynamics, and in that sense is reliable only in the vicinity of low-order bifurcations associated with the onset of instability. Center manifolds form a particular class of more general invariant manifolds associated with a fixed point, on which solutions obey *de facto* a slaving principle. A comprehensive treatment of the computational aspects relative to the underlying parameterizations can be found in [85]. The treatment in [85] is based on the so-called *parameterization method* [16–18] itself built upon the invariance equation (see Eq. (2.26) below) and the associated cohomological equations that the sought (slaving) parameterization solves at different orders. The parameterization method allows for efficient computations for not only the case of invariant manifolds associated with fixed points, but also for the cases of invariant tori for autonomous or quasi-periodically forced systems, averaging and periodic diffeomorphisms [27], invariant tori in Hamiltonian systems [85], as well as normally hyperbolic invariant tori. Other complementary approaches include

e.g. the Lyapunov–Schmidt reduction [77,125] and the Lyapunov–Perron method [88,125], as well as the usage of symmetries [77,83].

Despite the success for analyzing a broad class of bifurcations or detecting special solutions in dynamical systems such as quasi-periodic ones, these methods relying on invariant manifold theory, have failed to prove their efficiency for reducing complicated behaviors resulting from the presence of chaos. In a certain sense, the “story” of the inertial manifold (IM) constitutes perhaps an epitome of this failure. Despite appealing mathematical results showing existence of IMs for a broad class of dissipative systems [38,62,66,130,164], and convergence error estimates when e.g. slaving is not guaranteed to be satisfied (Approximate Inertial Manifold (AIM)) [48,52,98,131], early promises [55,64,65,95,96] have been challenged due to practical shortcomings pointed out for efficient closure by IMs or AIMs for turbulent flows and route chaos [46,68,72,80,87,97,137].

Essentially, the current IM theory [180] predicts that the underlying slaving of the high modes to the low modes, holds when the cutoff wavenumber,  $k_c$ , is taken sufficiently far within the dissipative range, especially in “strongly” turbulent regimes that correspond e.g. to the presence of many unstable modes. Still, as the AIM theory underlines, satisfactory closures may be expected to be derived for  $k_c$  corresponding to scales larger than what the IM theory predicts. Nevertheless, as one seeks to further decrease  $k_c$  within the inertial range, standard AIMs fail typically in providing relevant closures and one needs to rely on no longer a fixed cutoff but instead a dynamic one so as to avoid energy accumulation on the cutoff level [50,54,56].

In general, to aim at closing a given chaotic system at a fixed cutoff scale such that the neglected scales contain a non-negligible fraction of the energy,<sup>1</sup> makes, a priori, the closure problem difficult to address. This difficulty is often manifested by either an under- or over-parameterization of the small scales, i.e. a deficient or excessive parameterization of the small-scale energy, leading to an incorrect reproduction of the backscatter transfer of energy to the large scales [9,94,108,121,140]. Thus, a deficiency in the (nonlinear) parameterization of the high modes leads to errors in the backscatter transfer of energy which is due to nonlinear interactions between the modes, especially those near the cutoff scale. We can speak of an inverse error cascade, i.e. errors in the modeling of the parameterized (small) scales that contaminate gradually the larger scales, and may spoil severely the closure skills for the resolved variables.

To remedy such a pervasive issue, it is thus reasonable, given a cutoff scale to seek for nonlinear parameterizations (manifolds) that minimize as much as possible a defect of parameterization in order to reduce spurious backscatter transfer of energy to the large scales. Obviously such manifolds should coincide with the invariant ones as one approaches towards the first bifurcation.

This latter point explains the two-part structure of our article. We show here that an important class of dynamically-based formulas for our parameterizations are obtained as parametric variations of manifolds approximating the invariant ones. To better appreciate the origins of the modified manifolds thus obtained, the standard approximation theory of invariant manifolds is revisited in Part I of this article. A special emphasis is put on backward–forward (BF) systems naturally associated with the original system, whose asymptotic integration provides the leading-order approximation of invariant manifolds.

Part II presents then (i) the modifications of these approximating manifolds based also on integration of the same BF systems but this time over a finite time  $\tau$ , and (ii) the variational approach aimed at making an efficient selection of  $\tau$  per mode to parameterize, in order to

---

<sup>1</sup> Such as “cutting” within the inertial range of turbulence.

minimize a parameterization defect. The parametric class of *leading interaction approximation (LIA)* of the high modes obtained this way, is completed by another parametric class built from the *quasi-stationary approximation (QSA)*; close to the first criticality, the QSA is an approximation to the LIA, but differs as one moves away from criticality.

In this article our formulations are general, but our primary motivations are geophysical fluid dynamics, and our numerical illustrations are with simple systems of this type. With this in mind, we elaborate our approach for a broad class of ordinary differential equations (ODEs), that includes forced-dissipative systems of the form

$$\frac{dy}{dt} = Ay + B(y, y) + F, \quad y \in \mathbb{C}^N. \quad (1.1)$$

Here  $A$  denotes a linear  $N \times N$  matrix,  $B$  a quadratic nonlinearity (as in the fluid advection operator) and  $F$  a constant forcing, i.e. autonomous. Such systems with complex entries arise e.g. as equations for the perturbed variable around a mean state, when the latter are expressed in the eigenbasis  $\{e_j\}_{j=1}^N$  of the linearization at this mean state.

We decompose the phase space into the sum of the subspace,  $E_c$ , of resolved variables (“coarse-scale”), and the subspace,  $E_s$ , of unresolved variables (“small-scale”). In practice  $E_c$  is spanned by the first few eigenmodes with dominant real parts (e.g. unstable), and  $E_s$  by the rest. Within this framework, and given a cutoff dimension,  $m$  (i.e.  $\dim(E_c)=m$ ), we consider for systems such as (1.1) parametric families of nonlinear parameterizations of the form

$$\begin{aligned} H_\tau(\xi) &= \sum_{n \geq m+1} H_n(\tau_n, \xi) e_n, \quad \xi \in E_c, \\ \tau &= (\tau_{m+1}, \dots, \tau_N), \quad \tau_n \geq 0. \end{aligned} \quad (1.2)$$

The purpose is to dispose of parameterizations that cover situations of slaving between the resolved and unresolved variables as well as situations for which slaving is not expected to occur (e.g. far from criticality), as  $\tau$  is varied. In that respect, we aim at determining a family of parameterizations that include the leading-order approximation of invariant manifolds when the system is placed near the first bifurcation value. The theory of approximation of invariant manifolds revisited in Part I teaches us that such a family can be produced by finite time-integration of auxiliary BF systems derived from Eq. (1.1); see e.g. (2.29) and (4.12) below. This gives rise to the LIA class, for which taking the limit (under appropriate non-resonance conditions) of  $H_n(\tau_n, \xi)$  as  $\tau_n \rightarrow \infty$  provides the leading-order approximation of the invariant manifold; see Theorems 1 and 2 below.

We propose a variational approach to deal with situations far away from criticality. It consists of determining the optimal  $\tau_n$ -value,  $\tau_n^*$ , by minimizing (relevant) cost functionals that depend on solutions from direct numerical simulation (DNS) but over a training interval of length comparable to a characteristic recurrence or decorrelation time of the dynamics; see Sects. 5 and 6 below for applications.

Given a solution  $y(t)$  of Eq. (1.1) available over an interval  $I_T$  of length  $T$ , one such cost functional on which a substantial part of this article focuses on is given by the following parameterization defect

$$\mathcal{Q}_n(\tau_n, T) = \overline{|y_n(t) - H_n(\tau_n; y_c(t))|^2}. \quad (1.3)$$

Here  $\overline{(\cdot)}$  denotes the time-mean over  $I_T$  while  $y_n(t)$  and  $y_c(t)$  denote the projections onto the high-mode  $e_n$  and the reduced state space  $E_c$  of  $y(t)$ , respectively. Our goal is then to optimize  $\mathcal{Q}_n(\tau_n, T)$  by solving for each  $m+1 \leq n \leq N$ ,

$$\min_{\tau_n} Q_n(\tau_n, T). \tag{1.4}$$

This procedure corresponds to minimizing the variance of the residual error per high mode in case  $y_n$  and  $H_n$  are zero-mean, and to minimizing the residual error as measured in a least-square sense, in the general case.

Geometrically, as shown in Sect. 4.2 below, the graph of  $H_\tau$  gives rise to a manifold  $\mathfrak{M}_\tau$  that satisfies

$$\overline{\text{dist}(y(t), \mathfrak{M}_\tau)^2} \leq \sum_{n=m+1}^N Q_n(\tau_n, T), \tag{1.5}$$

where  $\text{dist}(y(t), \mathfrak{M}_\tau)$  denotes the distance of  $y(t)$  (lying on the attractor) to the manifold  $\mathfrak{M}_\tau$ .

Thus minimizing each  $Q_n(\tau_n, T)$  (in the  $\tau_n$ -variable) is a natural idea to enforce closeness of  $y(t)$  in a least-square sense to the manifold  $\mathfrak{M}_\tau$ . The left panel in Fig. 1 illustrates (1.5) for the  $y_n$ -component: The optimal parameterization,  $H_n(\tau_n^*, \xi)$ , minimizing (1.4) is shown; it illustrates a situation where the dynamics is transverse to it (i.e. absence of slaving) while  $H_n(\tau_n^*, \xi)$  provides the best (quadratic) parameterization in a least-square sense.

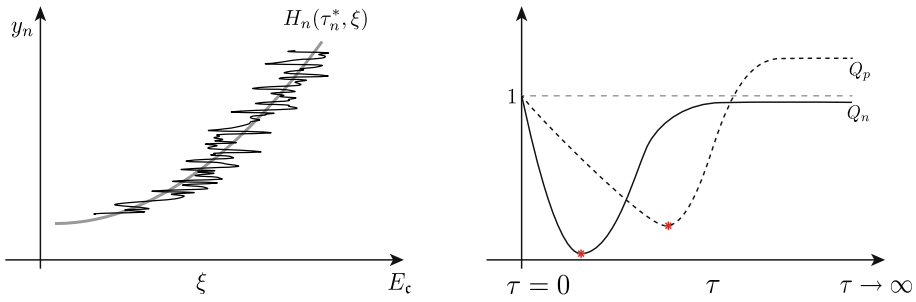
In practice, the following *normalized parameterizing defect* (for the  $n$ th mode),  $Q_n$ , is a useful tool to compare the different parameterizations  $H_n(\tau; \cdot)$  as  $\tau$  is varied. It is defined as

$$Q_n(\tau, T) = \frac{\overline{|y_n - H_n(\tau; y_c)|^2}}{\overline{|y_n|^2}}. \tag{1.6}$$

It provides a non-dimensional number to judge objectively of the quality of a parameterization. If  $Q_n(\tau, T) = 0$  for each  $n \geq m + 1$ , then  $H_\tau$  provides an exact slaving relation, and if  $H_n = 0$  i.e.  $H_\tau \equiv 0$ , corresponding to a standard Galerkin approximation, then  $Q_n(\tau, T) = 1$ . Thus, the notion of (normalized) parameterizing defect allows us to bring another perspective on criticisms brought to the (approximate) inertial manifold theory [72,90]: given a cutoff scale, if  $Q_n(\tau, T) > 1$  (over-parameterization) for several high modes, then a parameterization  $H_\tau$  may indeed lead to closure skills worse than those that would be obtained from a standard Galerkin scheme (cf.  $Q_p$  in Fig. 1; right). In other words, only a parameterization associated with a manifold that avoids such a situation is useful compared to a standard Galerkin scheme. This understanding alone is overlooked in the literature concerned with inertial manifolds and the like. We call such a manifold a *parameterizing manifold (PM)*; see Definition 1 for a precise characterization of a PM.

Minimizing the parameterization defects leads thus to an *optimal PM*, for the cost functionals  $Q_n$ . We emphasize that each component  $H_n$ , of the parameterization  $H_\tau$  given in (1.2), depends only on  $\tau_n$  (and not the other  $\tau_p$ 's for  $p \neq n$ ), and thus the cost functionals,  $Q_n$ , may be minimized independently from each other.

The parametric dependence on  $\tau$  of  $H_\tau$  is of practical importance. To understand this, let us consider for a moment a parameterization,  $H_n$ , given as a homogeneous quadratic polynomial of the  $m$ -dimensional  $\xi$ -variable with unknown coefficients (not depending on  $\tau_n$ ). To learn these coefficients via a standard regression would lead to  $m(m - 1)/2$  coefficients to estimate. Instead, adopting the parametric formulation given in (1.3), only the parameter  $\tau$  needs to be learned (per high-mode) in case each coefficient of  $H_n(\tau, \xi)$  is given by a function of  $\tau$ . This way, we benefit from a significant reduction of the amount  $N_T$  of snapshots  $y(t_k)$  required from numerical integration of Eq. (1.1) to obtain robust parameterizations (in a statistical sense). Roughly speaking, if  $N_T$  is smaller or comparable to  $m(m - 1)/2$ , then learning the unknown (and arbitrary) coefficients of a homogeneous quadratic parameterization (not given under the parametric form (1.3)) is either undetermined or not robust statistically.



**Fig. 1** Left panel: The optimal parameterization,  $H_n(\tau_n^*, \xi)$ , minimizing (1.4) is shown (in gray). Here the dynamics (black curve) is transverse to it (i.e. absence of slaving) while  $H_n(\tau_n^*, \xi)$  provides the best (quadratic) parameterization in a least-square sense. See Fig. 4 below for a concrete example in the case of a truncated Primitive Equation model due to Lorenz [123]. The parameter  $\tau_n^*$  corresponds to the argmin of  $Q_n$  (red asterisk) shown in the right panel. Right panel: Dependence on  $\tau$  shown for two parameterization defects  $Q_n$  and  $Q_p$  given by (1.6), with  $p, n \geq m + 1$ . The minimum is marked by a red asterisk (Color figure online)

Explicit formulas for the coefficients of  $H_n(\tau, \xi)$  are derived in Sects. 4.3 and 4.4 below. These formulas are dynamically-based in the sense that these coefficients involve structural elements of the right-hand side (RHS) of Eq. (1.1) such as the eigenvalues  $\beta_j$  of  $A$ , projections onto the  $n^{\text{th}}$  high-mode of nonlinear interactions  $B_{ij}^n$  between pairs of low eigenmodes ( $e_i, e_j$ ) of  $A$  ( $1 \leq i, j \leq m$ ), as well as possible nonlinear interactions between these modes and the forcing term.

For instance, for the LIA class, the coefficients of the  $H_n(\tau, \xi)$ 's monomials are given by  $D_{ij}^n(\tau)B_{ij}^n$  with

$$D_{ij}^n(\tau) = \frac{1 - e^{-\tau\delta_{ij}^n}}{\delta_{ij}^n}, \quad \tau > 0,$$

with  $\delta_{ij}^n = \beta_i + \beta_j - \beta_n$ . (1.7)

We emphasize that at an heuristic level, the coefficient  $D_{ij}^n(\tau)$  allows for balancing the denominator  $\delta_{ij}^n$  by the numerator  $1 - e^{-\tau\delta_{ij}^n}$  when the former is small. Such compensating  $\tau$ -factors are in general absent from parameterizations built from invariant manifold or (approximate) inertial manifolds techniques.

From the approximation theory of invariant manifolds revisited in Part I below, one notes that  $D_{ij}^n(\tau)$  is equal to  $1/\delta_{ij}^n$  in the case of standard approximation formulas of invariant manifolds (Theorem 2), corresponding thus to the asymptotic case  $\tau \rightarrow \infty$  if  $\delta_{ij}^n > 0$ . When adopting these approximation formulas outside their domain of applicability (i.e. not for approximating an underlying invariant manifold), it corresponds typically to small  $\delta_{ij}^n$ 's which without the compensating  $\tau$ -factors lead to an over-parameterization and an incorrect reproduction of the backscatter transfer of energy to the large scales. This problem is typically encountered in invariant manifold approximation when small spectral gaps are present, regardless of whether the solution dynamics is simple or complicated; see the Supplementary Material for a simple example. It turns out that, to seek for an optimal backward integration time  $\tau$  actually helps alleviate this problem by introducing numerators balancing the small denominators present in standard LIA parameterizations such as provided by Theorem 2 below.

At the same time,  $\tau = 0$  implies  $D_{ij}^n(\tau) = 0$ , which corresponds to the null parameterization, namely to a Galerkin approximation of dimension  $m$ . Thus, minimizing the  $Q_n$ 's gives

rise to an intermediate (and optimized) parameterization compared to a Galerkin approximation ( $H_n = 0$ ) or an invariant manifold approximation ( $Q_n = 0$ ).

The right panel in Fig. 1 shows a typical dependence on  $\tau$  of the  $Q_n$ 's defined in (1.6) for the LIA class. Similar dependences hold for the QSA class. On a practical ground, the minimization problem (1.4) is greatly facilitated by exploiting the explicit formulas of Sects. 4.3 and 4.4. An efficient minimization can be indeed operated by application of a simple gradient-descent algorithm in the real variable  $\tau$ , when the appropriate moments up to fourth order have been estimated; see Appendix.

We emphasize that the parameterization formulas of the LIA or QSA classes can be derived for dissipative nonlinear partial differential equations (PDEs) as well; see Sect. 6 below. The LIA class as rooted in the backward–forward method mentioned above was initially introduced for PDEs (possibly driven by a multiplicative linear noise) in [31, Chap. 4] and was applied to the closure of a stochastic Burgers equation in [31, Chaps. 6 & 7] and to optimal control in [26]. The main novelty compared to these previous works is the idea of optimizing per high mode the backward integration time,  $\tau_n$ , by minimization of the parameterization defect  $Q_n$ . Here, we also restrict ourselves to quadratic parameterizations that we prefer to optimize instead of computing higher-order terms that although being potentially useful make more cumbersome the numerical integration of the corresponding closure systems by adding too many extra terms in the RHS of the latter.

The justification of the variational approach proposed in this article relies on the ergodic theory of dissipative deterministic dynamical systems. In that respect, given the flow  $T_t$  associated with Eq. (1.1), we assume in Part II of this article that  $T_t$  possesses an invariant probability measure  $\mu$ , which is *physically relevant* [37,57], in the sense that time-average equals to ensemble average for trajectories emanating from Lebesgue almost every initial condition. More precisely, we say that the invariant measure,  $\mu$ , is physical if the following property holds for  $y$  in a positive Lebesgue measure set  $B(\mu)$  (of  $\mathbb{C}^N$ ) and for every continuous observable  $\varphi : \mathbb{C}^N \rightarrow \mathbb{C}$

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(T_t(y)) dt = \int \varphi(y) d\mu(y). \tag{1.8}$$

This property assures that meaningful averages can be calculated and the statistics of the dynamical system can be investigated by the asymptotic distribution of orbits starting from Lebesgue almost every initial condition in e.g. the basin of attraction  $B(\mu)$  of the statistical equilibrium,  $\mu$ .

It can be proven for e.g. Anosov flows [13], partially hyperbolic systems [1], Lorenz-like flows [12], and observed experimentally for many others [28,33,57,71] that a common feature of (dissipative) chaotic systems is the transformation (under the action of the flow) of the initial Lebesgue measure into a probability measure with finer and finer scales, reaching asymptotically an invariant measure  $\mu$  of Sinai–Ruelle–Bowen (SRB) type. This measure is singular with respect to the Lebesgue measure, is supported by the local unstable manifolds contained in the global attractor or the non-wandering set [37, Definition 6.14], and if it has no zero Lyapunov exponents it satisfies (1.8) [177]. This latter property is often referred to as the *chaotic hypothesis* that, roughly speaking, expresses an extension of the ergodic hypothesis to non-Hamiltonian systems [71].

At the core of our analysis, is the disintegration  $\mu_\xi$  of statistical equilibrium  $\mu$  with respect to the resolved variable  $\xi$  in  $E_c$ ; see [23, Sec. 3]. In our case, the probability measure  $\mu_\xi$  gives the conditional probability of the unresolved variables (in  $E_s$ ), contingent upon the value taken by the resolved variable  $\xi$ . Denoting by  $y_s(t)$  the high-mode projection of  $y(t)$ , Theorem 4 below shows, under a natural boundedness assumption on the 2nd-order moments,

that the optimal PM that minimizes the defect

$$Q_T(\Psi) = \overline{\|y_s(t) - \Psi(y_c(t))\|^2}, \tag{1.9}$$

with  $\Psi$  denoting a square-integrable mapping<sup>2</sup> from  $E_c$  to  $E_s$ , is given, when  $T \rightarrow \infty$ , by

$$\Psi^*(\xi) = \int_{E_s} \zeta \, d\mu_\xi(\zeta), \quad \xi \in E_c. \tag{1.10}$$

This formula shows that the optimal PM corresponds actually to the manifold that maps to each resolved variable  $\xi$  in  $E_c$ , the averaged value of the unresolved variable  $\zeta$  in  $E_s$  as distributed according to the conditional probability measure  $\mu_\xi$ . In other words, the optimal PM provides the best manifold (in a least-square sense) that averages out the fluctuations of the unresolved variable. The closure system that consists of approximating the unresolved variables by this optimal parameterization provides then, when the high-mode to high-mode interactions are small, the conditional expectation of the original system; see Theorem 5 below. The latter provides the best vector field of the reduced state space for which the effects of the unresolved variables are averaged out with respect to the probability measure  $\mu_\xi$  on the space of unresolved variables, itself conditioned on the resolved variables. For slow-fast systems, in the limit of infinite time-scale separation, it is well-known that the slow dynamics is approximated (on bounded time scales) by the conditional expectation of the multiscale system [100,101,138] and that slow trajectories may be obtained through a variational principle [119]. Nevertheless, the conditional expectation may be useful to approximate other global features of the multiscale dynamics when time-scale separation is lacking. For instance, the low-frequency variability dynamics may be well approximated for chaotic systems that do not exhibit distinguished fast variables but rather episodic bursts of fast oscillations punctuated by slow oscillations for each variable; see [32] and Sect. 3.4 below.

The optimal PM,  $\Psi^*$ , comes with a normalized parameterization defect,  $Q_T(\Psi^*) = Q_T(\Psi^*)/\overline{\|y_s(t)\|^2}$ , that satisfies necessarily (Theorem 4)

$$0 \leq \lim_{T \rightarrow \infty} Q_T(\Psi^*) \leq 1. \tag{1.11}$$

This variational view on the parameterization problem of the unresolved variables removes any sort of ambiguity that has surrounded the notion of (approximate) inertial manifold in the past. Indeed, within this paradigm shift, given an ergodic invariant measure  $\mu$  and a reduced dimension  $m$ , the optimal PM may have a parameterization defect very close to 1 and thus the best possible nonlinear parameterization one could ever imagine may not a priori do much better than a classical Galerkin approximation, and sometimes even worse. To the opposite, the smaller  $Q_T(\Psi^*)$  is (for  $T$  large), the better the parameterization. All sort of nuances are actually admissible, even when the parameterization defect is just below unity; see [32].

The parameterization defect analysis will be often completed by the evaluation of the *correlation parameterization*,  $c(t)$  (see (3.6)), that provides a measure of collinearity between the parameterized variable  $\Psi(y_c(t))$  and the unresolved variable  $y_s(t)$ , as time evolves. It allows thus for measuring how far from a slaving situation a given PM is on a more geometrical ground than with  $Q_T$  (Sect. 3.1). As we will see in applications, the parameterization correlation allows us, once an optimal PM has been determined, to select the dimension  $m$  of the reduced state space according to the following criterium:  $m$  should correspond to the lowest dimension of  $E_c$  for which the probability distribution function (PDF) of the corresponding *parameterization angle*,  $\alpha(t) = \arccos(c(t))$ , is the most skewed towards zero and

<sup>2</sup> With respect to the probability measure  $m$  obtained as a projection of  $\mu$  onto  $E_c$ .



the mode (i.e. the value that appears most often) of this PDF is the closest to zero. The basic idea is that one should not only parameterize properly the statistical effects of the neglected scales but also avoid to lose their phase relationships with the retained scales [132]. This is particularly important to derive closures that respect a certain phase coherence between the resolved and unresolved scales.

Although finite-time error estimates are easily accessible when PMs are used to derive surrogate low-dimensional systems in view of the optimal control of dissipative nonlinear PDEs (see e.g. [26, Theorem 1 & Corollary 2]), error estimates that relate the parameterization defect to the ability of reproducing the original dynamics's long term statistics by a surrogate system are difficult to produce for uncontrolled deterministic systems, in particular for chaotic regimes, due to the singular nature (with respect to the Lebesgue measure) of the invariant measure  $\mu$  satisfying (1.8). In the stochastic realm, this invariant measure becomes smooth for a broad class of systems and the tools of stochastic analysis make the obtention of such estimates more amenable albeit non trivial; see [21]. Nevertheless, as discussed above, considerations from ergodic theory and conditional expectations are already insightful for the deterministic systems dealt with in this article. They allow us to envision the addition of memory effects (non-Markovian terms) and/or stochastic parameterizations when a PM alone is not sufficient to provide an accurate enough closure. The addition of such ingredients are beyond the scope of this article, but are outlined in the Concluding Remarks (Sect. 7) as a natural direction to extend the present work. The latter sets up a framework for determining, via dynamically-based formulas to optimize, approximations of the Markovian terms arising in the Mori-Zwanzig formalism [34,79]; this formalism providing a conceptual framework to study the reduction of nonlinear autonomous systems.

The structure of this article is as follows. In Sect. 2 we revisit the approximation formulas of invariant manifolds for equilibria. The leading-order approximation  $h_k$  to these manifolds is obtained as the pullback limit of the high-mode part of the solution to an auxiliary backward–forward system (Theorem 1) and explicit formulas of  $h_k$  are derived (Theorem 2). The resulting invariant manifold approximation formulas are applied to an El Niño–Southern Oscillation ODE model in the Supplementary Material, in the case of a subcritical Hopf bifurcation. In Sect. 3, we introduce the measure-theoretic framework in which our variational approach is formulated. Theorem 4 characterizes the minimizers (optimal PMs) of the parameterization defect, and Theorem 5 shows that optimal PMs relate naturally to conditional expectations. As a first application, in Sect. 3.4 the closure results of [32] concerning the low-order model atmospheric Primitive Equations of [123], are enlightened by new insights introduced in this article. Building upon the backward–forward systems of Sect. 2, we derive in Sect. 4 parametric formulas of dynamically-based parameterizations aimed at being optimized.

Applications to the closure of a low-order model of Rayleigh–Bénard convection are then discussed in Sect. 5, for which a period-doubling regime and a chaotic regime are analyzed. In Sect. 6 the approach is finally illustrated—in the context of the Kuramoto–Sivashinsky turbulence—as providing efficient closures without slaving and for cutoff scales placed well within the inertial range, keeping only the unstable modes in the reduced state space. It is shown that the variational approach introduced in this article allows for fixing the excessive backscatter transfer of energy to the low modes encountered by standard parameterizations. We conclude in Sect. 7 by outlining future directions of research.

## Part I: Invariant Manifold Reduction Revisited

### 2 Approximation Formulas for Invariant Manifolds of Nonlinear ODEs

#### 2.1 Local Invariant Manifolds for Equilibria: Validity and Motivations for Other Parameterizations

Our framework takes place with autonomous systems of ordinary differential equations (ODEs) in  $\mathbb{R}^N$  of the form:

$$\frac{dY}{dt} = F(Y), \quad (2.1)$$

for which the vector field  $F$  is assumed to be sufficiently smooth in the state variable  $Y$ .

Invariant manifold theory allows for the rigorous derivation of low-dimensional surrogate systems from which not only the system's qualitative behavior near e.g. a steady state is preserved, but also quantitative features of the nonlinear dynamics are reasonably well approximated such as the solution's amplitude or possible dominant periods. This aspect of the theory is recalled below in the Supplementary Material, for the unfamiliar reader.

To set the ideas, assuming that  $\bar{Y}$  is a steady state of the system (2.1), we rewrite the system (2.1) in terms of the perturbed variable,  $y = Y - \bar{Y}$ , namely

$$\begin{aligned} \frac{dy}{dt} &= Ay + G(y), \quad \text{with} \\ A &= DF(\bar{Y}), \\ G(y) &= F(y + \bar{Y}) - Ay, \end{aligned} \quad (2.2)$$

where  $DF(x)$  denotes the Jacobian matrix of  $F$  at  $x$ .

From its definition, the nonlinear mapping,  $G: \mathbb{R}^N \rightarrow \mathbb{R}^N$ , satisfies

$$G(0) = 0, \quad \text{and} \quad DG(0) = 0. \quad (2.3)$$

As a consequence,  $G(y)$  admits the following expansion for  $y$  near the origin:

$$G(y) = G_k(\underbrace{y, \dots, y}_{k \text{ times}}) + O(\|y\|^{k+1}), \quad (2.4)$$

where

$$G_k: \underbrace{\mathbb{R}^N \times \dots \times \mathbb{R}^N}_{k \text{ times}} \rightarrow \mathbb{R}^N \quad (2.5)$$

denotes a homogenous polynomial of order  $k \geq 2$ . That is,  $G_k$  is the homogeneous part of lowest degree. Sometimes,  $G_k(y)$  will be used as a compact notation for  $G_k(y, \dots, y)$ .

The spectrum of  $A$  is denoted by  $\sigma(A)$ , i.e.

$$\sigma(A) = \{\beta_j \in \mathbb{C} : j = 1, \dots, N\}, \quad (2.6)$$

where the  $\beta_j$ s denote the eigenvalues of  $A$  for which we have accounted for their algebraic multiplicity in the sense that if  $\lambda$  is a root of multiplicity  $p$  of the characteristic polynomial  $\chi_A$ , then e.g.  $\beta_1 = \lambda, \dots, \beta_p = \lambda$ . The corresponding generalized eigenvectors are denoted by

$$\{e_j \in \mathbb{C}^N : j = 1, \dots, N\}. \quad (2.7)$$

The index in (2.6) also accounts for an arrangement of the eigenvalues in lexicographical order, that is the eigenvalues are ordered so that their real parts decrease as the index increases, and for eigenvalues with the same real parts, they are arranged so that the imaginary parts decrease.

Taking into account this ordering, grouping the first  $m$  eigenvalues of  $A$ , and assuming

$$\operatorname{Re}(\beta_m) \neq \operatorname{Re}(\beta_{m+1}), \tag{2.8}$$

the spectrum of  $A$  is decomposed as follows

$$\sigma(A) = \sigma_c(A) \cup \sigma_s(A), \tag{2.9}$$

where

$$\sigma_c(A) = \{\beta_j, j = 1, \dots, m\}, \tag{2.10}$$

and

$$\sigma_s(A) = \{\beta_j, j = m + 1, \dots, N\}. \tag{2.11}$$

Note that due to (2.8) and the aforementioned lexicographical order, we have

$$\operatorname{Re}(\beta_m) > \operatorname{Re}(\beta_{m+1}). \tag{2.12}$$

This spectral decomposition implies a natural decomposition of  $\mathbb{C}^N$ :

$$\mathbb{C}^N = E_c \oplus E_s, \tag{2.13}$$

in terms of the generalized eigenspaces

$$\begin{aligned} E_c &= \operatorname{span}\{e_j : j = 1, \dots, m\}, \\ E_s &= \operatorname{span}\{e_j : j = m + 1, \dots, N\}. \end{aligned} \tag{2.14}$$

This spectral decomposition of  $\mathbb{C}^N$  along with the corresponding canonical projectors  $\Pi_c$  and  $\Pi_s$  onto  $E_c$  and  $E_s$ , respectively, are at the core of our dimension reduction of Eq. (2.2).

The theory of local invariant manifolds for equilibria says that the simple condition (2.12) combined with the tangency condition (2.3) about the nonlinear term  $G$  ensure the existence of a local  $m$ -dimensional invariant manifold, namely a manifold obtained as the local graph over an open ball  $\mathfrak{B}$  in  $E_c$  centered at the origin, that is

$$\mathfrak{M} = \{\xi + h(\xi) : \xi \in \mathfrak{B} \subset E_c\}, \tag{2.15}$$

where  $h : E_c \rightarrow E_s$  is a  $C^1$ -smooth manifold function such that  $h(0) = 0$  and  $Dh(0) = 0$ , for which the following property holds:

- (i) any solution  $y(t)$  of Eq. (2.2) such that  $y(t_0)$  belongs to  $\mathfrak{M}$  for some  $t_0$ , stays on  $\mathfrak{M}$  over an interval of time  $[t_0, t_0 + \alpha)$ ,  $\alpha > 0$ , i.e.

$$y(t) = y_c(t) + h(y_c(t)), \quad t \in [t_0, t_0 + \alpha), \tag{2.16}$$

where  $y_c(t)$  denotes the projection of  $y(t)$  onto the subspace  $E_c$ .

Additionally, if  $\operatorname{Re}(\beta_{m+1}) < 0$  and  $\operatorname{Re}(\beta_m) \geq 0$ , then the local invariant manifold is the so-called local center-unstable manifold and the following property holds

- (ii) If there exists a trajectory  $t \mapsto y(t)$  such that  $y_c(t)$  belongs to  $\mathfrak{B}$  for all  $-\infty < t < \infty$ , then the trajectory must lie on  $\mathfrak{M}$ .

Property (ii) implies that an invariant set  $\Sigma$  of any type, e.g., equilibria, periodic orbits, invariant tori, must lie in  $\mathfrak{M}$  if its projection onto  $E_c$  is contained in  $\mathfrak{B}$ , i.e. if  $\Pi_c \Sigma \subset \mathfrak{B}$ . Property (2.16) holds then globally in time for the solutions that composed such invariant sets, and thus the knowledge of the  $m$ -dimensional variable,  $y_c(t)$ , is sufficient to entirely determine any solution  $y(t)$  that belongs to such an invariant set. Furthermore,  $y_c(t)$  is obtained as the solution of the following reduced  $m$ -dimensional problem

$$\frac{dx}{dt} = \Pi_c Ax + \Pi_c G(x + h(x)), \quad x(0) = y_c(0) \in \mathfrak{B}, \quad (2.17)$$

which in turn characterizes the solution  $y(t)$  in  $\Sigma$ , since the slaving relationship  $y_s(t) = h(y_c(t))$  holds for any solution  $y(t)$  that belongs to an invariant set  $\Sigma$  for which  $\Pi_c \Sigma \subset \mathfrak{B}$ .

More generally, property (i) allows for  $y_c(t)$  to leave the neighborhood  $\mathfrak{B}$  for some time instance,  $t$ , and thus to violate the parameterization (2.16) for  $y(t)$ , but does not exclude to have (2.16) to hold again over another interval  $[t_1, t_1 + \alpha_1]$  as soon as  $y(t_1)$  belongs to  $\mathfrak{M}$ .

Regarding the neighborhood  $\mathfrak{B}$ , the theory shows that it shrinks as the spectral gap,

$$\gamma_m = \operatorname{Re}(\beta_m) - \operatorname{Re}(\beta_{m+1}),$$

gets small and the nonlinear term  $G$  deviates quickly from the tangency condition as one moves away from the origin, leaving possible an (exact) parameterization only for solutions with sufficiently small amplitude. Indeed, the existence of such a (local) exact parameterization or say in other words, of a local  $m$ -dimensional invariant manifold is subject to the following *spectral gap condition*:

$$\gamma_m \geq CLip(G|_{\mathcal{V}}), \quad (2.18)$$

where  $Lip(G|_{\mathcal{V}})$  denotes the Lipschitz constant of the nonlinearity  $G$ , restricted to a neighborhood  $\mathcal{V}$  of the origin in  $\mathbb{C}^N$  such that  $\mathcal{V} \cap E_c = \mathfrak{B}$ , and  $C > 0$  is typically independent on  $\mathcal{V}$ . Due to the tangency condition (2.3), the condition (2.18) always holds once  $\mathcal{V}$  (and thus  $\mathfrak{B}$ ) is chosen sufficiently small. The theory of local invariant manifolds makes thus sense if solutions with sufficiently small amplitudes lie in the neighborhood  $\mathcal{V}$ . This situation is encountered for many bifurcations, near criticality for which the system's linear part has modes that become unstable, although a condition on the asymptotic stability of the origin is often required to have a local attractor that continuously unfolds from the origin as the bifurcation parameter is varied [125, Theorem 6.1]. In the context of e.g. nonlinear oscillations that bifurcate from a steady state, local invariant manifolds provide exact parameterizations<sup>3</sup> of stable limit cycles near criticality in the case of a supercritical Hopf bifurcation, whereas it is the parameterization of the unstable limit cycle that emerges continuously from the steady state that is guaranteed to be exact, at least sufficiently close to criticality in the case of a subcritical Hopf bifurcation. In the Supplementary Material, we show that the approximation formulas of Sect. 2.2, allow for approximating not only the unstable “inner” unstable limit cycle but also the “outer” stable limit cycle arising in an El Niño–Southern Oscillation (ENSO) model via subcritical Hopf bifurcation.

In any event, local invariant manifolds by their local nature, although useful in many applications do not allow for an efficient dimension reduction of arbitrary or at least generic solutions. Attempts to extend the theory to a more global setting, have failed dramatically to systematically provide nonlinear parameterizations of type (2.16) for a broader set of solutions, since, in general, the same type of spectral gap condition as (2.18) is also encountered in such an endeavor. For instance, the theory of inertial manifolds is known to be conditioned

<sup>3</sup> As provided for instance by a center manifold or the unstable manifold of the origin.

on spectral gap conditions such as given by (2.18) for which the Lipschitz constant is global or taken over a neighborhood  $\mathcal{V}$  that contains the (projection onto  $E_c$  of the) global attractor.

Part II proposes a new framework to provide manifolds which are no-longer locally invariant—and thus not subject to a spectral gap condition—but still provide meaningful nonlinear parameterizations of nonlinear dynamics; these manifolds being called *parameterizing manifolds (PMs)*. Nevertheless, the calculation of PMs departs from the theory of approximation of local invariant manifolds which we revisit in the next section, before presenting the main, new, analytical ingredients in Sect. 4.

The material presented in Sect. 2.2 below will serve to derive (approximate) parameterizations for perturbed variable taken with respect to a mean state  $\bar{Y}$ , instead of a steady state; see Sect. 4.3. To set the ideas, we consider  $F(Y)$  to be given by  $LY + B(Y, Y)$  with  $L$  linear, and  $B$  a quadratic homogeneous polynomial and symmetric,  $B(X, Y) = B(Y, X)$ . The equation for the perturbed variable  $y$  then becomes

$$\frac{dy}{dt} = (Ly + 2B(y, \bar{Y})) + B(y, y) + B(\bar{Y}, \bar{Y}), \tag{2.19}$$

which adopting the notations of Eq. (2.2), corresponds to  $A = Ly + 2B(y, \bar{Y})$  and  $G(y) = B(y, y) + L\bar{Y} + B(\bar{Y}, \bar{Y})$ . Since  $\bar{Y}$  is no longer a steady state,  $G(0) \neq 0$ , and  $L\bar{Y} + B(\bar{Y}, \bar{Y})$  is a time-independent forcing term. Thus the standard local invariant manifold theory for equilibria cannot be applied.

Nevertheless, as shown in Sect. 4 below, the theory underlying the derivation of approximation formulas for invariant manifolds is still relevant for their appropriate modification in view of providing approximate parameterizations in presence of forcing, once a good representation of these formulas is adopted; see Theorem 1 below for the representation of these approximation formulas (see (2.33)), and Sect. 4.3 for the modified parameterizations in presence of forcing.

### 2.2 Leading-Order Approximation of Invariant Manifolds

This section is devoted to the derivation of analytic formulas for the approximation of the (local) invariant manifold function  $h$  in (2.15). As shown below these formulas are easily obtained by relying only on the invariance property of  $\mathfrak{M}$ , responsible for the *invariance equation* to be satisfied by  $h$ . We recall first the derivation of this fundamental equation; see also [88, pp. 169–171] and [42, VII. A. 1]. For the existence of the invariant/center manifolds for ODEs, we refer to [172].

In that respect, note first that by applying respectively the projectors  $\Pi_c$  and  $\Pi_s$  on both sides of Eq. (2.2) and by using that  $A$  leaves invariant the eigensubspaces  $E_c$  and  $E_s$ , we obtain that Eq. (2.2) can be split as follows

$$\frac{dy_c}{dt} = A_c y_c + \Pi_c G(y_c + y_s), \tag{2.20a}$$

$$\frac{dy_s}{dt} = A_s y_s + \Pi_s G(y_c + y_s), \tag{2.20b}$$

with

$$y_c = \Pi_c y \in E_c, \quad y_s = \Pi_s y \in E_s, \quad A_c = \Pi_c A \quad \text{and} \quad A_s = \Pi_s A. \tag{2.21}$$

Since  $\mathfrak{M}$  is locally invariant, any solution  $y(t)$  of Eq. (2.2) with initial datum on  $\mathfrak{M}$  stays on  $\mathfrak{M}$  as long as  $y_c(t)$  stays in  $\mathcal{B}$  (where  $\mathcal{B}$  is given in (2.15)), i.e.

$$y(t) = y_c(t) + h(y_c(t)), \tag{2.22}$$

provided that  $y_c(t)$  lies in  $\mathcal{B}$ ; see (2.16).

This implies, as long as  $y_c(t)$  belongs to  $\mathcal{B}$ , that  $y_s(t) = h(y_c(t))$ , which, when substituted into Eq. (2.20b) gives

$$\frac{dh(y_c)}{dt} = A_s h(y_c) + \Pi_s G(y_c + h(y_c)). \tag{2.23}$$

On the other hand since  $h$  is differentiable, we have by using Eq. (2.20a),

$$\frac{dh(y_c)}{dt} = Dh(y_c) \frac{dy_c}{dt} = Dh(y_c)[A_c y_c + \Pi_c G(y_c + h(y_c))]. \tag{2.24}$$

Then (2.23) and (2.24) allow us to conclude that as long as  $y_c(t)$  belongs to  $\mathcal{B}$ ,  $h$  evaluated along the corresponding “segment” of trajectory satisfies

$$\begin{aligned} Dh(y_c(t))[A_c y_c(t) + \Pi_c G(y_c(t) + h(y_c(t)))] - A_s h(y_c(t)) \\ = \Pi_s G(y_c(t) + h(y_c(t))), \end{aligned} \tag{2.25}$$

which can be recast into the aforementioned *invariance equation* to be satisfied by  $h$ , namely

$$Dh(\xi)[A_c \xi + \Pi_c G(\xi + h(\xi))] - A_s h(\xi) = \Pi_s G(\xi + h(\xi)), \quad \xi \in \mathcal{B}. \tag{2.26}$$

This functional equation is a nonlinear system of first order PDEs that cannot be solved in closed form except in special cases. However, one can solve Eq. (2.26) approximately by representing  $h(\xi)$  as a formal power series. The solution is thus sought in terms of Taylor expansion in the  $\xi$ -variable and various numerical techniques—based, e.g., on the resolution of the multilinear Sylvester equations associated with the invariance equation—have been proposed in the literature to find the corresponding coefficients [10,58]. Once a power series approximation has been found, *a posteriori* error estimates can be checked by applying for instance [19, Theorem 3, p. 5].<sup>4</sup>

For a broad class of systems, the leading-order approximation of  $h$  can be efficiently and analytically calculated. It consists of dropping in Eq. (2.26) the terms involving nonlinear dependence on  $h$ . This operation leads to the following equation for the corresponding leading-order approximation  $h_k$  (see, e.g., [30,88]):

$$Dh_k(\xi)A_c \xi - A_s h_k(\xi) = \Pi_s G_k(\xi), \tag{2.27}$$

where  $G_k$  is the leading-order term in the Taylor expansion of  $G$  about the origin; cf. Eq. (2.4).

Easily checkable conditions on the eigenvalues of  $A$ , allows then for guaranteeing an analytic solution to Eq. (2.27). For instance, in the case  $A$  is self-adjoint, it simply requires certain *cross non-resonance conditions* to be satisfied as stated in Theorem 2 below. Namely, for any given set of resolved modes for which their self-interactions (through the leading-order nonlinear term  $G_k$ ) do not vanish when projected against an unresolved mode  $e_n$ , it is required that some specific linear combinations of the corresponding eigenvalues dominate the eigenvalue associated with  $e_n$ ; see (NR) below.

In the general case, when  $A$  is not necessarily diagonal, the cross non-resonance condition is strengthened to the requirement that  $\text{Re}(\beta_{m+1}) < k \text{Re}(\beta_m)$  which ensures that the following Lyapunov–Perron integral  $\mathfrak{J}: E_c \rightarrow E_s$ ,

$$\mathfrak{J}(\xi) = \int_{-\infty}^0 e^{-sA_s} \Pi_s G_k(e^{sA_c} \xi) ds, \tag{2.28}$$

<sup>4</sup> According to this theorem, a candidate to a (truncated) Taylor expansion has to be first determined, and then it has to be checked to satisfy the invariance equation up to some order to ensure to be a genuine Taylor approximation; see also [88, Thm. 6.2.3].

is well defined and in fact provides a solution  $h_k$  to Eq. (2.27); see Theorem 1 below. This solutions provides actually the leading-order approximation of the (local) invariant manifold function  $h$  if we assume furthermore that  $\text{Re}(\beta_{m+1}) < \min\{2k\text{Re}(\beta_m), 0\}$ ; see Theorem 1 again.

This Lyapunov–Perron integral itself possesses a flow interpretation: it is obtained as the pullback limit constructed from the solution of the following backward–forward auxiliary system

$$\frac{dy_c^{(1)}}{ds} = A_c y_c^{(1)}, \quad s \in [-\tau, 0], \tag{2.29a}$$

$$\frac{dy_s^{(1)}}{ds} = A_s y_s^{(1)} + \Pi_s G_k(y_c^{(1)}), \quad s \in [-\tau, 0], \tag{2.29b}$$

$$\text{with } y_c^{(1)}(s)|_{s=0} = \xi, \text{ and } y_s^{(1)}(s)|_{s=-\tau} = 0. \tag{2.29c}$$

Indeed, the solution to Eq. (2.29b) at  $s = 0$  is given by

$$h_\tau^{(1)}(\xi) = y_s^{(1)}[\xi](0; -\tau) = \int_{-\tau}^0 e^{-sA_s} \Pi_s G_k(e^{sA_c} \xi) ds, \tag{2.30}$$

and taking the limit formally in (2.30) as  $\tau \rightarrow \infty$ , leads to  $\mathfrak{J}$  given by (2.28).

The theorem below states more precisely the relationships between Eq. (2.27), the Lyapunov–Perron integral (2.28), and the solution to the backward–forward system (2.29).

**Theorem 1** Consider Eq. (2.2). Let the subspaces  $E_c$  and  $E_s$  be given by (2.14) and let  $m$  be the dimension of  $E_c$ . Assume (2.12) and furthermore that

$$\text{Re}(\beta_{m+1}) < k \text{Re}(\beta_m), \tag{2.31}$$

where  $k$  denotes the leading order of the nonlinearity  $G$ ; cf. (2.4).

Then, the Lyapunov–Perron integral

$$\mathfrak{J}(\xi) = \int_{-\infty}^0 e^{-sA_s} \Pi_s G_k(e^{sA_c} \xi) ds, \quad \xi \in E_c, \tag{2.32}$$

is well defined and is a solution to Eq. (2.27). Moreover,  $\mathfrak{J}$  is the pullback limit of the high-mode part of the solution to the backward–forward system (2.29):

$$\mathfrak{J}(\xi) = \lim_{\tau \rightarrow \infty} y_s^{(1)}[\xi](0; -\tau), \tag{2.33}$$

where  $y_s^{(1)}[\xi](0; -\tau)$  denotes the solution to Eq. (2.29b) at  $s = 0$ .

Finally, if we assume furthermore that

$$\text{Re}(\beta_{m+1}) < \min\{2k\text{Re}(\beta_m), 0\}, \tag{2.34}$$

then  $\mathfrak{J}$  provides the leading-order approximation of the invariant manifold function  $h$  in the sense that

$$\|\mathfrak{J}(\xi) - h(\xi)\|_{E_s} = o(\|\xi\|_{E_c}^k), \quad \xi \in E_c. \tag{2.35}$$

**Proof** First, we outline how condition (2.31) combined with the fact that  $G_k$  is a homogeneous polynomial of order  $k$ , ensure that the Lyapunov–Perron integral  $\mathfrak{J}$  is well defined. In that respect, we note first that natural estimates about  $\|e^{tA_s} \Pi_s\|_{L(\mathbb{C}^N)}$  and  $\|e^{tA_c} \Pi_c\|_{L(\mathbb{C}^N)}$  hold.

This is essentially a consequence of (2.12). Indeed, any choice of real constants  $\eta_1$  and  $\eta_2$  such that

$$\text{Re}(\beta_m) > \eta_1 > \eta_2 > \text{Re}(\beta_{m+1}), \tag{2.36}$$

ensures the existence of a constant  $K > 0$  (depending on  $\eta_1$  and  $\eta_2$ ) such that the following estimates hold:

$$\begin{aligned} \|e^{tA_c} \Pi_c\|_{L(\mathbb{C}^N)} &\leq K e^{\eta_1 t}, \quad \forall t \leq 0, \\ \|e^{tA_s} \Pi_s\|_{L(\mathbb{C}^N)} &\leq K e^{\eta_2 t}, \quad \forall t \geq 0. \end{aligned} \tag{2.37}$$

The latter inequalities resulting essentially from the fact that  $\|e^{tB}\|_{L(\mathbb{C}^N)}$  is bounded for  $t \geq 0$  if  $\text{Re} \lambda < 0$  for all  $\lambda$  in  $\sigma(B)$ .

Since  $G_k$  is a homogeneous polynomial of order  $k$ , there exists  $C > 0$  such that

$$\|G_k(\xi)\| \leq C \|\xi\|^k, \quad \forall \xi \in E_c. \tag{2.38}$$

Now, by using (2.37) and (2.38), we obtain for each  $s \leq 0$  that

$$\begin{aligned} \|e^{-sA_s} \Pi_s G_k(e^{sA_c} \xi)\| &\leq K e^{-s\eta_2} \|G_k(e^{sA_c} \xi)\| \\ &\leq C K e^{-s\eta_2} \|e^{sA_c} \xi\|^k \\ &\leq C K^2 e^{-s(\eta_2 - k\eta_1)} \|\xi\|^k. \end{aligned}$$

Assumption (2.31) allows us to choose  $\eta_1$  and  $\eta_2$  in (2.36) such that  $\eta_2 - k\eta_1 < 0$  which in turns leads to

$$\begin{aligned} \left\| \int_{-\infty}^0 e^{-sA_s} \Pi_s G_k(e^{sA_c} \xi) \, ds \right\| &\leq \int_{-\infty}^0 \|e^{-sA_s} \Pi_s G_k(e^{sA_c} \xi)\| \, ds \\ &\leq C K^2 \|\xi\|^k \int_{-\infty}^0 e^{-s(\eta_2 - k\eta_1)} \, ds \\ &= \frac{C K^2 \|\xi\|^k}{k\eta_1 - \eta_2}, \quad \forall \xi \in E_c. \end{aligned} \tag{2.39}$$

We have thus shown that  $\mathfrak{J}$  is well defined.

We show next that  $\mathfrak{J}$  satisfies Eq. (2.27). To do so, for any  $\xi$  in  $E_c$  we introduce the following function

$$\begin{aligned} \psi : (-\infty, 0] &\rightarrow E_s \\ t &\mapsto \mathfrak{J}(e^{tA_c} \xi) = \int_{-\infty}^t e^{(t-s)A_s} \Pi_s G_k(e^{sA_c} \xi) \, ds. \end{aligned} \tag{2.40}$$

On one hand, by differentiating  $\psi(t) = \int_{-\infty}^t e^{(t-s)A_s} \Pi_s G_k(e^{sA_c} \xi) \, ds$ , we obtain

$$\frac{d\psi}{dt} = \Pi_s G_k(e^{tA_c} \xi) + A_s \int_{-\infty}^t e^{(t-s)A_s} \Pi_s G_k(e^{sA_c} \xi) \, ds. \tag{2.41}$$

On the other, using that  $\psi(t) = \mathfrak{J}(e^{tA_c} \xi)$ , we have

$$\frac{d\psi}{dt} = D\mathfrak{J}(e^{tA_c} \xi) A_c e^{tA_c} \xi. \tag{2.42}$$

It follows then that

$$D\mathfrak{J}(e^{tA_c} \xi) A_c e^{tA_c} \xi = \Pi_s G_k(e^{tA_c} \xi) + A_s \int_{-\infty}^t e^{(t-s)A_s} \Pi_s G_k(e^{sA_c} \xi) \, ds, \quad \forall t \leq 0. \tag{2.43}$$



Set  $t = 0$  in the above equality, we then obtain

$$D\mathcal{J}(\xi)A_c\xi = \Pi_s G_k(\xi) + A_s \int_{-\infty}^0 e^{-sA_s} \Pi_s G_k(e^{sA_c}\xi) ds, \quad \forall \xi \in E_c,$$

which is equivalent to

$$D\mathcal{J}(\xi)A_c\xi - A_s\mathcal{J}(\xi) = \Pi_s G_k(\xi), \quad \forall \xi \in E_c.$$

We have thus verified that  $\mathcal{J}$  is a solution to Eq. (2.27).

Recall from Eq. (2.30) that the high-mode part of the solution to the backward–forward system (2.29) is given (at  $s = 0$ ) by:

$$y_s^{(1)}[\xi](0; -\tau) = \int_{-\tau}^0 e^{-sA_s} \Pi_s G_k(e^{sA_c}\xi) ds, \tag{2.44}$$

By using the same type of estimates as in (2.39), it is easy to show that the limit,  $\lim_{\tau \rightarrow \infty} y_s^{(1)}[\xi](0; -\tau)$ , exists and it is equal to  $\mathcal{J}(\xi)$ .

The leading-order approximation property stated in (2.35) under the assumption (2.34) is a direct consequence of the general result [30, Corollary 7.1] proved for stochastic evolution equations in infinite dimension, driven by a multiplicative white noise which thus applies to our finite dimensional and deterministic setting. Indeed, to apply [30, Corollary 7.1], we are only left with the checking of constants  $\eta_1$  and  $\eta_2$  for which [30, condition (7.1)] is verified, namely

$$\eta_s < \eta_2 < \eta_1 < \eta_c, \quad \eta_2 < 2k\eta_1 < 0, \tag{2.45}$$

with  $\eta_s = \text{Re}(\beta_{m+1})$  and  $\eta_c = \text{Re}(\beta_m)$  here. One can readily check that this condition is guaranteed under the assumptions (2.12) and (2.34). Indeed, if  $\text{Re}(\beta_{m+1}) < 2k\text{Re}(\beta_m) < 0$ , we just need to choose

$$\eta_1 = \text{Re}(\beta_m) - \epsilon \text{ and } \eta_2 = \text{Re}(\beta_{m+1}) + \epsilon,$$

with sufficiently small positive  $\epsilon$ ; and if  $\text{Re}(\beta_{m+1}) < 0 < 2k\text{Re}(\beta_m)$ , we just need to choose  $\eta_1 = -\epsilon$  and  $\eta_2 = \text{Re}(\beta_{m+1}) + \epsilon$  with again  $\epsilon$  sufficiently small.  $\square$

The next Theorem shows, under a slightly relaxed spectral condition (see (NR) below), that if the matrix  $A$  is assumed to be diagonal, then even when the Lyapunov–Perron integral (2.32) is no longer defined, a solution  $h_k$  to Eq. (2.27) can still be derived and that this solution possesses even an explicit expression.

This expression consists of an expansion in terms of the eigenvectors  $e_n$  lying in the eigenspace  $E_s$ , and whose coefficients are homogeneous polynomials of order  $k$  in the  $\xi$ -variable lying in eigenspace  $E_c$ ; the coefficients of these polynomials being themselves expressed in terms of ratios between the linear combinations of eigenvalues of  $A$  and the corresponding eigenmodes interactions through the leading-order nonlinear term  $G_k$ ; see (2.48). More precisely, we have

**Theorem 2** *Consider Eq. (2.2). Let the subspaces  $E_c$  and  $E_s$  be given by (2.14) and let  $m$  be the dimension of  $E_c$ . Assume (2.12) and that the matrix  $A$  is diagonal under its eigenbasis  $\{e_j \in \mathbb{C}^N : j = 1, \dots, N\}$ . We denote by  $\{e_j^*, j = 1, \dots, N\}$  the eigenvectors of the conjugate transpose  $A^*$ .*

*Recalling that  $G_k$  denotes the leading-order homogeneous polynomial in the expansion of  $G$  (see (2.4)), let us assume furthermore that the eigenvalues  $\beta_j$  of  $A$  satisfies the following cross non-resonance condition:*

$$\forall (i_1, \dots, i_k) \in \mathcal{I}^k, n \in \{m + 1, \dots, N\}, \text{ it holds that} \tag{NR}$$

$$\left( \langle G_k(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_k}), \mathbf{e}_n^* \rangle \neq 0 \right) \implies \left( \sum_{j=1}^k \beta_{i_j} - \beta_n \neq 0 \right),$$

where  $\mathcal{I} = \{1, \dots, m\}$ , and  $\langle \cdot, \cdot \rangle$  denotes the inner product on  $\mathbb{C}^N$  defined by

$$\langle a, b \rangle = \sum_{i=1}^N a_i \bar{b}_i, \quad a, b \in \mathbb{C}^N. \tag{2.46}$$

Then, a solution to Eq. (2.27) exists, and is given by

$$h_k(\xi) = \sum_{n=m+1}^N h_{k,n}(\xi) \mathbf{e}_n, \quad \xi = (\xi_1, \dots, \xi_m) \in E_c, \tag{2.47}$$

where  $h_{k,n}(\xi)$  is a homogeneous polynomial of degree  $k$  in the variables  $\xi_1, \dots, \xi_m$  given by

$$h_{k,n}(\xi) = \sum_{(i_1, \dots, i_k) \in \mathcal{I}^k} \frac{\langle G_k(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_k}), \mathbf{e}_n^* \rangle}{\sum_{j=1}^k \beta_{i_j} - \beta_n} \xi_{i_1} \dots \xi_{i_k}. \tag{2.48}$$

**Remark 1** (i) The formulas (2.47)–(2.48) for the case of real and symmetric matrices, are known; see e.g. [126, Appendix A]. The result presented in Theorem 2 extends nevertheless these formulas to cases for which  $A$  is diagonalizable in  $\mathbb{C}$ , allowing in particular for an arbitrary number of complex conjugate eigenpairs. The case when the neutral/unstable modes correspond to a single complex conjugate pair has been dealt with in [126, Appendix A]. Even in this special case, our formulas are in contradistinction simpler than those given in [126, Eq. (A.1.15)]. This is due to the use of generalized eigenvectors adopted here and the method of proof of Theorem 2 which relies on the calculation of spectral elements of the homological operator  $\mathcal{L}_A$  naturally associated with Eq. (2.27); see (2.54) below.

- (ii) The case of eigenvalues of higher-order multiplicity is more involved. The presence of Jordan blocks makes indeed the derivation of general analytic formulas challenging but still possible by the method used in the derivation of the formulas (2.47)–(2.48). Communication about these formulas will be pursued elsewhere.
- (iii) By only assuming the (NR) condition, the solution to Eq. (2.27) given by the formulas (2.47)–(2.48) is not necessarily unique. This situation happens for instance when we have a  $k$ -uple  $(i_1, \dots, i_k)$  and an index  $n$  for which  $\langle G_k(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_k}), \mathbf{e}_n^* \rangle = 0$  while  $\sum_{j=1}^k \beta_{i_j} - \beta_n = 0$ . In this case, we can add to any solution  $h_k$  to Eq. (2.27) a monomial  $cx_{i_1} \dots x_{i_k}$  with any scalar coefficient  $c$  and get another solution; see (2.63)–(2.64) below.
- (iv) Note that if the (NR) condition is strengthened to

$$\forall (i_1, \dots, i_k) \in \mathcal{I}^k, n \in \{m + 1, \dots, N\}, \text{ it holds that}$$

$$\left( \langle G_k(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_k}), \mathbf{e}_n^* \rangle \neq 0 \right) \implies \left( \sum_{j=1}^k \operatorname{Re}(\beta_{i_j}) - \operatorname{Re}(\beta_n) > 0 \right), \tag{2.49}$$

then the expression of  $h_k$  given by (2.47)–(2.48) results directly from the expression of Lyapunov–Perron integral  $\mathfrak{J}$ . Indeed,

$$\begin{aligned} \mathcal{J}(\xi) &= \int_{-\infty}^0 e^{-sA_s} \Pi_s G_k \left( \sum_{i=1}^m e^{\beta_i s} \xi_i \mathbf{e}_i \right) ds \\ &= \int_{-\infty}^0 \sum_{j=m+1}^N e^{-s\beta_j} \left\langle G_k \left( \sum_{i=1}^m e^{\beta_i s} \xi_i \mathbf{e}_i \right), \mathbf{e}_n \right\rangle \mathbf{e}_n ds \end{aligned} \tag{2.50}$$

i.e.

$$\mathcal{J}(\xi) = \sum_{j=m+1}^N \sum_{(i_1, \dots, i_k) \in \mathcal{I}^k} \left\langle G_k(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_k}), \mathbf{e}_n^* \right\rangle \xi_{i_1} \cdots \xi_{i_k} \mathbf{e}_n \int_{-\infty}^0 e^{(\beta_{i_1} + \dots + \beta_{i_k} - \beta_j)s} ds, \tag{2.51}$$

recalling that  $G_k(u)$  denotes  $G_k(u, \dots, u)$ , a homogeneous polynomial of order  $k$ . The condition (2.49) ensures that the integrals in (2.51) are well-defined, leading to (2.47)–(2.48) after integration.

Of course, by assuming only (NR) instead of (2.49), the Lyapunov–Perron integral may not be well defined anymore. But as shown below, the solution to Eq. (2.27) still exists, and is given again by (2.47)–(2.48).

(v) Finally, it is worth mentioning that cross non-resonance conditions of the form

$$\sum_{j=1}^k \beta_{i_j} - \beta_n \neq 0, \quad \forall (i_1, \dots, i_k) \in \mathcal{I}^k, \quad n \in \{m + 1, \dots, N\},$$

is also encountered for the study of normal forms on an invariant manifolds; see, e.g. [84, Sect. 3.2.1], [60, Thm. 2.4] and also [11, Thm. 3.1].

**Proof of Theorem 2** The proof is inspired by Lie algebra techniques used in the derivation of normal forms for ODEs (see, e.g., [5, Chap. 5] and [11, Chap. 1]). We proceed in three steps.

**Step 1** We seek a solution to Eq. (2.27) as a mapping  $h_k : E_c \rightarrow E_s$  that admits the following expansion:

$$h_k(\xi) = \sum_{n=m+1}^N \left( \sum_{(i_1, \dots, i_k) \in \mathcal{I}^k} \Psi_{i_1, \dots, i_k}^n(\xi) \right) \mathbf{e}_n, \quad \xi = (\xi_1, \dots, \xi_m) \in E_c. \tag{2.52}$$

Here, for each  $(i_1, \dots, i_k) \in \mathcal{I}^k$ , the function  $\Psi_{i_1, \dots, i_k}^n(\xi)$  is a complex-valued homogeneous polynomial of degree  $k$  given by

$$\Psi_{i_1, \dots, i_k}^n(\xi) = \Gamma_{i_1, \dots, i_k}^n \xi_{i_1} \cdots \xi_{i_k}. \tag{2.53}$$

The task is then to determine the coefficients  $\Gamma_{i_1, \dots, i_k}^n$  (in  $\mathbb{C}$ ) by using Eq. (2.27).

**Step 2** In that respect, we introduce the following homological operator  $\mathcal{L}_A$ :

$$\mathcal{L}_A[\phi](\xi) = D\phi(\xi)A_c\xi - A_s\phi(\xi), \quad \xi \in E_c, \tag{2.54}$$

where  $\phi : E_c \rightarrow E_s$  is a smooth function.

A key observation consists of noting that the  $E_s$ -valued function,  $\xi \mapsto \Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n$ , provides an eigenfunction of  $\mathcal{L}_A$  corresponding to the eigenvalue  $\sum_{j=1}^k \beta_{i_j} - \beta_n$ , in other words that the following identity holds

$$\mathcal{L}_A[\Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n](\xi) = \left[ \sum_{j=1}^k \beta_{i_j} - \beta_n \right] \Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n. \tag{2.55}$$

In order to check (2.55), we first calculate  $D\phi(\xi)A_c\xi$  when  $\phi(\xi) = \Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n$ . In that respect, denoting by  $e_j^n$  the  $j^{\text{th}}$  component of  $\mathbf{e}_n$ , the Jacobian matrix  $D[\Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n]$ , given by the following  $N \times m$  matrix,

$$D[\Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n] = \begin{pmatrix} \frac{\partial \Psi_{i_1, \dots, i_k}^n(\xi)}{\partial \xi_1} e_1^n & \dots & \dots & \frac{\partial \Psi_{i_1, \dots, i_k}^n(\xi)}{\partial \xi_m} e_1^n \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \Psi_{i_1, \dots, i_k}^n(\xi)}{\partial \xi_1} e_N^n & \dots & \dots & \frac{\partial \Psi_{i_1, \dots, i_k}^n(\xi)}{\partial \xi_m} e_N^n \end{pmatrix}, \tag{2.56}$$

possesses the following representation

$$\begin{aligned} D[\Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n] &= \mathbf{e}_n \left( \frac{\partial \Psi_{i_1, \dots, i_k}^n(\xi)}{\partial \xi_1}, \dots, \frac{\partial \Psi_{i_1, \dots, i_k}^n(\xi)}{\partial \xi_m} \right) \\ &= \Gamma_{i_1, \dots, i_k}^n \mathbf{e}_n \mathbf{B}(\xi). \end{aligned} \tag{2.57}$$

where  $\mathbf{B}(\xi) = (B_1(\xi), \dots, B_m(\xi))$  is an  $m$ -dimensional row vector whose components are given for any  $j$  in  $\{1, \dots, m\}$  by

$$B_j(\xi) = \frac{\partial}{\partial \xi_j} (\xi_{i_1} \dots \xi_{i_k}) = \begin{cases} p \xi_j^{p-1} \prod_{i_\ell \neq j} \xi_{i_\ell}, & \text{if } j \in \{i_1, \dots, i_k\}, \\ 0, & \text{otherwise,} \end{cases} \tag{2.58}$$

where  $p$  denotes the number of indices in the set  $\{i_1, \dots, i_k\}$  that equal  $j$ .

Thus,

$$D[\Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n]A_c\xi = \Gamma_{i_1, \dots, i_k}^n \mathbf{e}_n \mathbf{B}(\xi)A_c\xi. \tag{2.59}$$

which leads to

$$D[\Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n]A_c\xi = \Gamma_{i_1, \dots, i_k}^n \mathbf{e}_n \mathbf{B}(\xi) (\beta_1 \xi_1, \dots, \beta_m \xi_m)^{\text{tr}}, \tag{2.60}$$

since  $A$  is assumed to be diagonal.

By noting that the product  $\mathbf{B}(\xi) (\beta_1 \xi_1, \dots, \beta_m \xi_m)^{\text{tr}}$  is nothing else that  $\sum_{j=1}^k \beta_j \xi_{i_1} \dots \xi_{i_k}$ , and recalling the expression of  $\Psi_{i_1, \dots, i_k}^n(\xi)$  in (2.53), we infer from (2.60) that

$$D[\Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n]A_c\xi = \sum_{j=1}^k \beta_{i_j} \Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n. \tag{2.61}$$

On the other hand,

$$A_S \Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n = \beta_n \Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n, \tag{2.62}$$

and recalling the definition of  $\mathcal{L}_A$  in (2.54), the identity (2.55) follows.

**Step 3** By using the expansion of  $h_k(\xi)$  given by (2.52) in Eq. (2.27), and by using the fact that  $\Psi_{i_1, \dots, i_k}^n(\xi)\mathbf{e}_n$  are eigenvectors of the homological operator  $\mathcal{L}_A$  with eigenvalue  $\sum_{j=1}^k \beta_{i_j} - \beta_n$  (cf. (2.55)), we get

$$\sum_{n=m+1}^N \left[ \sum_{(i_1, \dots, i_k) \in \mathcal{I}^k} \left( \sum_{j=1}^k \beta_{i_j} - \beta_n \right) \Psi_{i_1, \dots, i_k}^n(\xi) \right] \mathbf{e}_n = \Pi_S G_k(\xi).$$

Recalling from (2.53) that  $\Psi_{i_1, \dots, i_k}^n = \Gamma_{i_1, \dots, i_k}^n \xi_{i_1} \dots \xi_{i_k}$ , we obtain

$$\sum_{n=m+1}^N \left[ \sum_{i_1, \dots, i_k \in \mathcal{I}^k} \left( \sum_{j=1}^k \beta_{i_j} - \beta_n \right) \Gamma_{i_1, \dots, i_k}^n \xi_{i_1} \dots \xi_{i_k} \right] \mathbf{e}_n = \Pi_S G_k(\xi). \tag{2.63}$$

At the same time, since  $G_k$  is a homogeneous polynomial of order  $k$  and  $\xi = \sum_{i=1}^m \xi_i e_i$ , we obtain

$$\begin{aligned} \Pi_{\mathfrak{s}} G_k(\xi) &= \sum_{n=m+1}^N \langle G_k(\xi), e_n^* \rangle e_n \\ &= \sum_{n=m+1}^N \sum_{(i_1, \dots, i_k) \in \mathcal{I}^k} \xi_{i_1} \cdots \xi_{i_k} \langle G_k(e_{i_1}, \dots, e_{i_k}), e_n^* \rangle e_n. \end{aligned} \tag{2.64}$$

By using the above identity in (2.63), we obtain the following formulas for the coefficients  $\Gamma_{i_1, \dots, i_k}^n$  in (2.53):

$$\Gamma_{i_1, \dots, i_k}^n = \frac{\langle G_k(e_{i_1}, \dots, e_{i_k}), e_n^* \rangle}{\sum_{j=1}^k \beta_{i_j} - \beta_n}. \tag{2.65}$$

The formula of  $h_k$  given in (2.47)–(2.48) is thus derived by combining (2.52), (2.53) and (2.65). The proof is complete.  $\square$

### 2.3 Analytic Formulas for Higher-Order Approximations

We discuss briefly here simple considerations to derive higher-order approximations of an invariant manifold. The approach relies on the use of a power series expansion of the manifold function  $h$  in the invariance equation (2.26). However, instead of keeping all the monomials at a given degree arising from this expansion, we filter out terms that carries significantly less energy compared with those that are kept. This elimination procedure relies on the assumption that the projected ODE dynamics onto the resolved subspace  $E_c$  contains most of the energy; an assumption which is often met in practical applications concerned with invariant manifold reduction. To present the idea in a simple setting, we consider below the case for which  $G(y) = G_2(y, y) + G_3(y, y, y)$  and a cubic approximation is sought.

When  $G = G_2 + G_3$ , the leading-order approximation of  $h$  is  $h_2$  given by (2.47)–(2.48) with  $k = 2$ . Recall also  $h_2$  satisfies (2.27). To determine the approximation of order 3, we replace  $h$  in the invariance equation (2.26) by  $h^{\text{app}} = h_2 + \psi$ , where  $\psi$  represents the homogeneous cubic terms in the power expansion of  $h$ , to be determined. By identifying all the terms of order two, we recover (2.27) with  $k = 2$  to be satisfied for  $h_2$ , and by identifying all the terms of order three, we obtain the following equation for  $\psi$ :

$$\begin{aligned} D\psi(\xi)A_c\xi - A_{\mathfrak{s}}\psi(\xi) &= -Dh_2(\xi)\Pi_c G_2(\xi) + \Pi_{\mathfrak{s}}G_2(\xi, h_2(\xi)) \\ &\quad + \Pi_{\mathfrak{s}}G_2(h_2(\xi), \xi) + \Pi_{\mathfrak{s}}G_3(\xi). \end{aligned} \tag{2.66}$$

Notice that the LHS of (2.66) is  $\mathcal{L}_A\psi$ , and that the RHS is a homogeneous cubic polynomial in the  $\xi$ -variable. If most of the energy of the ODE dynamics is contained in the low modes, one gets that the energy carried by  $y_{\mathfrak{s}}$  is much smaller than  $\|y_c\|^2$ . It is then reasonable to expect that the energy carried by  $h_2(\xi)$  is much smaller than  $\|\xi\|^2$  for  $\xi = y_c(t)$  as  $t$  varies. This energy consideration implies that on the RHS of (2.66), the term  $\Pi_{\mathfrak{s}}G_3(\xi)$  dominates the other three terms provided that  $\|G_2(y, y)\|/\|y\|^2$  is on the same order of magnitude as  $\|G_3(y, y, y)\|/\|y\|^3$ . Thus, it is reasonable to seek for a good approximation of  $\psi$  by simply solving the equation:

$$Dh_3(\xi)A_c\xi - A_{\mathfrak{s}}h_3(\xi) = \Pi_{\mathfrak{s}}G_3(\xi). \tag{2.67}$$

Note that this is exactly (2.27) with  $k = 3$ . In virtue of Theorem 2, the existence of  $h_3$  is guaranteed under the non-resonance condition (NR), and  $h_3$  is given by (2.47)–(2.48). We denote this cubic parameterization by

$$\begin{aligned} \Phi(\xi) &= h_2(\xi) + h_3(\xi) \\ &= \sum_{n=m+1}^N \left( \sum_{(i_1, i_2) \in \mathcal{I}^2} \frac{\langle G_2(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}), \mathbf{e}_n^* \rangle}{\beta_{i_1} + \beta_{i_2} - \beta_n} \xi_{i_1} \xi_{i_2} + \sum_{(i_1, i_2, i_3) \in \mathcal{I}^3} \frac{\langle G_3(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \mathbf{e}_{i_3}), \mathbf{e}_n^* \rangle}{\beta_{i_1} + \beta_{i_2} + \beta_{i_3} - \beta_n} \xi_{i_1} \xi_{i_2} \xi_{i_3} \right) \mathbf{e}_n. \end{aligned} \tag{2.68}$$

with  $\mathcal{I} = (1, \dots, m)$ . See the Supplementary Material for an application to the derivation of effective reduced models able to capture a subcritical Hopf bifurcation arising in an ENSO model.

In what precedes, we considered the case  $G$  of order 3, and determined approximations of order 3. We could nevertheless, seek for higher-order approximations of invariant manifolds, independently of the nonlinearity to be of high-order or not. For instance if  $G(y) = B(y, y)$ , i.e. quadratic, we outline hereafter how recursive solutions to a hierarchy of homological equations arise naturally once we look for higher-order approximations.

In that respect, we introduce some notations. We denote by  $\text{Poly}_k(E_c; E_s)$  (resp.  $\text{Poly}_k(E_c; E_c)$ ) the space of vectors in  $E_s$  (resp.  $E_c$ ) whose components are homogeneous polynomials of order  $k$  in the  $E_c$ -variable. Given a polynomial  $\mathcal{P}$  in  $\text{Poly}_k(E_c; E_s)$  or in  $\text{Poly}_k(E_c; E_c)$ , the symbol  $[\mathcal{P}(\xi)]_k$  represents the collection of terms of order  $k$  in  $\mathcal{P}$ .

By seeking a solution,  $\Psi$ , to the invariance equation Eq. (2.26) under the form,

$$\Psi(\xi) = \sum_{k \geq 2} \Psi_k(\xi), \quad \Psi_k \in \text{Poly}_k(E_c; E_s). \tag{2.69}$$

we infer that the  $\Psi_k$ 's satisfy the following recursive homological equations given by

$$\mathcal{L}[\Psi_k](\xi) = \left[ \Pi_s B(\Phi_{<k}(\xi), \Phi_{<k}(\xi)) \right]_k - \sum_{\ell=2}^{k-1} D\Psi_{k-\ell+1}(\xi) \left[ \Pi_c B(\Phi_{<\ell}(\xi), \Phi_{<\ell}(\xi)) \right]_\ell \tag{2.70}$$

where  $\Phi_{<\ell}(\xi)$  denotes

$$\Phi_{<\ell}(\xi) = \xi + \sum_{j=2}^{\ell-1} \Psi_j(\xi). \tag{2.71}$$

Note that with the convention  $\sum_2^1 \equiv 0$ , we recover the first homological equation, namely

$$\mathcal{L}[\Psi_2](\xi) = \Pi_s B(\xi, \xi). \tag{2.72}$$

In other words  $\Psi_2 = h_2$ . We refer to [85] for a detailed account regarding the rigorous and computational aspects for the determination of solutions to Eq. (2.70). [109, Chap. 11] contains also a detailed survey of algorithms to compute numerically invariant manifolds for fast-slow systems.

## Part II: Variational Approach to Closure

### 3 Optimal Parameterizing Manifolds

#### 3.1 Variational Formulation

##### 3.1.1 Parameterizing Manifolds (PM) and Parameterization Defect

A cornerstone of our approach presented below is the notion of *parameterizing manifold* (PM) that we recall below from [26,31,32]. Our framework takes place in finite dimension as in Part I, however here we consider more general systems of the form

$$\frac{dy}{dt} = Ay + G(y) + F, \quad y \in \mathbb{C}^N, \tag{3.1}$$

where  $F$  denotes a time-independent forcing in  $\mathbb{C}^N$ ,  $A$  is a  $N \times N$  matrix with complex entries, while  $G$  is assumed to be a smooth nonlinearity for which we do not assume  $G(0) = 0$  anymore. In practice Eq. (3.1) can be thought as derived in the perturbed variable from an original system, for which  $A$  is either the Jacobian matrix at a mean state ( $F \neq 0$ ) or at a steady state ( $F = 0$ ), although the concepts presented below do not restrict to such situations. Hereafter we assume that  $A$ ,  $F$  and  $G$  are such that classical solutions (at least  $C^1$ ) exist and that the corresponding initial value problem possesses a unique solution, at least for initial data taken in an open domain  $\mathcal{D}$  of  $\mathbb{C}^N$ . Dynamically-based formulas to design PMs for Eq. (3.1) are given in Sects. 4.3 and 4.4 below. For the moment we recall the definition of a PM, and introduce the notion of parameterization defect that will be used for the optimization of PMs.<sup>5</sup>

**Definition 1** Let  $T > 0$  and  $0 \leq t_1 < t_2 \leq \infty$ . Let  $y$  be a solution to Eq. (3.1), and  $\Psi : E_c \rightarrow E_s$  be a continuous mapping satisfying the following energy inequality for all  $t$  in  $[t_1, t_2)$

$$\int_t^{t+T} \|y_s(s) - \Psi(y_c(s))\|^2 ds < \int_t^{t+T} \|y_s(s)\|^2 ds, \tag{3.2}$$

where  $y_c(s) = \Pi_c y(s)$  and  $y_s(s) = \Pi_s y(s)$ , with  $\Pi_c$  and  $\Pi_s$  that denote the canonical projectors onto  $E_c$  and  $E_s$ , respectively ( $E_c$  and  $E_s$  being defined in (2.14)).

Then, the manifold,  $\mathfrak{M}_\Psi$ , defined as the graph of  $\Psi$ , i.e.

$$\mathfrak{M}_\Psi = \{\xi + \Psi(\xi) \mid \xi \in E_c\}, \tag{3.3}$$

is a finite-horizon parameterizing manifold associated with the system of ODEs (3.1), over the time interval  $[t_1, t_2)$ . The time-parameter  $T$  measuring the length of the “finite-horizon” is independent on  $t_1$  and  $t_2$ . If (3.2) holds for  $t_2 = \infty$ , then  $\mathfrak{M}_\Psi$  is simply called a finite-horizon parameterizing manifold, and if it holds furthermore for all  $T$ , it is called a parameterizing manifold (PM).

Given a parameterization  $\Psi$  of the unresolved variables (in  $E_s$ ) in terms of the resolved ones (in  $E_c$ ), a natural non-dimensional number, the *parameterization defect*, is defined as

$$Q_T(t, \Psi) = \frac{\int_t^{t+T} \|y_s(s) - \Psi(y_c(s))\|^2 ds}{\int_t^{t+T} \|y_s(s)\|^2 ds}, \quad t \in [t_1, t_2). \tag{3.4}$$

<sup>5</sup> Note however that other cost functionals may be considered at this stage; see Sect. 4.4 below.

Sometimes, the dependence on  $t$  will be secondary, and by making  $t = t_1$  in (3.4) with  $t_1$  sufficiently large so that for instance transient dynamics has been removed, we will denote  $Q_T(t, \Psi)$  simply by  $Q_T(\Psi)$ . In any event, either  $Q_T(t, \Psi)$  or  $Q_T(\Psi)$  allows us to compare objectively two manifolds in their ability to parameterize the variables that lie in the subspace  $E_s$  by those that lie in the subspace  $E_c$ . Clearly a situation corresponding to an exact slaving of the variables in  $E_s$  by those in  $E_c$  as encountered in the invariant manifold theory revisited in Part I, corresponds to  $Q_T(\Psi) \equiv 0$  for any solution  $y$  that lies on the invariant manifold,  $\mathfrak{M}_\Psi$ , associated with the parameterization  $\Psi$ . If furthermore  $\mathfrak{M}_\Psi$  attracts e.g. exponentially any trajectory like in the case of an inertial manifold, then  $Q_T(\Psi) \rightarrow 0$ , as  $T \rightarrow \infty$  whatever the solution  $y$ .

A standard  $m$ -dimensional Galerkin approximation based on the modes in  $E_c$  (with  $\dim(E_c) = m$ ), corresponds to  $\Psi = 0$  and thus to  $Q_T(\Psi) \equiv 1$ . Thus,

$$\mathfrak{M}_\Psi \text{ is a PM if and only if } Q_T(\Psi) < 1 \text{ for all } T > 0.$$

Clearly, given a parameterization  $\Psi$ , it may happen that the corresponding parameterization defect  $Q_T(\Psi)$  fluctuates from solutions to solutions, and depends also substantially on the time interval  $[t_1, t_2)$  over which the initial time  $t$  is taken to compute the integrals in (3.4), as well as the horizon  $T$ .

Nevertheless, given a set of solutions of interest, a horizon  $T$ , an interval  $[t_1, t_2)$ , and a set dimension of the reduced state space (i.e.  $\dim(E_c) = m$ ), one is naturally inclined for seeking for parameterizations,  $\Psi$ , that come with the smallest parameterization defect. In other words, we aim at solving the following minimization problem

$$\min_{\Psi \in \mathcal{E}} \int_t^{t+T} \|y_s(s) - \Psi(y_c(s))\|^2 ds, \tag{3.5}$$

for which  $\mathcal{E}$  denotes a space of parameterizations that makes not only tractable the determination of a minimizer, but also that is not too greedy in terms of data. This latter requirement comes from important practical considerations. For instance, for high-dimensional systems (e.g.  $N$  of about few hundred thousands), one has typically  $y(t)$  available over a relatively small interval of time, and thus if e.g.  $m \sim N/100$  and the choice of  $\mathcal{E}$  is too naive, such as homogeneous polynomials in the  $E_c$ -variable, with arbitrary coefficients, one might easily face an overfitting problem in which too many coefficients have to be determined while not enough snapshots of  $y(s)$  are available over  $[t, t + T]$ . Section 4 below shows that the backward–forward system (2.29) provides a space  $\mathcal{E}$  of dynamically-based parameterizations that allow to bypass this difficulty as the coefficients to be determined are dependent only on a scalar parameter, the backward integration time  $\tau$  in (2.29).

These practical considerations are central in our approach but before providing their details, we consider in the next section other important theoretical questions. These questions deal with the existence (and uniqueness) of minimizers to (3.5) on one hand, and with the characterization of the closure system that is reached once (3.5) is solved, on the other. Thus, we show in Sect. 3.2 below that, under assumptions of ergodicity, reasonable for a broad class of forced-dissipative nonlinear systems such as arising in fluid dynamics, the minimization problem (3.5) possesses a unique solution, as  $T \rightarrow \infty$ ; see Theorem 4 and also [32, Theorem A.1 and Remark 4.1]. We call the corresponding minimizer, the *optimal parameterizing manifold*. We conclude finally by showing that an optimal PM, once used as a substitute of the unresolved variables, leads to a reduced system in  $E_c$  that gives the conditional expectation of the original system, i.e. the best vector field of the reduced state space resulting from averaging of the unresolved variables with respect to a probability measure conditioned on the resolved variables; see Theorem 5 below.



We emphasize that PMs have already demonstrated their utility in other applications. For instance, PMs have shown their usefulness for the effective determination of surrogate low-dimensional systems in view of the optimal control of dissipative nonlinear PDEs. In this case, rigorous error estimates show that parameterization defects arise naturally in the efficient model reduction of optimal control problems (see [26, Thm. 1 and Cor.2]) as furthermore supported by detailed numerical results (see [26, Sec. 5.5] and [22]). Speaking roughly, these estimates show that the smaller is the parameterization defect, the better a low-dimensional controller designed from the surrogate system, behaves. Error estimates that relate the parameterization defect to the ability of reproducing the original dynamics' long term statistics by a surrogate system are difficult to produce for uncontrolled deterministic systems, in particular for chaotic regimes such as considered hereafter in Sects. 5 and 6, due to the singular nature (with respect to the Lebesgue measure) of the underlying invariant measure. In the stochastic realm, this invariant measure becomes smooth for a broad class of systems and the tools of stochastic analysis make the obtention of such estimates more amenable albeit non trivial; see [21]. Nevertheless, considerations from ergodic theory and conditional expectations are already insightful for the deterministic systems dealt with in this article as explained in Sect. 3.2 below.

### 3.1.2 Parameterization Correlation and Angle

Given a parameterization  $\Psi$  that is not trivial (i.e.  $\Psi \neq 0$ ), we define the *parameterization correlation* as,

$$c(t) = \frac{\text{Re}(\Psi(y_c(t)), y_s(t))}{\|\Psi(y_c(t))\| \|y_s(t)\|}. \tag{3.6}$$

It provides a measure of collinearity between the parameterized variable  $\Psi(y_c(t))$  and the unresolved variable  $y_s(t)$ , as time evolves. In case of exact slaving,  $y_s(t) = \Psi(y_c(t))$  and thus  $c(t) \equiv 1$ .

The parameterization correlation,  $c(t)$ , is another key quantity in our approach. Speaking roughly, we aim for not only at finding a PM with the smallest parameterization defect but also with a parameterization correlation,  $c(t)$ , to be as much close to one as possible. The basic idea is to find parameterizations that approximate as much as possible an ideal slaving situation, for regimes in which slaving does not hold necessarily.

In particular, the parameterization correlation allows us, once an optimal PM has been determined, to select the dimension  $m$  of the reduced phase space according to the following criterium:  $m$  should correspond to the lowest dimension of  $E_c$  for which the probability distribution function (PDF) of the corresponding *parameterization angle*,

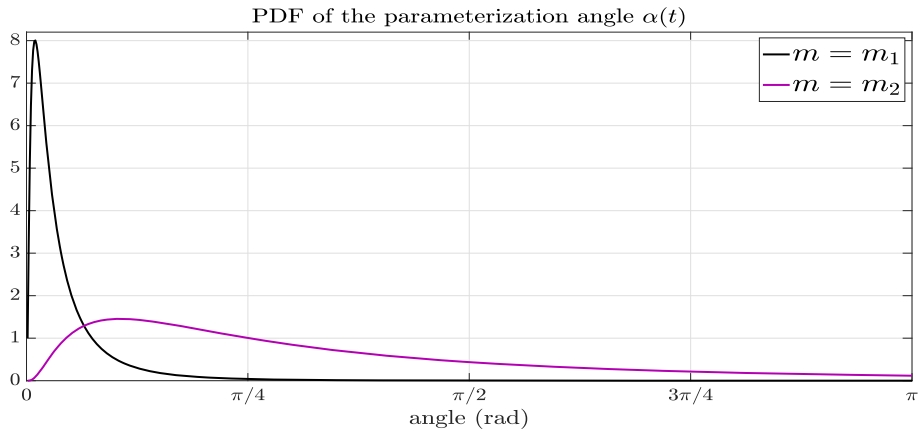
$$\alpha(t) = \arccos(c(t)), \tag{3.7}$$

is the most skewed towards zero and the mode of this PDF (i.e. the value that appears most often) is the closest to zero; see Fig. 2.

As a rule of thumb, we aim at finding PMs,  $\Psi$ , such that:

1. The parameterization defect,  $Q_T(\Psi)$ , is as small as possible, and
2. The PDF of the parameterization angle  $\alpha(t)$  is skewed towards zero as much as possible, and its mode (i.e. the value that appears most often) is close to zero.

We illustrate in Sects. 3.4 and 5 below that, when breakdown of slaving principle occurs, these rules manifest as a natural framework to diagnose and select a parameterization. Nevertheless as the dimension of the original problem gets large, one may have to inspect a modewise



**Fig. 2** Effect of the reduced dimension  $m$ : schematic. This effect is schematically shown here on the PDF of the parameterization angle  $\alpha(t)$ . Here a case corresponding to  $m_1 > m_2$ , is depicted:  $m_1$  is large enough to be a successful PM while  $m_2$  is not

version of  $Q_T$  (as discussed in Sect. 4.2) as well as of  $\alpha(t)$ ; see Sect. 6.3 for the latter. In any case, the idea is that one should not only parameterize properly the statistical effects of the neglected scales but also avoid to lose their phase relationships with the retained scales [132]. This is particularly important to derive closures that respect a certain phase coherence between the resolved and unresolved scales.

### 3.2 Optimal Parameterizing Manifold and Conditional Expectation

We present in this section the main results that serve as a foundational basis for the applications discussed hereafter. We denote by  $X$  the vector field associated with Eq. (3.1) i.e.

$$X(y) = Ay + G(y) + F, \quad \text{for all } y \in \mathbb{C}^N. \tag{3.8}$$

To simplify the presentation, we assume this vector field to be sufficiently smooth and dissipative on  $\mathbb{C}^N$ , such that the corresponding flow,  $T_t$ , is well-defined. We assume, furthermore, that  $T_t$  possesses an invariant probability measure  $\mu$ , which is *physically relevant* [37,57], in the sense that the following property holds for  $y$  in a positive Lebesgue measure set  $B(\mu)$  (of  $\mathbb{C}^N$ ) and for every continuous observable  $\varphi : \mathbb{C}^N \rightarrow \mathbb{C}$

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(T_t(y)) dt = \int \varphi(y) d\mu(y). \tag{3.9}$$

This property assures that meaningful averages can be calculated and the statistics of the dynamical system can be investigated by the asymptotic distribution of orbits starting from Lebesgue almost every initial condition in e.g. the basin of attraction,  $B(\mu)$ , of the statistical equilibrium  $\mu$ .

Recall that, like all probability measures invariant under  $T_t$ , an invariant measure that satisfies (3.9) is supported by the global attractor  $\mathcal{A}$  when the latter exists; e.g. [24, Lemma 5.1]. In the case a global attractor is not known to exist, an invariant measure has its support in the *non-wandering set*,  $\Lambda$ ; see [69, Remark 1.4, p. 197].

It can be proven for e.g. Anosov flows [13], partially hyperbolic systems [1], Lorenz-like flows [12], and observed experimentally for many others [28,33,57,71] that a common

feature of (dissipative) chaotic systems is the transformation (under the action of the flow) of the initial Lebesgue measure into a probability measure with finer and finer scales, reaching asymptotically an invariant measure  $\mu$  of Sinai–Ruelle–Bowen (SRB) type. This measure is singular with respect to the Lebesgue measure, is supported by the local unstable manifolds contained in  $\mathcal{A}$  or in  $\Lambda$  [37, Def. 6.14], and if it has no zero Lyapunov exponents it satisfies (3.9) [177]. This latter property is often referred to as the *chaotic hypothesis* that, roughly speaking, expresses an extension of the ergodic hypothesis to non-Hamiltonian systems [71]. We work thus hereafter within this hypothesis and we assume furthermore that (3.9) holds for  $\varphi$  that lies in the space of integrable function,  $L^1_\mu(\mathbb{C}^N)$ , with respect to the invariant measure  $\mu$ .

Having clarified the ergodic framework within which we will frame our variational approach, we consider now a high-mode parameterization of the form

$$\Psi(\xi) = \sum_{n=m+1}^N \Psi_n(\xi)e_n, \quad \xi \in E_c, \tag{3.10}$$

with the  $e_n$ 's denoting the eigenmodes of the linear part,  $A$ , that span the subspace  $E_s$ . The regularity assumption made on  $\Psi$  is clarified hereafter; see Theorem 3. In practice,  $\Psi$  does not need to cover the whole range  $[m + 1, N]$  and some  $\Psi_n$  may be zero.

We denote by  $m$  the push-forward of the measure  $\mu$  by the projector  $\Pi_c$  onto  $E_c$ , namely

$$m(B) = \mu(\Pi_c^{-1}(B)), \quad B \in \mathcal{B}(E_c), \tag{3.11}$$

where  $\mathcal{B}(E_c)$  denotes the family of Borel sets of  $E_c$ ; i.e. the family of sets that can be formed from open sets (for the topology on  $E_c$  induced by the norm  $\|\cdot\|_{E_c}$ ) through the operations of countable union, countable intersection, and relative complement.

In what follows (see Sect. 4), given a solution  $y(t)$  that emanates from  $y_0$  in  $B(\mu)$ , we also consider the parameterization defect,  $\mathcal{Q}_n$ , associated with the parameterization  $\Psi_n$  of the  $n^{\text{th}}$ -eigenmode, namely

$$\mathcal{Q}_n(T) = \frac{1}{T} \int_0^T \left| \langle y_s(t), e_n^* \rangle - \Psi_n(y_c(t)) \right|^2 dt, \tag{3.12}$$

where we recall that  $\{e_j^*\}_{j=1}^N$  denotes the eigenvectors of the conjugate transpose  $A^*$ .

In the case  $\{e_n\}$  forms an orthonormal basis of  $\mathbb{C}^N$ , namely when  $A$  is a Hermitian matrix, we have due to the Parseval's identity,

$$\mathcal{Q}_T(\Psi) = \frac{1}{T} \int_0^T \|y_s(t) - \Psi(y_c(t))\|^2 dt = \sum_{n=m+1}^N \mathcal{Q}_n(T). \tag{3.13}$$

However this equality does not always hold, in general. Indeed, by writing  $y_s(t) = \sum_{n=m+1}^N y_n(t)e_n$  with  $y_n(t) = \langle y_s(t), e_n^* \rangle$ , we remark that

$$\|y_s(t) - \Psi(y_c(t))\|^2 = \sum_{n_1, n_2=m+1}^N \left\langle \left( y_{n_1}(t) - \Psi_{n_1}(y_c(t)) \right) e_{n_1}, \left( y_{n_2}(t) - \Psi_{n_2}(y_c(t)) \right) e_{n_2} \right\rangle,$$

and the latter identity is reduced to  $\sum_{n=m+1}^N |y_n(t) - \Psi_n(y_c(t))|^2$  when  $\langle e_j, e_k \rangle = \delta_{j,k}$  for all  $j, k = m + 1, \dots, N$ .

Thus, solving (3.5) is not always equivalent to solving the following family of variational problems

$$\min_{\Psi_n \in \mathcal{E}} \int_0^T \left| \langle y_s(t), e_n^* \rangle - \Psi_n(y_c(t)) \right|^2 dt, \quad m + 1 \leq n \leq N. \tag{3.14}$$

As we will see, for practical reasons we will often prefer to solve (3.14) rather than (3.5); see Sect. 4.2 below. Nevertheless, the existence and uniqueness of minimizers for either (3.14) or (3.5), are dealt with in the same way. Hereafter, we present the latter only in the case of (3.5) (allowing for the simplification of certain statements) and leave to the reader the corresponding statements and proofs in the case of the minimization problems (3.14).

In that respect, we select the space of parameterizations,  $\mathcal{E}$ , to be the Hilbert space constituted by  $E_s$ -valued functions of the resolved variables  $\xi$  in  $E_c$ , that are square-integrable with respect to  $m$ , namely

$$\mathcal{E} = L^2_m(E_c; E_s) = \left\{ \Psi : E_c \rightarrow E_s \text{ measurable and such that } \int_{E_c} \|\Psi(\xi)\|^2 dm(\xi) < \infty \right\}. \tag{3.15}$$

Our approach to minimize,  $\mathcal{Q}_T(\Psi)$  (in  $\mathcal{E}$ ), and to identify parameterizations for which the normalized parameterization defect

$$\mathcal{Q}_T(\Psi) = \mathcal{Q}_T(\Psi) \langle \|y_s\|^2 \rangle_T^{-1}, \tag{3.16}$$

satisfies

$$0 < \lim_{T \rightarrow \infty} \mathcal{Q}_T(\Psi) < 1, \tag{3.17}$$

relies substantially on the general disintegration theorem of probability measures; see e.g. [51, p. 78]. In (3.16), we have denoted by  $\langle \|y_s\|^2 \rangle_T$  the time-mean of  $y_s$  over  $[0, T]$ . The disintegration theorem states that given a probability measure  $\mu$  on  $\mathbb{C}^N$ , a vector subspace  $V$  of  $\mathbb{C}^N$ , and a Borel-measurable mapping  $p : \mathbb{C}^N \rightarrow V$ , then there exists a uniquely determined family of probability measures  $\{\mu_x\}_{x \in V}$  such that, for  $m$ -almost all  $x$  in  $V$ ,  $\mu_x$  is concentrated on the pre-image  $p^{-1}(\{x\})$  of  $x$ , i.e.  $\mu_x(\mathbb{C}^N \setminus p^{-1}(\{x\})) = 0$ , and such that for every Borel-measurable function  $\phi : \mathbb{C}^N \rightarrow \mathbb{C}$ ,

$$\int \phi(y) d\mu(y) = \int_V \left( \int_{y \in p^{-1}(\{x\})} \phi(y) d\mu_x(y) \right) dm(x). \tag{3.18}$$

Here  $m$  denotes the *push-forward* in  $V$  of the measure  $\mu$  by the mapping  $p$ , i.e.  $m$  is given by (3.11) where  $\Pi_c$  is replaced by  $p$ . Note that when  $p$  is the projection onto  $V$ , the probability measure  $\mu_x$  is the conditional probability of the unresolved variables, contingent upon the value of the resolved variable to be  $x$ ; see also [29, Supporting Information].

Hereafter, we apply this theorem with the reduced phase space,  $V$ , to be the subspace of the resolved variables,  $E_c$ , and the mapping  $p$  to be the projector  $\Pi_c$  onto  $E_c$ . In this case, a decomposition analogous to (3.18) holds for the measure  $\mu$  itself, namely

$$\mu(B \times F) = \int_F \mu_\xi(F) dm(\xi), \quad B \times F \in \mathcal{B}(E_c) \otimes \mathcal{B}(E_s). \tag{3.19}$$

First, we state a result identifying natural conditions under which,  $\lim_{T \rightarrow \infty} \mathcal{Q}_T(\Psi)$  exists.

---

<sup>6</sup> i.e. up to an exceptional set of null measure with respect to  $m$ .

**Theorem 3** Assume that Eq. (3.1) admits an invariant probability measure  $\mu$  satisfying (3.9) and that the unresolved variable  $\zeta$  in  $E_s$  has a finite energy in the sense that

$$\int \|\zeta\|^2 \, d\mu < \infty. \tag{3.20}$$

If  $\Psi$  lies in  $L^2_m(E_c, E_s)$ , then for a.e. solution  $y(t)$  of Eq. (3.1) that emanates from an initial datum  $y_0$  in the basin of attraction  $B(\mu)$ , the limit  $\lim_{T \rightarrow \infty} Q_T(\Psi)$  exists, and is given by

$$\lim_{T \rightarrow \infty} Q_T(\Psi) = \int_{(\xi, \zeta) \in E_c \times E_s} \|\zeta - \Psi(\xi)\|^2 \, d\mu. \tag{3.21}$$

**Proof** This theorem is a direct consequence of the ergodic property (3.9) applied to the observable

$$\varphi(\xi, \zeta) = \|\zeta - \Psi(\xi)\|^2. \tag{3.22}$$

Indeed, first, let us note that  $\varphi(\xi, \zeta) = \|\zeta\|^2 - 2\langle \zeta, \Psi(\xi) \rangle + \|\Psi(\xi)\|^2$  satisfies

$$\int \varphi(\xi, \zeta) \, d\mu \leq \int \|\zeta\|^2 \, d\mu_\xi(\zeta) + \int \|\Psi(\xi)\|^2 \, d\mu + \int (\|\zeta\|^2 + \|\Psi(\xi)\|^2) \, d\mu, \tag{3.23}$$

by application of (3.19) and the Fubini’s theorem for the two first integrals in the RHS of (3.23), and of the Cauchy–Schwarz and Young inequalities for the third integral. Another application of (3.19) and the Fubini’s theorem for this latter integral shows that  $\varphi$  lies in  $L^1_\mu(\mathbb{C}^N)$ , since  $\Psi$  belongs to  $L^2_m(E_c, E_s)$  and (3.20) holds.  $\square$

We are now in position to show the existence of a unique minimizer to the minimization problem

$$\min_{\Psi \in \mathcal{E}} \left( \lim_{T \rightarrow \infty} Q_T(\Psi) \right), \tag{3.24}$$

i.e. to ensure the existence of an optimal manifold minimizing the parameterization defect. The minimizer is also characterized; see (3.26) below. An earlier version of such results may be found in [32, Theorem A.1] for the special case of a truncated Primitive Equation model due to Lorenz [123]. The general case is dealt with below.

**Theorem 4** Assume that the assumptions of Theorem 3 hold. Then the minimization problem

$$\min_{\Psi \in \mathcal{E}} \int_{(\xi, \zeta) \in E_c \times E_s} \|\zeta - \Psi(\xi)\|^2 \, d\mu, \tag{3.25}$$

possesses a unique solution in  $\mathcal{E} = L^2_m(E_c, E_s)$  whose argmin is given by

$$\Psi^*(\xi) = \int_{E_s} \zeta \, d\mu_\xi(\zeta), \quad \xi \in E_c. \tag{3.26}$$

Furthermore

$$\lim_{T \rightarrow \infty} Q_T(\Psi^*) \leq \lim_{T \rightarrow \infty} Q_T(\Psi), \quad \forall \Psi \in L^2_m(E_c, E_s). \tag{3.27}$$

**Proof** The proof is a direct consequence of the disintegration theorem applied to the ergodic measure  $\mu$ . Let us introduce the following Hilbert space of  $E_s$ -valued functions

$$L^2_\mu(E_c \times E_s; E_s) = \left\{ f : E_c \times E_s \rightarrow E_s, \text{ measurable and s.t.} \right. \\ \left. \int_{E_c \times E_s} \|f(\xi, \zeta)\|^2 \, d\mu(\xi, \zeta) < \infty \right\}. \tag{3.28}$$

Let us define the expectation  $\mathbb{E}_\mu(g)$  with respect to the invariant measure  $\mu$  by

$$\mathbb{E}_\mu(g) = \int_{E_c \times E_s} g(\xi, \zeta) \, d\mu(\xi, \zeta), \quad g \in L^2_\mu(E_c \times E_s; E_s). \tag{3.29}$$

By applying to the ambient Hilbert space  $L^2_\mu(E_c \times E_s; E_s)$ , the standard projection theorem onto closed convex sets [14, Theorem 5.2], one defines (given  $\Pi_c$ ) the conditional expectation  $\mathbb{E}_\mu[g|\Pi_c]$  of  $g$  as the unique function in  $\mathcal{E}$  that satisfies the inequality

$$\mathbb{E}_\mu[\|g - \mathbb{E}_\mu[g|\Pi_c]\|^2] \leq \mathbb{E}_\mu[\|g - \Psi\|^2], \text{ for all } \Psi \in \mathcal{E}. \tag{3.30}$$

The general disintegration theorem of probability measures, applied to  $\mu$  (see (3.18)), provides the following explicit representation of the conditional expectation

$$\mathbb{E}_\mu[g|\Pi_c] = \int_{E_s} g(\xi, \zeta) \, d\mu_\xi(\zeta), \tag{3.31}$$

with  $\mu_\xi$  denoting the disintegrated measure of  $\mu$  in (3.19).

Now let us take  $g(\xi, \zeta) = \zeta$ , then

$$\mathbb{E}_\mu[\zeta|\Pi_c] = \Psi^*, \tag{3.32}$$

with  $\Psi^*$  defined by (3.26). We have then

$$\|\Psi^*(\xi)\|^2 \leq \int \|\zeta\|^2 \, d\mu_\xi(\zeta), \tag{3.33}$$

and by using (3.18) we have

$$\int \|\Psi^*(\xi)\|^2 \, dm(\xi) \leq \int \|\zeta\|^2 \, d\mu. \tag{3.34}$$

This inequality shows that  $\Psi^*$  lies in  $L^2_m(E_c, E_s)$  due to assumption (3.20).

We have then from (3.30),

$$\mathbb{E}_\mu[\|\zeta - \Psi^*\|^2] \leq \mathbb{E}_\mu[\|\zeta - \Psi\|^2], \text{ for all } \Psi \in \mathcal{E}. \tag{3.35}$$

By recalling that

$$\mathbb{E}_\mu[\|\zeta - \Psi^*\|^2] = \int_{E_s \times E_s} \|\zeta - \Psi^*(\xi)\|^2 \, d\mu(\xi, \zeta) = \int \|\zeta - \Psi^*(\xi)\|^2 \, d\mu(\xi, \zeta), \tag{3.36}$$

one obtains then, by applying respectively (3.9) to  $\varphi = \|\zeta - \Psi^*\|^2$  and  $\varphi = \|\zeta - \Psi\|^2$ , that for all  $\Psi$  in  $\mathcal{E}$ ,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \|y_s(t) - \Psi^*(y_c(t))\|^2 \, dt \leq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \|y_s(t) - \Psi(y_c(t))\|^2 \, dt. \tag{3.37}$$

The proof is complete. □

The manifold obtained as the graph of  $\Psi^*$  given by (3.26) will be called the *optimal PM*. Formula (3.26) shows that the optimal PM corresponds actually to the manifold that maps to each resolved variable  $\xi$  in  $E_c$ , the averaged value of the unresolved variable  $\zeta$  in  $E_s$  as distributed according to the conditional probability measure  $\mu_\xi$ . In other words, the optimal PM provides the best manifold (in a least-square sense) that averages out the fluctuations of the unresolved variable.

By making  $\Psi \equiv 0$  in (3.27), this optimal PM comes with a (normalized) parameterization defect (3.16) that satisfies necessarily

$$0 \leq \lim_{T \rightarrow \infty} Q_T(\Psi^*) \leq 1. \quad (3.38)$$

This variational view on the parameterization problem of the unresolved variables removes any sort of ambiguity that has surrounded the notion of (approximate) inertial manifold in the past. Indeed, within this paradigm shift, given an ergodic invariant measure  $\mu$  and a reduced dimension  $m$  (defining thus a projector  $\Pi_c$ ), the optimal PM may have a parameterization defect very close to 1 and thus the best possible nonlinear parameterization one could ever imagine cannot a priori do much better than a classical Galerkin approximation, and sometimes even worse. To the opposite, the smaller  $Q_T(\Psi^*)$  is (for  $T$  large), the best the parameterization. All sort of nuances are actually admissible, even when the parameterization defect is just below unity; see [32] and Sect. 3.4 below.

We emphasize that although the theory presented in this section has been shaped for asymptotic values of  $T$ , in practice we will be instead interested to seek for optimal PMs learned over a training length as short as possible (to rely on as few as possible DNS snapshots). In that respect, it is where the parametric families of dynamically-based parameterizations derived in Sect. 4 below (and relying on Part I) become useful. We will indeed show that by applying these formulas in practice, we are able to derive optimal PMs trained over short training intervals of length comparable to a characteristic recurrence or decorrelation time of the dynamics; see Sects. 5 and 6 below.

**Remark 2** (i) The ergodic property (3.9) can be relaxed into weaker forms such as considered in e.g. [24,69]. These relaxed versions hold for a broad class of dissipative systems including systems of ODEs and even PDEs, as long as a global attractor exists [24, Theorem 2.2]. However these weaker forms do not guarantee the existence of the limit in (3.21) and the latter would be replaced instead by a notion of generalized limit involving e.g. averaging over accumulations points. The statistical equilibrium  $\mu$  is then not guaranteed to be unique.

Nevertheless, bearing these changes in mind, the proof presented above can be easily adapted and the conclusion of Theorem 4 remains valid with however a form of optimality that is now subject to the choice of the statistical equilibrium. Within this ergodic framework, several optimal parameterizing manifolds may co-exist but for each statistical equilibrium there is only one optimal parameterizing manifold. The same is true if a global attractor  $\mathcal{A}$  is not guaranteed to exist:  $\mathcal{A}$  must be replaced by the non-wandering set  $\Lambda$ , and the optimal PM is unique for trajectories sampled according to the statistical equilibrium  $\mu$ .

- (ii) With the nuances brought up in (i) above, Theorem 4 applies thus to any relevant Galerkin truncations of systems of PDEs arising in fluid dynamics; see [32] and Sect. 3.4 below for an application to a 9D Galerkin truncation of the Primitive Equations of the atmosphere due to Lorenz [123].
- (iii) Theorem 4 is fundamental for understanding and interpretation but is of little interest for computing the optimal PM in practice, except in specific problems for which  $\mu$  is known explicitly (see e.g. [23, Sec. 4]) or can be approximated semi-analytically [128,129]; see also [171] for an alternative approach to estimate numerically  $\mu_\xi$  in the context of slow-fast systems. In Sect. 4 below we introduce instead explicit dynamically-based parameterizations that, once optimized according to a mode-adaptive approach, provide an efficient way to determine PMs that although suboptimal (for (3.25)) will be shown to be skillful for closure in practice; see Sects. 5 and 6 below.

We have then the following result relating the conditional expectation to the optimal PM. We state this theorem in the case of quadratic interactions, motivated by applications in fluid dynamics; see also [32, Sec. 4.3] and Sect. 3.4 below, for an illustration.

**Theorem 5** *Under the conditions of Theorem 4 if  $G$  is a quadratic nonlinearity  $B$  in Eq. (3.1), the conditional expectation,  $\mathbb{E}_\mu[X|\Pi_c]$ , satisfies*

$$\mathbb{E}_\mu[X|\Pi_c](\xi) = A_c \xi + \Pi_c B(\xi, \xi) + \Pi_c (B(\xi, \Psi^*(\xi)) + B(\Psi^*(\xi), \xi)) + F_c + \eta(\xi), \quad \xi \in E_c, \tag{3.39}$$

where  $X$  is the vector field given by (3.8),  $\Psi^*$  is the optimal PM guaranteed by Theorem 4, and  $\eta$  is given by

$$\eta(\xi) = \int_{\zeta \in E_s} \Pi_c B(\zeta, \zeta) \, d\mu_\xi(\zeta). \tag{3.40}$$

Thus in the case  $\eta = 0$ , the optimal PM,  $\Psi^*$ , provides the conditional expectation  $\mathbb{E}_\mu[X|\Pi_c]$ , i.e.

$$\mathbb{E}_\mu[X|\Pi_c](\xi) = A_c \xi + \Pi_c B(\xi, \xi) + \Pi_c (B(\xi, \Psi^*(\xi)) + B(\Psi^*(\xi), \xi)) + F_c. \tag{3.41}$$

**Proof** Expanding  $X(\xi + \zeta)$  (with  $(\xi, \zeta)$  in  $E_c \times E_s$ ) and integrating with respect to the disintegrated probability measure,  $\mu_\xi$ , we get (by using that  $\int d\mu_\xi = 1$ )

$$\begin{aligned} \mathbb{E}_\mu[X|\Pi_c](\xi) &= A_c \xi + \Pi_c B(\xi, \xi) + F_c + \eta(\xi) + \int \left( \Pi_c (B(\xi, \zeta) + B(\zeta, \xi)) \right) d\mu_\xi(\zeta), \\ &= A_c \xi + \Pi_c B(\xi, \xi) + F_c + \eta(\xi) + \Pi_c B \left( \xi, \int \zeta \, d\mu_\xi(\zeta) \right) \\ &\quad + \Pi_c B \left( \int \zeta \, d\mu_\xi(\zeta), \xi \right), \end{aligned} \tag{3.42}$$

which given the expression of  $\Psi^*$  in (3.26), gives (3.39). □

### 3.3 Inertial Manifolds and Optimal PMs

To avoid any confusion, we clarify the distinction between the concept of an inertial manifold (IM) and that of an optimal parameterizing manifold (PM). First of all, an IM is a particular case of an asymptotic PM since when an inertial manifold  $\Psi$  exists,  $Q_T(\Psi) = 0$  for all  $T$  sufficiently large. We list below some important points to better appreciate the differences between the two concepts.

- (i) When an IM,  $\Psi$ , exists, then  $\Psi = \Psi^*$  in (3.26) with  $\mu_\xi$  being the Dirac mass (in  $E_s$ ) concentrated on  $\Psi(\xi)$ , i.e.  $\mu_\xi = \delta_{\Psi(\xi)}$ . Furthermore in this case, the probability distribution  $p_\alpha$  of the parameterization angle,  $\alpha(t)$  given by (3.7), is given by the Dirac mass  $\delta_0$  (on the real line) concentrated at 0.
- (ii) Working with the eigenbasis of the linear part of Eq. (3.1) and assuming that an IM exists, let  $m_*$  denote the minimal dimension of the reduced state space required for an IM to exist. If  $m = \dim(E_c) < m_*$  then there is no inertial manifold but a PM still exists in general as supported by Theorem 3. One may wonder however whether more can be said when  $m < m_*$ .

This is where the parameterization defect,  $Q_T$ , and the parameterization angle,  $\alpha(t)$ , provide useful mutual informations. Typically when  $m < m_*$ , seeking for a manifold that minimizes  $Q_T$  allows for parameterizing optimally (in a least square sense) the



statistical effects of the neglected scales in terms of those retained. However one should keep in mind to avoid losing the phase relationships between the resolved and unresolved scales, and in that sense the distribution  $p_\alpha$  should not be too spread. For systems with a high-dimensional global attractor one may need to inspect a modewise version of  $Q_T$  (as discussed in Sect. 4.2 below) as well as of  $\alpha(t)$  for the design of the nonlinear parameterization; see Sect. 6.3 for the latter in the context of 1D Kuramoto-Sivashinsky turbulence.

Thus, even for systems that admit an IM, an optimal PM often provides an efficient closure based on much fewer modes compared to an inertial form. Such an observation about efficient reduced dimension is known by the practitioner familiar with the notion of approximate inertial manifold (AIM). An AIM provides a manifold such that the attractor lies within a neighborhood of it that shrinks as the reduced dimension  $m$  is increased [48,52,131]. Nevertheless, as the reduced dimension is set too low, a given AIM may suffer from e.g. an over-parameterization of the small scales resulting into dramatic errors backscattering to the large scales; see Sect. 6. This is because the AIM approach does not address the question of finding an optimal manifold that minimizes the parameterization defect while keeping the reduced dimension as low as possible. This is the focus of the PM approach proposed in this article which is thus, in essence, variational rather than concerned with the rate of convergence with  $m$  as in standard AIM theory.

### 3.4 A Reduced-Order Primitive Equation Example: PM and Breakdown of Slaving Principles

The conditional expectation is related to the optimal PM according to Theorem 5, making thus the optimal PM an essential ingredient for the closure problem. Depending on the problem at hand, the conditional expectation provides e.g. the reduced equations that filter out the fast gravity waves from truncated Primitive Equations (PE) of the atmosphere; see [32]. Truncations corresponding to  $\eta = 0$  in (3.39), i.e. when the high-high interactions do not contribute to the low-mode dynamics, is particularly favorable for the conditional expectation to provide such a filtering property. As shown numerically in [32], the conditional expectation provides indeed such a “low-pass filter” closure for the truncated PE proposed by Lorenz in 1980 [123], when a critical Rossby number,  $\epsilon^*$ , is crossed. We reproduce hereafter some of these numerical results and provide new, complementary understanding based on the theory of PMs such as discussed in this article.

The model of [123], when rescaled following [32], becomes

$$\begin{aligned}
 \epsilon^2 a_i \frac{dX_i}{dt} &= \epsilon^3 a_i b_i X_j X_k - \epsilon^2 c(a_i - a_k) X_j Y_k + \epsilon^2 c(a_i - a_j) Y_j X_k \\
 &\quad - 2\epsilon c^2 Y_j Y_k - \epsilon^2 N_0 a_i^2 X_i + a_i(Y_i - Z_i), \\
 a_i \frac{dY_i}{dt} &= -\epsilon a_k b_k X_j Y_k - \epsilon a_j b_j Y_j X_k + c(a_k - a_j) Y_j Y_k - a_i X_i - N_0 a_i^2 Y_i, \\
 \frac{dZ_i}{dt} &= -\epsilon b_k X_j (Z_k - H_k) - \epsilon b_j (Z_j - H_j) X_k + c Y_j (Z_k - H_k) \\
 &\quad - c(Z_j - H_j) Y_k + g_0 a_i X_i - K_0 a_i Z_i + \mathcal{F}_i.
 \end{aligned}
 \tag{3.43}$$

The above equations are written for each cyclic permutation of the set of indices (1, 2, 3), namely, for

$$(i, j, k) \in \{(1, 2, 3), (2, 3, 1), (3, 1, 2)\}.
 \tag{3.44}$$

We refer to [32] for a detailed description of this model and its parameters. For our purpose, it is sufficient to know that the time,  $t$ , is an  $\mathcal{O}(1)$ -slow time, and that  $X_i$ 's,  $Y_i$ 's, and  $Z_i$ 's are  $\mathcal{O}(1)$ -amplitudes for the divergent velocity potential, streamfunction, and dynamic height, respectively. In this setting  $N_0$  and  $K_0$  are rescaled damping coefficients in the slow time. The  $\mathcal{F}_i$ 's are  $\mathcal{O}(1)$  control parameters that, in combination with variations of  $\epsilon$ , can be used to affect regime transitions/bifurcations. In a general way,  $\epsilon$ , can be identified with the Rossby number.

Solutions of higher-order accuracy in  $\epsilon > 0$  that are entirely slow in their evolution are, by definition, balanced solutions, and [73] showed by construction several examples of explicitly specified, approximate balanced models. One of these, the Balance Equations (BE), was conspicuously more accurate than the others when judged in comparison with apparently slow solutions of (3.43). The BE approximation consists of a parameterization of the  $X_i$ 's and  $Z_i$ 's variables, in terms of the  $Y_i$ 's variables. The  $\mathbf{Z}$ -component of this parameterization has an explicit expression. The  $\mathbf{X}$ -component of this parameterization, denoted by  $\Phi$ , is however obtained implicitly, by solving a system of differential-algebraic equations derived from Eq. (3.43) under a balance assumption that consists of replacing the dynamical equation for the  $X_i$ 's by algebraic relations. Eventually, we arrive at a 3D reduced system of ODEs, simply called the BE, and that takes the form

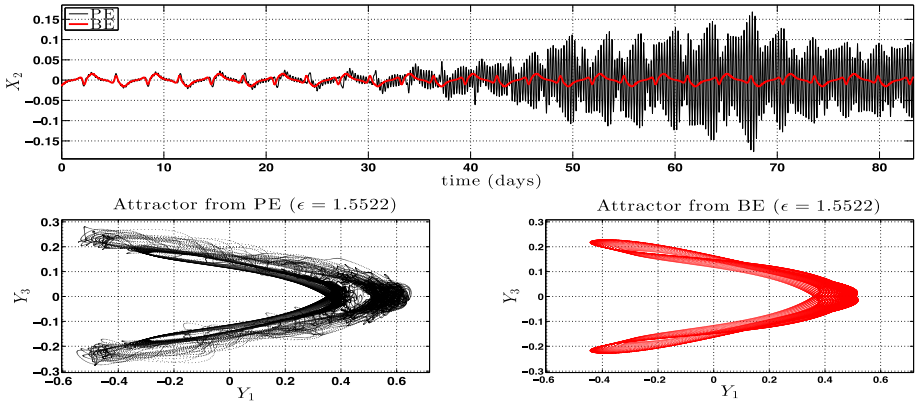
$$a_i \frac{dY_i}{dt} = -\epsilon a_k b_k \Phi_j(\mathbf{Y}) Y_k - \epsilon a_j b_j Y_j \Phi_k(\mathbf{Y}) + c(a_k - a_j) Y_j Y_k - a_i \Phi_i(\mathbf{Y}) - N_0 a_i^2 Y_i, \quad (3.45)$$

with  $(i, j, k)$  as in (3.44). We refer to [32, Sec. 3.1] for a derivation.

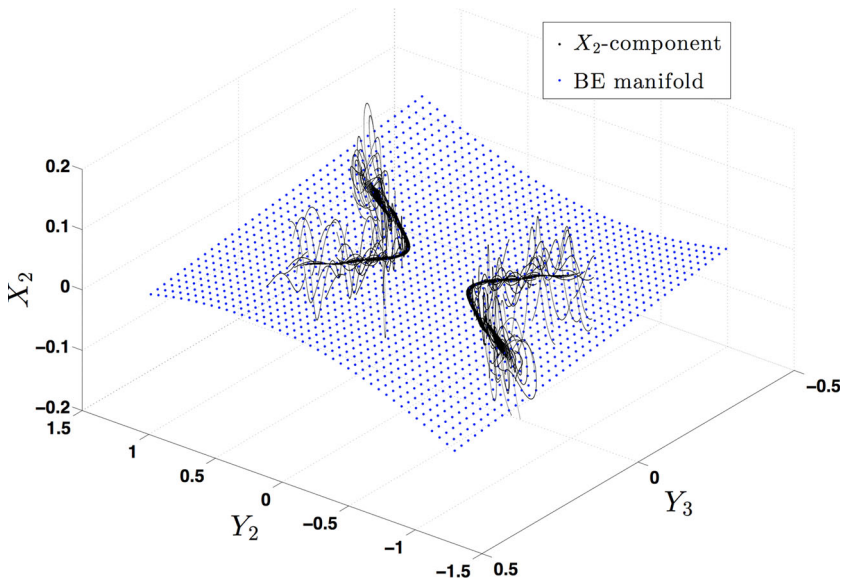
For certain Rossby numbers for which energetic bursts of fast oscillations occur in the course of time (occurring for  $\epsilon > \epsilon^*$ ), Chekroun et al. [32] have shown that the underlying BE manifold (associated with the BE parameterization of the  $\mathbf{X}$ - and  $\mathbf{Z}$ -variables), provides a very good approximation of the optimal PM for this problem, and thus of the conditional expectation in virtue of Theorem 5, i.e. the best approximation in the  $\mathbf{Y}$ -variable for which the “fast”  $\mathbf{X}$ - and  $\mathbf{Z}$ -variables are averaged out. In other words, the BE (3.45) provides a nearly optimal reduced vector field that averages out the fast oscillations contained in the  $\mathbf{Y}$ -variable. Figure 3, reproduced from [32], illustrates this feature for the model (3.43). The lower-right panel shows that the BE reduced model is able to capture the coarse-grained topological features of the projected attractor onto the “slow” variables,  $Y_1$  and  $Y_3$ , when compared with the projection onto the same variables of the attractor associated with the full Eq. (3.43). For the rest of this section we will use the BE as if it were the optimal PM. All the results presented hereafter correspond to  $\epsilon = 1.5522 > \epsilon^*$ ; see [32].

The underlying BE manifold is a 6D manifold obtained as graph of a 6D-valued mapping of a 3D-variable ( $\mathbf{Y}$ ), and as such only slices can be represented in 3D. Such a slice is shown in Fig. 4. More exactly, it shows the  $X_2$ -variable as parameterized by the slow  $Y_2$ - and  $Y_3$ -variables. Note that in order to obtain this representation, the  $Y_1$ -variable, involved also in the BE parameterization  $\Phi$  along with the  $Y_2$ - and  $Y_3$ -variables, has been set to its most probable value conferring to Fig. 4 a certain “typicalness.” This being kept in mind, the slice thus obtained of the BE manifold (and shown in Fig. 4) will be simply called the BE manifold, for simplifying the discourse.

As evidenced in Fig. 4, a PE solution on the attractor—as observed through the  $X_2$ -variable—possesses an intricate transversal component to the BE manifold that seems to exclude its parameterization by a smooth manifold, whereas, at the same time, a substantial portion of the trajectory lies very close to the BE manifold. It is this latter portion of the dynamics that is well captured by the BE manifold and that allows for approximating the aforementioned conditional expectation. Here Fig. 4 reveals thus simple geometric features

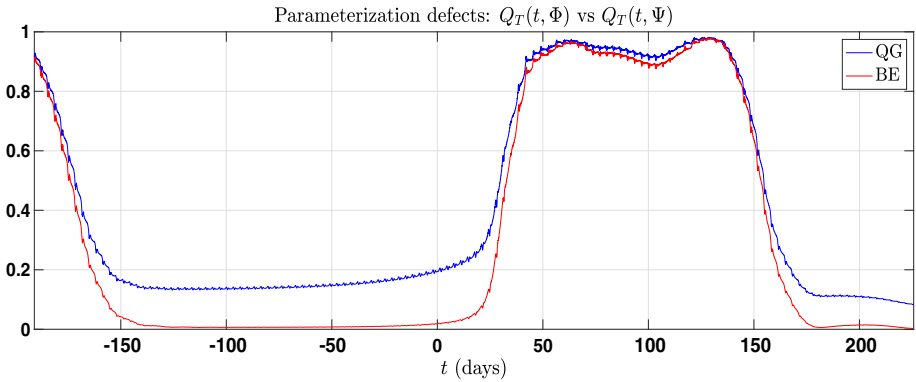


**Fig. 3** Attractor comparison between PE and BE (reproduced from [32], with permission from Elsevier). A slow-variable projection of the global attractor associated with Eq. (3.43) (lower-left panel) and its approximation obtained from the BE reduced model (lower-right panel). Even in presence of energetic bursts of fast oscillations in the fast variables (here such an episode is shown in the upper panel for the  $X_2$ -variable (black curve)), the BE model (3.45) is able to capture the coarse-grained topological features of the projected attractor onto the slow variables. This is because the BE manifold provides a good approximation of the optimal PM given in (3.26) that averages here out (optimally) the fast oscillations



**Fig. 4** The BE manifold for the  $X_2$ -variable. Note that in order to obtain this representation, the  $Y_1$ -variable, involved also in the BE parameterization  $\Phi$  along with the  $Y_2$ - and  $Y_3$ -variables, has been set to its most probable value. The black curve shows the resulting  $X_2$ -variable obtained after solving Eq. (3.43) while the blue dots correspond to the BE parameterization  $\Phi$  involved in (3.45)

(not identified in [32]), which are responsible for the BE to provide in the space of slow variables, a vector field that approximates the PE dynamics. It does so by filtering out the (fast) oscillations contained in the PE solutions; the fast dynamics corresponding, in this representation, to the transversal part of the dynamics. Indeed, a closer inspection reveals

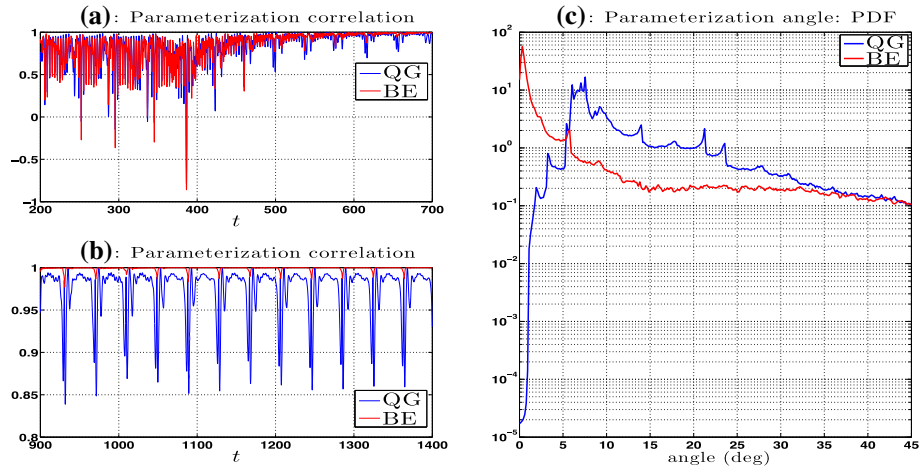


**Fig. 5** Parameterization defects of the BE manifold  $\Phi$  and the QG manifold  $\Psi$ . Here the parameterization defects as given by (3.4),  $Q_T(t, \Phi)$  (red curve) and  $Q_T(t, \Psi)$  (blue curve), are computed for the BE manifold,  $\Phi$ , and for the QG manifold  $\Psi$  [32, Eq. (4.22)]; each with  $T = 80$  (for the rescaled system (3.43)) which corresponds to 10 days in the time-variable of the original Lorenz model [123] (Color figure online)

that this transversal part of the dynamics corresponds exactly to the aforementioned burst of fast oscillations. This is confirmed by computing the parameterization defect. In that respect, Fig. 5 shows the parameterization defect  $t \mapsto Q_T(t, \Phi)$  (given by (3.4)) of the BE manifold  $\Phi$  for a time horizon set to  $T = 80$  (for the rescaled system (3.43)) which corresponds to 10 days in the time-variable of the original Lorenz model [32]. Figure 5 shows that  $Q_T(t, \Phi)$  oscillates, as  $t$  evolves, between values right above zero and right below one (red curve). The rising of values taken by  $Q_T(t, \Phi)$  occurs over time windows for which the parameterized  $X$ -variable contains a significant fraction of the total energy, such as manifested by bursts of fast oscillations in the  $X_2$ -variable shown in the upper panel of Fig. 3 between 40 and 80 days. To the contrary, when the PE solutions get very close to the BE manifold, the dynamics is almost slaved to this manifold and  $Q_T(t, \Phi) \approx 0$ .

Complementarily, the parameterization defect  $Q_T(t, \Psi)$  has been computed for the standard Quasigeostrophic (QG) manifold [32, Eq. (4.22)] that can be derived for  $\epsilon = 0$  and is associated with the famous quadratic Lorenz system [122]; see [32, Sec. 4.2]. Here again a similar behavior is observed for  $Q_T(t, \Psi)$  (blue curve in Fig. 5) with the noticeable difference that  $Q_T(t, \Psi)$  stays further away from zero than  $Q_T(t, \Phi)$  does, as  $t$  evolves.

The parameterization correlation,  $c(t)$  given by (3.6), has been also computed for the BE and the QG manifolds. The results are shown in Panels (a) and (b) of Fig. 6, over different time intervals. Although when an episode of fast (gravity waves) oscillations occurs in the PE solutions, the parameterization correlation can deviate substantially from 1 for the BE and QG manifolds (panel (a)), the parameterization correlation gets, comparatively, much closer to 1 for the BE than for the QG manifold over time intervals for which the slow, Rossby waves dominate the dynamics (panel (b)). This phenomenon is confirmed statistically at the level of the probability distribution for the corresponding parameterization angle,  $\alpha(t) = \arccos(c(t))$ . The PDF of the latter is much more skewed towards zero for the BE manifold than for the QG manifold supporting thus, at a quantitative level, the visual rendering of Fig. 4 which suggests that a substantial portion of the PE trajectory lies very close to the BE manifold. More precisely, Fig. 6c shows that the mode of the PDF of  $\alpha(t)$  (i.e. the value that appears most often) for the BE manifold is located very close to zero,



**Fig. 6** Parameterization correlation and angle. The parameterization correlation,  $c(t)$  given by (3.6), is shown for the BE manifold ( $\Psi = \Phi$ , red curve) and the QG manifold (“ $\Psi = \text{QG}$  manifold,” blue curve), over two consecutive time windows for panels **a** and **b**; the range of fluctuations over the 2nd window (panel **b**) being smaller to the range shown in the 1st window (panel **a**). The time-episode shown in panel **a** corresponds to the presence of energetic bursts of fast oscillations in the solutions ( $Q_T \approx 1$  for the BE), whereas panel **b** corresponds to a time-episode devoid of such oscillations ( $Q_T \approx 0$  for the BE). The PDFs of the corresponding parameterization angle  $\alpha(t)$  given by (3.7), estimated after long integration of Eq. (3.43), are shown in panel **c** (Color figure online)

whereas  $\alpha(t)$  almost never reaches such a level of proximity to zero for the QG manifold. In that sense, the BE manifold is a manifold that is close to be locally invariant in the sense of (i) of Sect. 2.1, that is a slaving relationship like (2.16) almost takes place over time, while being brutally violated from time to time (transversal part of the PE dynamics to the manifold; see Fig. 4).

Thus the BE manifold provides an example of a manifold that is close to be locally invariant and that provides a (nearly optimal) PM. However, nothing excludes the existence of dynamics that although getting very close to a given manifold over certain time windows (almost slaving situation), experiences excursions far away from it so often that in average the parameterization defect gets greater than one, making this manifold to be a non-parameterizing one. Situations for which the dynamics lies in the vicinity of a given manifold (without large excursions) is also a favorable context for this manifold to be a PM; see Sect. 5.3 below for such an example.

Noteworthy are also the tails of the PDFs of the parameterization angle  $\alpha(t)$  for both, the BE and QG manifolds, which do not drop off suddenly as  $\alpha$  increases: this is symptomatic of the fact that the PE solutions get frequently far away from these manifolds as time evolves. As a comparison, we refer to Sect. 5.3 below for an example of parameterization angle  $\alpha$  whose PDF drops suddenly as  $\alpha$  increases.

Although enlightening, this example of (excellent) approximation of the optimal PM (and thus of the conditional expectation) that the BE manifold provides, exploits specific aspects of the problem at hand, encapsulated in the very derivation of the BE manifold. The question of efficient dynamically-based formulas for the approximation of an optimal PM in a general context, thus remains. The next section addresses this issue.

### 4 Parameterizing Manifolds and Mode-Adaptive Minimization: Dynamically-Based Formulas

In this section we derive dynamically-based formulas for designing parameterizing manifolds in practice. The formulas derived in Sect. 4.3 below take their origin in the pullback representation (2.33) (in Theorem 1) and the associated backward–forward system (2.29) that arise in the approximation theory of invariant manifolds revisited in Part I. The parametric class of leading interaction approximation (LIA) of the high modes obtained this way is completed by another parametric class built from the quasi-stationary approximation (QSA) in Sect. 4.4; close to the first criticality, the QSA is an approximation to the LIA, but differs as one moves away from criticality. We also make precise hereafter the corresponding minimization problems to solve in order to optimize our parameterizations in practice, within a mode-adaptive optimization procedure (Sect. 4.2).

#### 4.1 Backward–Forward Method: General Considerations

We first show that the parameterization  $h_\tau^{(1)}$  given in (2.30), as obtained by finite-time integration of the backward–forward system (2.29), satisfies an equation analogous to Eq. (2.27) satisfied by  $h_k$ .

**Lemma 1** *The manifold function  $h_\tau^{(1)}$  defined by (2.30) satisfies the following system of first order quasilinear PDEs:*

$$\mathcal{L}_A[h](\xi) = \Pi_s G_k(\xi) - e^{\tau A_s} \Pi_s G_k(e^{-\tau A_c} \xi). \tag{4.1}$$

with  $\mathcal{L}_A[h](\xi) = Dh(\xi)A_c\xi - A_s h(\xi)$  and  $A_c, A_s$  defined in (2.21).

**Proof** In (2.30), by replacing  $\xi$  with  $e^{tA_c}\xi$ , we get

$$\begin{aligned} \Phi(t) &= h_\tau^{(1)}(e^{tA_c}\xi) = \int_{-\tau}^0 e^{-sA_s} \Pi_s G_k(e^{sA_c} e^{tA_c}\xi) ds \\ &= \int_{-\tau}^0 e^{-sA_s} \Pi_s G_k(e^{(s+t)A_c}\xi) ds \\ &= \int_{t-\tau}^t e^{-(s'-t)A_s} \Pi_s G_k(e^{s'A_c}\xi) ds'. \end{aligned} \tag{4.2}$$

We obtain then

$$\begin{aligned} \frac{d\Phi(t)}{dt} &= \Pi_s G_k(e^{tA_c}\xi) - e^{\tau A_s} \Pi_s G_k(e^{(t-\tau)A_c}\xi) \\ &\quad + A_s \int_{t-\tau}^t e^{-(s'-t)A_s} \Pi_s G_k(e^{s'A_c}\xi) ds' \\ &= \Pi_s G_k(e^{tA_c}\xi) - e^{\tau A_s} \Pi_s G_k(e^{(t-\tau)A_c}\xi) + A_s \Phi(t). \end{aligned} \tag{4.3}$$

On the other hand, we also have

$$\frac{d\Phi(t)}{dt} = [Dh_\tau^{(1)}(e^{tA_c}\xi)]A_c e^{tA_c}\xi. \tag{4.4}$$

Equation (4.1) follows by equating the RHSs of (4.3) and (4.4) and by taking the limit  $t \rightarrow 0$ . □

This lemma provides the equation satisfied by the parameterization  $h_\tau^{(1)}$  given by (2.30). However this parameterization is built from the backward–forward system (2.29) associated with Eq. (2.2) that does not include forcing terms, unlike for more general systems of ODEs such as Eq. (3.1) dealt with in Sect. 3.

To extend the parameterization  $h_\tau^{(1)}$  to systems that include forcing terms, we thus naturally seek for solution of the backward–forward system associated with Eq. (3.1), namely

$$\frac{dy_c^{(1)}}{ds} = A_c y_c^{(1)} + \Pi_c F, \quad s \in [-\tau, 0], \tag{4.5a}$$

$$\frac{dy_s^{(1)}}{ds} = A_s y_s^{(1)} + \Pi_s G_k(y_c^{(1)}) + \Pi_s F, \quad s \in [-\tau, 0], \tag{4.5b}$$

$$\text{with } y_c^{(1)}(s)|_{s=0} = \xi, \text{ and } y_s^{(1)}(s)|_{s=-\tau} = 0. \tag{4.5c}$$

Here  $\Pi_s = \text{Id}_{\mathbb{C}^N} - \Pi_c$  with  $\Pi_c$  denoting the canonical projector onto the eigensubspace,  $E_c$ , spanned by the dominant eigenmodes of  $A$ .

By going through similar calculations than for the proof of Lemma 1, the high-mode solution of (4.5),  $y_s^{(1)}[\xi](0; -\tau)$ , denoted here by  $\Psi_\tau^{(1)}(\xi)$ , satisfies then

$$\begin{aligned} \mathcal{L}_A[\Psi_\tau^{(1)}](\xi) + D\Psi_\tau^{(1)}(\xi)\Pi_c F &= \Pi_s G_k(\xi) - e^{\tau A_s} \Pi_s G_k(S_F(-\tau)\xi) \\ &+ (\text{Id} - e^{\tau A_s})\Pi_s F, \end{aligned} \tag{4.6}$$

with

$$S_F(t)\xi = e^{tA_c} \xi - A_c^{-1}(\text{Id} - e^{tA_c})\Pi_c F. \tag{4.7}$$

Obviously  $\Psi_\tau^{(1)} = h_\tau^1$  when  $F \equiv 0$ .

In practice, in order to find an explicit expression of the parameterization  $\Psi_\tau^{(1)}$ , one prefers to solve (4.5) rather than solving Eq. (4.6) directly. Note that we could have adopted the same strategy for deriving the formulas of Theorem 2, i.e. by solving the backward–forward system (2.29) in this case.

The manifold  $\mathfrak{M}_\tau$  associated with  $\Psi_\tau^{(1)}$  possesses a natural geometric interpretation. Given a solution  $y(t)$  of Eq. (3.1) and denoting by  $U_\tau y_c(t)$  the lift of  $y_c(t)$  onto the manifold  $\mathfrak{M}_\tau$ , i.e.  $U_\tau y_c(t) = y_c(t) + \Psi_\tau^{(1)}(y_c(t))$ , we obtain

$$\overline{\text{dist}(y(t), \mathfrak{M}_\tau)^2} \leq \overline{\|y(t) - U_\tau y_c(t)\|^2} = \overline{\|y_s(t) - \Psi_\tau^{(1)}(y_c(t))\|^2}, \tag{4.8}$$

where the overbar denotes the time average over  $[0, T]$ . In other words,

$$\overline{\text{dist}(y(t), \mathfrak{M}_\tau)^2} \leq \mathcal{Q}_T(\Psi_\tau^{(1)}), \tag{4.9}$$

with  $\mathcal{Q}_T$  that denotes the parameterization defect

$$\mathcal{Q}_T(\Psi_\tau^{(1)}) = \frac{1}{T} \int_0^T \left\| y_s(t) - \Psi_\tau^{(1)}(y_c(t)) \right\|^2 dt. \tag{4.10}$$

Thus, we understand a practical advantage in restricting ourself to the  $\Psi_\tau^{(1)}$ -class of parameterizations instead of the more general  $\mathcal{E}$ -class considered in (3.15). Indeed, once an explicit expression for  $\Psi_\tau^{(1)}$  is derived, it allows us to greatly simplify the minimization problem involved in Theorem 4, by replacing it with the minimization in the scalar variable  $\tau$  of the cost functional  $\mathcal{Q}_T$  given by (4.10). Although the corresponding minimizer is a priori suboptimal compared to the more general minimization problem (3.25), we will see in applications that it provides in various instances an efficient parameterization.

Furthermore, based on (4.9), minimizing  $\mathcal{Q}_T(\Psi_\tau^{(1)})$  in the  $\tau$ -variable has the following useful interpretation: it forces, within the  $\Psi_\tau^{(1)}$ -parametrization class, the manifold  $\mathfrak{M}_\tau$  to get the closest to the trajectory  $y(t)$ , in a least-square sense. As mentioned earlier, an alternative approach, the AIM approach, has been proposed in the literature, but the latter is *asymptotic* in essence rather than the PM approach presented here which is *variational*. The AIM approach consists indeed of seeking for a family of manifolds,  $\mathcal{M}_m$ , for which  $\overline{\text{dist}(u(t), \mathcal{M}_m)}$  vanishes to zero as  $m = \dim(\mathcal{M}_m) \rightarrow \infty$ ; see e.g. [48,162,163,166]. In contradistinction, the PM approach consists for a given reduced dimension,  $m$ , of seeking for a manifold  $\mathfrak{M}$  within a certain parametric class of dynamically-based parameterizations, for which  $\overline{\text{dist}(u(t), \mathfrak{M})}$  is minimized.

Thus, given a reduced dimension,  $m$ , seeking for the best approximation within a parameterization class is at the core of the PM approach and, as shown in Sect. 3, is quintessential to address closure problems, in the sense that it relates naturally to the conditional expectation i.e. to the best closure that can be derived out of nonlinear parameterizations alone; see Theorem 5.

**Remark 3** Given the limitations on our ability to estimate the norms, it is in general hard to derive sharp estimates of  $\mathcal{Q}_T(\Psi_\tau^{(1)})$ . Nevertheless, some related estimates have been produced about  $\overline{\text{dist}(y(t), \mathcal{M})^2/\|y(t)\|^2}$ , for the 2D Navier–Stokes equations [20,68] when  $\mathcal{M}$  denotes the manifold associated with the quasi-stationary approximation; see (4.40) below.

### 4.2 Mode-Adaptive Optimization

Although the minimization in the scalar variable  $\tau$  of the cost functional  $\mathcal{Q}_T$  in (4.10) is more appealing than solving the general minimization problem (3.25), we may suffer from the fact that the parameter  $\tau$  to be optimized, is chosen globally, irrespectively e.g. to the content of energy of a particular high mode to parameterize. To better account for the distribution of energy across the modes, we propose instead to optimize parameterizations of the form

$$\Phi_\tau^{(1)}(\xi) = \sum_{n=m+1}^N \Phi_n(\tau_n, \beta, \xi)e_n, \quad \tau = (\tau_{m+1}, \dots, \tau_N), \tag{4.11}$$

in the multivalued  $\tau$ -variable. We emphasize that each parameterization  $\Phi_n$  depends only on  $\tau_n$  (and not the other  $\tau_p$ 's for  $p \neq n$ ), and thus each  $\Phi_n$  may be optimized independently from each other.

This way, we are left for each of the  $n^{\text{th}}$  mode, with a parameterization to optimize,  $\Phi_n(\tau_n, \beta, \xi)$ , that is a scalar function of the scalar variable  $\tau_n$ . Following Sect. 4.1 and assuming  $A$  diagonalizable (in  $\mathbb{C}^N$ ), we obtain  $\Phi_n(\tau_n, \beta, \xi)$ , for each  $m + 1 \leq n \leq N$ , as the high-mode part  $y_n^{(1)}$  of the solution (at  $s = 0$ ) to the backward–forward system

$$\frac{dy_c^{(1)}}{ds} = A_c y_c^{(1)} + \Pi_c F, \quad s \in [-\tau_n, 0], \tag{4.12a}$$

$$\frac{dy_n^{(1)}}{ds} = \beta_n y_n^{(1)} + \Pi_n G_k(y_c^{(1)}) + \Pi_n F, \quad s \in [-\tau_n, 0], \tag{4.12b}$$

$$\text{with } y_c^{(1)}(s)|_{s=0} = \xi, \text{ and } y_n^{(1)}(s)|_{s=-\tau_n} = 0, \tag{4.12c}$$

in which the RHS in Eq. (4.5b) has been replaced by  $\beta_n y_n^{(1)} + \Pi_n G_k(y_c^{(1)}) + \Pi_n F$ . Here  $\Pi_n X = \langle X, e_n^* \rangle$ , for any  $X$  in  $\mathbb{C}^N$ .



Explicit formulas of the  $\Phi_n(\tau_n, \beta, \xi)$ 's are given in Sect. 4.3 below when  $G_k$  is a quadratic nonlinearity. We show hereafter that minimizing for each  $n$  the parameterization defect naturally associated with  $\Phi_n$  leads to an optimal parameterization,  $\Phi_\tau^{(1)}$ , with a clear geometrical interpretation. To do so—given a fully resolved solution  $y(t)$  of the underlying  $N$ -dimensional ODE system (4.16) available over a training interval  $[0, T]$ —we consider for each  $n \geq m + 1$ , the parameterization defect

$$Q_n(\tau_n, T) = \frac{1}{T} \int_0^T |\Pi_n y(t) - \Phi_n(\tau, \beta, y_c(t))|^2 dt, \tag{4.13}$$

with  $y_c(t) = \Pi_c y(t)$ .

Denoting by  $\mathfrak{M}_\tau$  the manifold associated with the parameterization  $\Phi_\tau^{(1)}$  given by (4.11), we have

$$\begin{aligned} \overline{\text{dist}(y(t), \mathfrak{M}_\tau)^2} &\leq \left\| y(t) - \left( y_c(t) + \sum_{n \geq m+1} \Phi_n(\tau_n, \beta, y_c(t)) e_n \right) \right\|^2 \\ &\leq \left\| \sum_{n \geq m+1} (\Pi_n y(t) - \Phi_n(\tau_n, \beta, y_c(t))) e_n \right\|^2. \end{aligned} \tag{4.14}$$

Taking the eigenvectors of  $A$  to be normalized, we are thus left, thanks to the triangular inequality, with the following estimate

$$\overline{\text{dist}(y(t), \mathfrak{M}_\tau)^2} \leq \sum_{n \geq m+1} \left| \Pi_n y(t) - \Phi_n(\tau_n, \beta, y_c(t)) \right|^2 = \sum_{n \geq m+1} Q_n(\tau_n, T). \tag{4.15}$$

Thus minimizing each  $Q_n(\tau_n, T)$  (in the  $\tau_n$ -variable) is a natural idea to enforce closeness of  $y(t)$  in a least-square sense to the corresponding manifold  $\mathfrak{M}_\tau$ . Note that we could have chosen to minimize  $Q_T$  as given in (4.10) but with  $\Phi_\tau^{(1)}$  replacing  $\Psi_\tau^{(1)}$ . The resulting minimization would become however more challenging in high-dimension as it would require to minimize  $Q_T(\Phi_\tau^{(1)})$  in the multidimensional variable  $\tau$ . Except when the basis  $\{e_j\}_{j=1}^N$  is orthonormal (see (3.13)), the two approaches are not equivalent, i.e. minimizing  $Q_T(\Phi_\tau^{(1)})$  in the vector  $\tau$ , vs minimizing  $Q_n(\tau_n, T)$  in the scalar  $\tau_n$  for each  $n \geq m + 1$ . We opted for the latter as a simple algorithm can be proposed to minimize  $Q_n$  efficiently; see Appendix. Nevertheless, even in this scalar case, a certain care must be paid, as the mapping  $\tau \mapsto Q_n(\tau, T)$  is not guaranteed to be convex; see Sect. 5. Furthermore, depending on the dynamics (and the training interval  $[0, T]$ ) local minima may appear that require also a special care in order to properly design an efficient parameterization for the problem at hand; see Remark 8 below.

### 4.3 Parametric Leading-Interaction Approximation

In this section, we focus on the case of quadratic nonlinear interactions under constant forcing, for which we derive parameterization formulas by solving the backward–forward systems (4.12) (for  $G_k$  quadratic) presented in Sect. 4.2 above. Our approach allows for deriving parameterizations that take into account interactions between the forcing components and the nonlinear terms, at the leading order. As already pointed out in Sect. 4.2, these parameterizations are conditioned on the choice of a finite collection  $\tau$  of scalar parameters. For these reasons we will refer to  $\Phi_\tau^{(1)}$  given by (4.36) as the *parametric Leading-Interaction Approx-*

imation (LIA). As  $\tau$  varies, the corresponding class of parameterizations will be referred to as the  $\Phi^{(1)}$ -class or simply the LIA class.

The ODE system considered here is of the form:

$$\frac{dy}{dt} = Ay + B(y, y) + F, \quad y \in \mathbb{C}^N, \tag{4.16}$$

where  $A$  is an  $N \times N$  matrix with complex entries,  $B$  denotes quadratic nonlinear interactions with complex coefficients, and  $F$  is a constant forcing term in  $\mathbb{C}^N$ .

Given the spectral elements  $(\beta_j, e_j)$  of the matrix  $A$  that we assume diagonalizable (in  $\mathbb{C}^N$ ), we decompose the state space into resolved and unresolved subspaces as follows

$$\mathbb{C}^N = E_c \oplus E_s, \tag{4.17}$$

where

$$\begin{aligned} E_c &= \text{span}\{e_i : i = 1, \dots, m\}, \\ E_s &= \text{span}\{e_i : i = m + 1, \dots, N\}, \end{aligned} \tag{4.18}$$

see also (2.6)–(2.14).

We define the projection of a vector  $X$  in  $\mathbb{C}^N$  onto  $e_j$  as follows

$$\Pi_j X = \langle X, e_j^* \rangle, \tag{4.19}$$

with  $\{e_j^*\}$  denoting the eigenvectors of the conjugate transpose,  $A^*$ . The projectors  $\Pi_c$  is then explicitly given by

$$\Pi_c X = \sum_{j=1}^m (\Pi_j X) e_j \text{ and } A_c = \text{diag}(\beta_1, \dots, \beta_m). \tag{4.20}$$

Recall that according to the convention (2.8) (of Sect. 2.1) made throughout this article, the reduced state space  $E_c$  is spanned by modes that come either as conjugate pairs or as a real eigenvector. As a result,  $\Pi_c X$  is real if  $X$  is real.

For each given unresolved mode  $e_n$  ( $n \geq m + 1$ ), a parameterization  $y_n^{(1)}$  of the corresponding unresolved variable

$$Y_n = \Pi_n y, \tag{4.21}$$

is obtained from the following backward–forward system:

$$\frac{dy_c^{(1)}}{ds} = A_c y_c^{(1)} + \Pi_c F, \quad s \in [-\tau, 0], \tag{4.22a}$$

$$\frac{dy_n^{(1)}}{ds} = \beta_n y_n^{(1)} + \Pi_n B(y_c^{(1)}, y_c^{(1)}) + \Pi_n F, \quad s \in [-\tau, 0], \tag{4.22b}$$

$$\text{with } y_c^{(1)}(s)|_{s=0} = \xi \in E_c, \text{ and } y_n^{(1)}(s)|_{s=-\tau} = 0. \tag{4.22c}$$

Note that the solution to (4.22a) is given by:

$$y_c^{(1)}(t) = e^{A_c t} \xi - \int_t^0 e^{A_c(t-s)} \Pi_c F \, ds, \quad t \in [-\tau, 0], \tag{4.23}$$

which admits the following explicit expression:

$$y_c^{(1)}(t) = \sum_{j=1}^m \left( e^{\beta_j t} \xi_j + \gamma_j(t) \Pi_j F \right) e_j, \tag{4.24}$$

where

$$\gamma_j(t) = \begin{cases} \frac{\exp(\beta_j t) - 1}{\beta_j}, & \text{if } \beta_j \neq 0, \\ t, & \text{otherwise.} \end{cases} \tag{4.25}$$

The solution to (4.22b) is given by:

$$y_n^{(1)}[\xi](t) = \int_{-\tau}^t e^{\beta_n(t-s)} \Pi_n B(y_c^{(1)}(s), y_c^{(1)}(s)) ds + \int_{-\tau}^t e^{\beta_n(t-s)} \Pi_n F ds, \quad t \in [-\tau, 0], \tag{4.26}$$

which leads to the following parameterization for the high mode  $e_n$ :

$$\Phi_n(\tau, \xi) = \int_{-\tau}^0 e^{-\beta_n s} \Pi_n B(y_c^{(1)}(s), y_c^{(1)}(s)) ds + \int_{-\tau}^0 e^{-\beta_n s} \Pi_n F ds. \tag{4.27}$$

By using (4.24) in the nonlinear term  $\Pi_n B(y_c^{(1)}(s), y_c^{(1)}(s))$  and expanding this term, the first integral  $I$  in the RHS of (4.27) becomes after simplification

$$\begin{aligned} I &= \sum_{i,j=1}^m U_{i,j}^n(\tau, \boldsymbol{\beta}) B_{i,j}^n F_i F_j + \sum_{i,j=1}^m V_{i,j}^n(\tau, \boldsymbol{\beta}) F_j (B_{i,j}^n + B_{j,i}^n) \xi_i \\ &+ \sum_{i,j=1}^m D_{i,j}^n(\tau, \boldsymbol{\beta}) B_{i,j}^n \xi_i \xi_j, \end{aligned} \tag{4.28}$$

where

$$B_{i,j}^n = \langle B(\mathbf{e}_i, \mathbf{e}_j), \mathbf{e}_n^* \rangle, \tag{4.29}$$

the coefficients  $D_{i,j}^n(\tau, \boldsymbol{\beta})$  of the quadratic terms (in the  $\xi$ -variable) are given by

$$D_{i,j}^n(\tau, \boldsymbol{\beta}) = \begin{cases} \frac{1 - \exp(-(\beta_i + \beta_j - \beta_n)\tau)}{\beta_i + \beta_j - \beta_n}, & \text{if } \beta_i + \beta_j - \beta_n \neq 0, \\ \tau, & \text{otherwise,} \end{cases} \tag{4.30}$$

while the coefficients in the constant and linear terms are given respectively by

$$U_{i,j}^n(\tau, \boldsymbol{\beta}) = \begin{cases} \frac{1}{\beta_i \beta_j} \left( D_{i,j}^n(\tau, \boldsymbol{\beta}) - \frac{1 - \exp(-\tau(\beta_i - \beta_n))}{\beta_i - \beta_n} - \frac{1 - \exp(-\tau(\beta_j - \beta_n))}{\beta_j - \beta_n} - \frac{1 - \exp(\tau\beta_n)}{\beta_n} \right), & \text{if } \beta_i \neq 0 \text{ and } \beta_j \neq 0, \\ \frac{1}{\beta_i} \left( \frac{\tau \exp(-\tau(\beta_i - \beta_n))}{\beta_i - \beta_n} - \frac{1 - \exp(-\tau(\beta_i - \beta_n))}{(\beta_i - \beta_n)^2} + \frac{\tau \exp(\tau\beta_n)}{\beta_n} + \frac{1 - \exp(\tau\beta_n)}{(\beta_n)^2} \right), & \text{if } \beta_i \neq 0 \text{ and } \beta_j = 0, \\ \frac{1}{\beta_j} \left( \frac{\tau \exp(-\tau(\beta_j - \beta_n))}{\beta_j - \beta_n} - \frac{1 - \exp(-\tau(\beta_j - \beta_n))}{(\beta_j - \beta_n)^2} + \frac{\tau \exp(\tau\beta_n)}{\beta_n} + \frac{1 - \exp(\tau\beta_n)}{(\beta_n)^2} \right), & \text{if } \beta_i = 0 \text{ and } \beta_j \neq 0, \\ -\frac{(\tau)^2 \exp(\tau\beta_n)}{\beta_n} - \frac{2}{\beta_n} \left( \frac{\tau \exp(\tau\beta_n)}{\beta_n} + \frac{1 - \exp(\tau\beta_n)}{(\beta_n)^2} \right), & \text{if } \beta_i = 0 \text{ and } \beta_j = 0, \end{cases} \tag{4.31}$$

and

$$V_{i,j}^n(\tau, \boldsymbol{\beta}) = \begin{cases} \frac{1 - \exp(-\tau(\beta_i + \beta_j - \beta_n))}{\beta_j(\beta_i + \beta_j - \beta_n)} - \frac{1 - \exp(-\tau(\beta_i - \beta_n))}{\beta_j(\beta_i - \beta_n)}, & \text{if } \beta_j \neq 0, \\ \frac{\tau \exp(-\tau(\beta_i - \beta_n))}{\beta_i - \beta_n} - \frac{1 - \exp(-\tau(\beta_i - \beta_n))}{(\beta_i - \beta_n)^2}, & \text{otherwise.} \end{cases} \tag{4.32}$$

By adding  $\int_{-\tau}^0 e^{\beta_n(t-s)} \Pi_n F ds$  to the constant and linear terms in  $I$ , we can form

$$\Gamma_n(F, \boldsymbol{\beta}, \tau, \xi) = \sum_{i,j=1}^m U_{i,j}^n(\tau, \boldsymbol{\beta}) B_{i,j}^n F_i F_j + \sum_{i,j=1}^m V_{i,j}^n(\tau, \boldsymbol{\beta}) F_j (B_{i,j}^n + B_{j,i}^n) \xi_i - \frac{1 - e^{\tau\beta_n}}{\beta_n} \Pi_n F, \tag{4.33}$$

leading thus to

$$\Phi_n(\tau, \boldsymbol{\beta}, \xi) = \Gamma_n(F, \boldsymbol{\beta}, \tau, \xi) + \sum_{i,j=1}^m D_{i,j}^n(\tau, \boldsymbol{\beta}) B_{i,j}^n \xi_i \xi_j. \tag{4.34}$$

The optimal  $\tau$  value for each of the unresolved mode is obtained by minimizing the corresponding parameterization defect  $\mathcal{Q}_n$  defined in (4.13). In other words, given a fully resolved solution  $y(t)$  of the underlying  $N$ -dimensional ODE system (4.16) available over a training interval  $[0, T]$  (after possible removal of transient dynamics), we solve for each  $m + 1 \leq n \leq N$  the following minimization problem

$$\begin{cases} \min_{\tau} \int_0^T |\Pi_n y(t) - \Phi_n(\tau, \boldsymbol{\beta}, \Pi_c y(t))|^2 dt, \\ \text{where } \Phi_n(\tau, \boldsymbol{\beta}, \xi) \text{ is given by (4.34)}. \end{cases} \tag{4.35}$$

The resulting minimizers  $\tau_n^*$  whose collection is denoted by  $\boldsymbol{\tau}^*$ , allows us then to define the following optimal parameterization within the LIA class

$$\Phi_{\boldsymbol{\tau}^*}^{(1)}(\xi) = \sum_{n=m+1}^N \Phi_n(\tau_n^*, \boldsymbol{\beta}, \xi) \mathbf{e}_n. \tag{4.36}$$

In what follows we will sometimes denote by  $\text{LIA}(\boldsymbol{\tau})$ , the parameterization  $\Phi_{\boldsymbol{\tau}}^{(1)}$  (see 4.36) with  $\Phi_n$  given by (4.34).

Although providing in general only a suboptimal solution to the more general family of minimization problems (3.14) discussed in Sect. 3.1, we will refer to the optimal LIA,  $\Phi_{\boldsymbol{\tau}^*}^{(1)}$ , as the optimal PM when the context is clear; see Sect. 5 below. As mentioned above, Appendix presents a simple gradient-descent method to determine efficiently, the  $\tau_n^*$ 's (and thus  $\boldsymbol{\tau}^*$ ) in practice; as pointed out above, see however Remark 8 below in the presence of local minima.

**Remark 4** Note that for  $F = 0$ , and when  $\beta_i + \beta_j > \beta_n$ , the LIA class includes the leading-order approximation,  $h_2$ , given by (2.47)–(2.48) (with  $k = 2$ ) of the invariant manifold dealt with in Sect. 2.2, in the sense that then for all  $\xi$  in  $E_c$ ,

$$\lim_{\tau \rightarrow \infty} \Phi_{\boldsymbol{\tau}}^{(1)}(\xi) = h_2(\xi). \tag{4.37}$$

Furthermore  $\Phi_{\boldsymbol{\tau}}^{(1)} \equiv 0$  when  $\tau = 0$ , i.e. the LIA class contains Galerkin approximations of dimension  $m = \dim(E_c)$ .

**Remark 5** Note that in the expression of  $\Phi_n$  given by (4.34), the term  $\Gamma_n(F, \boldsymbol{\beta}, \tau, \xi)$  takes into account interactions between the low-mode components of the forcing,  $F$ , as well as cross-interactions between the low-mode components of  $F$  and the low-mode variable  $\xi$  in  $E_c$ . It also includes the  $n^{\text{th}}$  high-mode component of the forcing.

We emphasize that these formulas can be derived for PDEs as well, as rooted in the backward–forward method recalled above and initially introduced for PDEs (possibly driven

by a multiplicative linear noise) in [31, Chap. 4]; see also [26, Sec. 3.2]. The main novelty compared to [31, Chap. 4] is the idea of optimizing, high-mode by high-mode, the backward integration time,  $\tau_n$ , of Eq. (4.22), by minimization of the parameterization defect  $\mathcal{Q}_n$ .

**Remark 6** Note that when  $\beta_{n+1} = \overline{\beta_n}$ , we have  $e_{n+1}^* = \overline{e_n^*}$  and therefore  $\Pi_{n+1}X = \overline{\Pi_n X}$  when  $X$  is real according to (4.19). Furthermore when  $B(u_c^{(1)}(s), u_c^{(1)}(s))$  and  $F$  are real, we have according to (4.27), that  $\Phi_{n+1} = \overline{\Phi_n}$  when evaluated on a real vector  $\xi$  of  $E_c$ .

#### 4.4 Parametric Quasi-stationary Approximation and Another Cost Functional

Other cost functionals than  $\mathcal{Q}_n(\tau_n, T)$  could have been considered to seek for optimal LIA. For instance,

$$\mathcal{J}_n(\tau, T; \Phi_n) = \left| \overline{[\Pi_n y(t)]^2} - \overline{[\Phi_n(\tau, \beta, y_c(t))]}^2 \right|. \tag{4.38}$$

Here  $\overline{(\cdot)}$  denotes a time-averaging over an interval of length  $T$ . The minimization of the  $\mathcal{J}_n$ 's leads in general to different optimal LIA compared to the one obtained by solving the minimization problems (4.35).

If the mean value of  $y_n(t)$  is zero, minimizing  $\mathcal{Q}_n$  consists of minimizing the variance of the residual error, i.e.  $|y_n - f(\tau, y_c)|^2$ , for a given parameterization  $f(\tau, \cdot)$ . By construction, minimizing  $\mathcal{J}_n$  consists instead of minimizing the residual error of the variance approximation, i.e.  $||y_n|^2 - |f(\tau, y_c)|^2|$ . The latter cost functional better accounts for the distribution of energy across the modes; see Sect. 6.3 for an illustration.

Although a geometric interpretation like (4.15) is not available for such a cost functional, minimizing (4.38) leads in general to a better reproduction of the energy budget across the high modes. For this reason, the cost functional (4.38) will be adopted for certain applications; see Sect. 6 below.

While the LIA class may be preferred when forcing terms are present (especially when e.g. only the low modes are forced), another class of parameterization is particularly suited to systems that do not include forcing terms. Still, in presence of such terms this other class may be relevant in certain applications (when e.g. only the high modes are forced) and thus we present hereafter the derivation of the corresponding formulas that take into account (constant) forcing as for LIA.

This class is rooted in the following *Quasi-Stationary approximation (QSA)* for Eq. (4.16)

$$\Pi_s A z + \Pi_s B(\xi, \xi) + \Pi_s F = 0, \quad \xi \in E_c, z \in E_s. \tag{4.39}$$

The QSA arises in homogeneous turbulence theory [64]; see Remark 7 below. It consists of neglecting the terms  $\Pi_s [B(y_s, y_c) + B(y_s, y_s)]$  in virtue of the energy content of the small structures being small, and following a suggestion of Kraichnan balancing  $dy_s/dt$  with  $\Pi_s B(y_c, y_s)$ , i.e., with the advection of small eddies by large eddies; see [68].

After solving (4.39), the QSA parameterization is then obtained as  $z = K(\xi)$  with  $K$  given by

$$K(\xi) = (-A_s)^{-1}(\Pi_s B(\xi, \xi) + \Pi_s F). \tag{4.40}$$

In contrast, the standard LIA is obtained by solving the backward-system (4.5) asymptotically, and the parameterization LIA( $\tau$ ) is obtained after solving the backward-systems (4.22).

Similar to what precedes, we use a dynamic version of Eq. (4.39) to get access to a parametric family of dynamically-based parameterizations such that  $K$  belongs to this family,

as in Remark 4 regarding the LIA class that includes  $h_2$ . By assuming  $A$  diagonal (in  $\mathbb{C}$ ), we consider thus for  $\tau > 0$

$$\begin{aligned} \frac{dz_n}{ds} &= \beta_n z_n + \Pi_n B(\xi, \xi) + \Pi_n F, \\ z_n(-\tau) &= 0. \end{aligned} \tag{4.41}$$

Solving Eq. (4.41) for each  $n$ , leads then to the following high-mode parameterization

$$\Psi_n(\tau, \beta, \xi) = \delta_n(\tau) \left( \sum_{i,j=1}^m B_{ij}^n \xi_i \xi_j + \Pi_n F \right), \tag{4.42}$$

with  $B_{ij}^n$  given by (4.29) and where

$$\delta_n(\tau) = \begin{cases} \beta_n^{-1} (e^{\beta_n \tau} - 1), & \text{if } \beta_n \neq 0, \\ \tau, & \text{otherwise.} \end{cases} \tag{4.43}$$

We arrive then at the following *parametric* QSA or simply denoted  $\text{QSA}(\tau)$ :

$$\Psi_\tau(\xi) = \sum_{n=m+1}^N \Psi_n(\xi, \beta, \xi) e_n. \tag{4.44}$$

In particular, if  $\beta_n < 0$  for all  $n \geq m + 1$ , since  $\delta_n(\tau) \xrightarrow{\tau \rightarrow \infty} -\beta_n^{-1}$ , then for all  $\xi$  in  $E_c$ ,

$$\lim_{\tau \rightarrow \infty} \Psi_\tau(\xi) = K(\xi), \tag{4.45}$$

with  $K$  given by (4.40). Furthermore  $\Psi_\tau \equiv 0$  when  $\tau = 0$ , i.e. the QSA class contains also Galerkin approximations of dimension  $m = \dim(E_c)$ .

In Sect. 6 below, we show applications of this parameterization class (called the QSA class), from which the *optimal* QSA is determined by solving for each  $m + 1 \leq n \leq N$  the following minimization problem

$$\left\{ \begin{array}{l} \min_{\tau} \left| \left[ \overline{\Pi_n y(t)} \right]^2 - \left[ \overline{\Psi_n(\tau, \beta, y_c(t))} \right]^2 \right| \\ \text{where } \Psi_n(\tau, \beta, \xi) \text{ is given by (4.42).} \end{array} \right. \tag{4.46}$$

The algorithm presented in Appendix to solve (4.35), can be easily adapted to solve (4.46) (after smoothing) and thus to determine the minimizers  $\tau_n^*$ ; the details are left to the reader.

As recalled above, Remark 4 emphasizes that the leading-order approximation  $h_2(\xi)$  (given by (2.32) with  $G_k = B$ ) of the invariant manifold dealt with in Sect. 2.2 may be obtained as a limit  $\text{LIA}(\tau)$ : here (4.45) shows that the standard QSA,  $K(\xi)$ , may also be obtained as a limit of  $\text{QSA}(\tau)$ . It is noteworthy that the theory of approximation of invariant manifolds shows that these two limiting objects,  $h_2(\xi)$  and  $K(\xi)$ , are actually related. More precisely, [31, Lemma 4.1] shows that near the first criticality and when  $F = 0$ , the QSA and the leading-order approximation  $h_2(\xi)$ , are linked according to the following approximation relation

$$h_2(\xi) = (-A_s)^{-1} \Pi_s B(\xi, \xi) + O(\|\xi\|^2), \quad \forall \xi \in E_c. \tag{4.47}$$

Thus when  $F = 0$ , one should not expect much difference between the parameterizations  $\text{LIA}(\tau)$  and  $\text{QSA}(\tau)$  for large values of  $\tau$  (and under the appropriate conditions on the  $\beta_k$ 's).

However, if  $\tau$  has components with small values, differences are expected to occur between the corresponding  $\text{LIA}(\tau)$  and  $\text{QSA}(\tau)$  parameterizations. To better appreciate these differences, let us introduce the function  $f(\tau) = p^{-1}(1 - e^{-p\tau})$  and note that  $f(\tau) = \delta_n(\tau)$  when

$p = -\beta_n$  and that  $f(\tau) = D_{ij}^n(\tau)$  (given by (4.30)) when  $p = \beta_i + \beta_j - \beta_n$ . Thus when  $F = 0$  the LIA and QSA classes differ only by these coefficients.

To simplify, let us assume that the eigenvalues of  $A$  are real and that  $E_c$  contains all and only the unstable modes. In this case,  $p = \beta_i + \beta_j - \beta_n$  is always bigger than  $p = -\beta_n$ . Now if we assume furthermore that  $p > 0$  (in either case) we have

$$0 \leq f(\tau) < p^{-1}, \quad (4.48)$$

and therefore due to (4.42) and (4.34) (with  $F = 0$ ), the range of the coefficient in front of each monomial is larger for  $\Psi_n(\tau, \xi)$  than for  $\Phi_n(\tau, \xi)$ , in this case. This allows in practice for  $\Psi_n(\tau, \xi)$  to span a larger range of values which in turn may lead to smaller values of  $\mathcal{Q}_n$  or  $\mathcal{J}_n$ . The situation described here is exactly what happens for the closure problem considered below in Sect. 6 within the context of Kuramoto-Sivashinsky turbulence, when one sets the cutoff wavenumber to be the highest wavenumber among the unstable modes. As we will show in Sect. 6 for different turbulent regimes, the QSA( $\tau$ ) when optimized (either for  $\mathcal{Q}_n$  or  $\mathcal{J}_n$ ) provides a drastic improvement compared to the standard QSA,  $K(\xi)$ , for such cutoff scales.

**Remark 7** As mentioned right after (4.39), the QSA is a well-known parameterization in homogeneous turbulence and has been rigorously proved to provide an AIM in [64] for the 2D Navier–Stokes equations. The QSA also arises in atmospheric turbulence in the so-called nonlinear normal-mode initialization [6,46,47,74,120,127,167]; see [49] for rigorous results. Nevertheless, when the cutoff wavelength is too low within the inertial range it is known that the standard QSA suffers from over-parameterization leading then to errors in the backscatter transfer of energy, i.e. errors in the modeling of the parameterized (small) scales that contaminate gradually the larger scales. We show in Sect. 6, in the context of KS turbulence that by solving the minimization problems (4.46), the optimal QSA fixes this problem remarkably.

## 5 Applications to a Reduced-Order Rayleigh–Bénard System

In this section, we apply the PM approach—as presented in its practical aspects in Sect. 4—to a Galerkin system of nine nonlinear ODEs examined in [145] and obtained from a triple Fourier expansion to the Boussinesq equations governing thermal convection in a 3D spatial domain.

The PM approach is applied to two parameter regimes for this 9D Rayleigh–Bénard (RB) convection system: (i) a regime located right after the first period-doubling bifurcation occurring for this system (Sect. 5.2), and (ii) a regime corresponding to chaotic dynamics that takes place right after the period-doubling cascade (Sect. 5.3).

We show hereafter for both cases, that, given a reduced state space,  $E_c$ , the dynamically-based parameterization, LIA( $\tau$ ), of Sect. 4.3 when optimized in the  $\tau$ -variable, by minimizing<sup>7</sup> the parameterization defects (4.35), provides efficient low-dimensional closures of the original RB system.

To prepare the numerical results of Sects. 5.2 and 5.3, we first recall the 9D RB system and give the details of its LIA( $\tau$ )-closure in Sect. 5.1. We emphasize that the closures are determined in each case with respect to a mean state  $\overline{C}$ , leading in particular to equations for the perturbed variable,  $C - \overline{C}$ , of the form (2.19).

<sup>7</sup> While maximizing, in certain circumstances, the parameterization correlation,  $c(t)$ , given by (3.6); see Sect. 5.2.

### 5.1 Optimal PM Closure

Like [145], our study below deals with three-dimensional cells with square planform in dissipative Rayleigh-Bénard convection. In that respect, the 9D RB system derived in [145, Section 2] takes the form:

$$\begin{aligned}
 \dot{C}_1 &= -\sigma b_1 C_1 - C_2 C_4 + b_4 C_4^2 + b_3 C_3 C_5 - \sigma b_2 C_7, \\
 \dot{C}_2 &= -\sigma C_2 + C_1 C_4 - C_2 C_5 + C_4 C_5 - \frac{\sigma}{2} C_9, \\
 \dot{C}_3 &= -\sigma b_1 C_3 + C_2 C_4 - b_4 C_2^2 - b_3 C_1 C_5 + \sigma b_2 C_8, \\
 \dot{C}_4 &= -\sigma C_4 - C_2 C_3 - C_2 C_5 + C_4 C_5 + \frac{\sigma}{2} C_9, \\
 \dot{C}_5 &= -\sigma b_5 C_5 + \frac{1}{2} C_2^2 - \frac{1}{2} C_4^2, \\
 \dot{C}_6 &= -b_6 C_6 + C_2 C_9 - C_4 C_9, \\
 \dot{C}_7 &= -b_1 C_7 - r C_1 + 2 C_5 C_8 - C_4 C_9, \\
 \dot{C}_8 &= -b_1 C_8 + r C_3 - 2 C_5 C_7 + C_2 C_9, \\
 \dot{C}_9 &= -C_9 - r C_2 + r C_4 - 2 C_2 C_6 + 2 C_4 C_6 + C_4 C_7 - C_2 C_8.
 \end{aligned} \tag{5.1}$$

Here  $\sigma$  denotes the Prandtl number, and  $r$  denotes the reduced Rayleigh number defined to be the ratio between the Rayleigh number  $R$  and its critical value  $R_c$  at which the convection sets in. The coefficients  $b_i$ 's are given by

$$\begin{aligned}
 b_1 &= \frac{4(1+a^2)}{1+2a^2}, & b_2 &= \frac{1+2a^2}{2(1+a^2)}, & b_3 &= \frac{2(1-a^2)}{1+a^2}, \\
 b_4 &= \frac{a^2}{1+a^2}, & b_5 &= \frac{8a^2}{1+2a^2}, & b_6 &= \frac{4}{1+2a^2},
 \end{aligned} \tag{5.2}$$

with  $a = \frac{1}{2}$  being the critical horizontal wavenumber of the square convection cell.

With the purpose to derive a closure for Eq. (5.1), we first put Eq. (5.1) into the following compact form:

$$\dot{\mathbf{C}} = \mathbf{A}\mathbf{C} + \mathbf{B}(\mathbf{C}, \mathbf{C}), \tag{5.3}$$

where  $\mathbf{C} = (C_1, \dots, C_9)^T$ ,  $\mathbf{A}$  is the  $9 \times 9$  matrix given by

$$\mathbf{A} = \begin{pmatrix}
 -\sigma b_1 & 0 & 0 & 0 & 0 & 0 & -\sigma b_2 & 0 & 0 \\
 0 & -\sigma & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\sigma}{2} \\
 0 & 0 & -\sigma b_1 & 0 & 0 & 0 & 0 & \sigma b_2 & 0 \\
 0 & 0 & 0 & -\sigma & 0 & 0 & 0 & 0 & \frac{\sigma}{2} \\
 0 & 0 & 0 & 0 & -\sigma b_5 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & -b_6 & 0 & 0 & 0 \\
 -r & 0 & 0 & 0 & 0 & 0 & -b_1 & 0 & 0 \\
 0 & 0 & r & 0 & 0 & 0 & 0 & -b_1 & 0 \\
 0 & -r & 0 & r & 0 & 0 & 0 & 0 & -1
 \end{pmatrix}, \tag{5.4}$$



and the quadratic nonlinearity  $B$  is defined by

$$B(\boldsymbol{\phi}, \boldsymbol{\psi}) = \begin{pmatrix} -\phi_2\psi_4 + b_4\phi_4\psi_4 + b_3\phi_3\psi_5 \\ \phi_1\psi_4 - \phi_2\psi_5 + \phi_4\psi_5 \\ \phi_2\psi_4 - b_4\phi_2\psi_2 - b_3\phi_1\psi_5 \\ -\phi_2\psi_3 - \phi_2\psi_5 + \phi_4\psi_5 \\ \frac{1}{2}\phi_2\psi_2 - \frac{1}{2}\phi_4\psi_4 \\ \phi_2\psi_9 - \phi_4\psi_9 \\ 2\phi_5\psi_8 - \phi_4\psi_9 \\ -2\phi_5\psi_7 + \phi_2\psi_9 \\ -2\phi_2\psi_6 + 2\phi_4\psi_6 + \phi_4\psi_7 \end{pmatrix} \tag{5.5}$$

for any  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_9)^{\text{tr}}$  and  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_9)^{\text{tr}}$  in  $\mathbb{C}^9$ .

We consider next fluctuations defined with respect to a mean state. In that respect, we subtract from  $\mathbf{C}(t) = (C_1(t), \dots, C_9(t))$  its mean value  $\bar{\mathbf{C}}$ , which is estimated, in practice, from simulation of Eq. (5.1) on the same training interval  $T$  than used to optimize our parameterizations hereafter. The corresponding ODE system for the fluctuation variable,  $\mathbf{D} = \mathbf{C} - \bar{\mathbf{C}}$ , is then given by:

$$\frac{d\mathbf{D}}{dt} = L\mathbf{D} + B(\mathbf{D}, \mathbf{D}) + A\bar{\mathbf{C}} + B(\bar{\mathbf{C}}, \bar{\mathbf{C}}), \tag{5.6}$$

with

$$L\mathbf{D} = A\mathbf{D} + B(\bar{\mathbf{C}}, \mathbf{D}) + B(\mathbf{D}, \bar{\mathbf{C}}). \tag{5.7}$$

Denote the spectral elements of the matrix  $L$  by  $\{(\beta_j, \mathbf{e}_j) : 1 \leq j \leq 9\}$  and those of  $L^*$  by  $\{(\beta_j^*, \mathbf{e}_j^*) : 1 \leq j \leq 9\}$ . By taking the expansion of  $\mathbf{D}$  under the eigenbasis of  $L$ ,

$$\mathbf{D} = \sum_{j=1}^9 y_j \mathbf{e}_j \quad \text{with} \quad y_j = \langle \mathbf{D}, \mathbf{e}_j^* \rangle, \tag{5.8}$$

and assuming that  $L$  is diagonal under its eigenbasis, we rewrite Eq. (5.6) in the variable  $\mathbf{y} = (y_1, \dots, y_9)^{\text{tr}}$  as follows:

$$\dot{y}_j = \beta_j y_j + \sum_{k,\ell=1}^9 \langle B(\mathbf{e}_k, \mathbf{e}_\ell), \mathbf{e}_j^* \rangle y_k y_\ell + \langle A\bar{\mathbf{C}} + B(\bar{\mathbf{C}}, \bar{\mathbf{C}}), \mathbf{e}_j^* \rangle, \quad j = 1, \dots, 9. \tag{5.9}$$

Now we take the reduced state space  $E_c$  to be spanned by the first  $m$  eigenvectors of  $A$  for some  $m < 9$ , where the eigenvalues are ranked according to the ordering (2.12) adopted here from Sect. 2.1, i.e. the modes are ordered according to their linear rate of growth/decay. For each  $m + 1 \leq n \leq 9$ , we approximate the (unresolved) variable  $y_n$  by the parameterization  $\Phi_n(\tau_n^*, \boldsymbol{\beta}, \cdot)$  obtained from (4.34) after minimization of (4.35), given a training interval of length  $T$  that will be specified hereafter depending on the context.

The resulting  $m$ -dimensional optimal PM closure (in the LIA class) reads then

$$\begin{aligned} \dot{x}_j &= \beta_j x_j + \sum_{k,\ell=1}^m \langle B(\mathbf{e}_k, \mathbf{e}_\ell), \mathbf{e}_j^* \rangle x_k x_\ell \\ &+ \sum_{k=1}^m \sum_{\ell=m+1}^9 \left( \langle B(\mathbf{e}_\ell, \mathbf{e}_k), \mathbf{e}_j^* \rangle + \langle B(\mathbf{e}_k, \mathbf{e}_\ell), \mathbf{e}_j^* \rangle \right) x_k \Phi_\ell(\tau_\ell^*, \boldsymbol{\beta}, x_1, \dots, x_m) \end{aligned}$$

$$\begin{aligned}
 &+ \sum_{k,\ell=m+1}^9 \langle B(\mathbf{e}_\ell, \mathbf{e}_k), \mathbf{e}_j^* \rangle \Phi_k(\tau_k^*, \boldsymbol{\beta}, x_1, \dots, x_m) \Phi_\ell(\tau_\ell^*, \boldsymbol{\beta}, x_1, \dots, x_m) \\
 &+ \langle A\bar{\mathbf{C}} + B(\bar{\mathbf{C}}, \bar{\mathbf{C}}), \mathbf{e}_j^* \rangle, \quad j = 1, \dots, m.
 \end{aligned}
 \tag{5.10}$$

Once the optimal PM closure (5.10) is solved, an approximation,  $\mathbf{C}^{\text{PM}}(t)$ , of the solution  $\mathbf{C}(t)$  to the original system (5.1) is obtained as follows,

$$\mathbf{C}^{\text{PM}}(t) = \sum_{j=1}^m x_j(t) \mathbf{e}_j + \sum_{n=m+1}^9 \Phi_n(\tau_n^*, \boldsymbol{\beta}, x_1(t), \dots, x_m(t)) \mathbf{e}_n + \bar{\mathbf{C}}.
 \tag{5.11}$$

### 5.2 Closure in a Period-Doubling Regime

As the reduced Rayleigh number  $r$  increases, the first period-doubling bifurcation for Eq. (5.1) occurs at approximately  $r = 13.97$ , and the dynamics becomes chaotic at approximately  $r = 14.22$  after successive periodic-doubling bifurcations. We have set  $r = 14.1$  to examine how the PM approach operates in a period-doubling regime. As a benchmark, for the same reduced dimension,  $m$ , as used for the optimal PM closure (5.10), we determine the reduced system of the form (2.17) in which  $h$  is replaced by the approximation  $h_2$  given by (2.47)–(2.48) (with  $k = 2$ ) in Theorem 2, i.e. the parameterization that provides the leading-order approximation of the local invariant manifold for an equilibrium. Applying the ideas of Sect. 2.1 to Eq. (5.1), the calculations of  $h_2$  are made about a steady state of Eq. (5.1), taken here to be the closest steady state  $\bar{\mathbf{Y}}$  to the mean state,  $\bar{\mathbf{C}}$ . If one denotes by  $F$  the RHS of Eq. (5.1), the linear part  $A$  in (2.2) is then taken to be given by  $DF(\bar{\mathbf{Y}})$ .

Thus, denoting by  $(\lambda_j, \mathbf{f}_j)$  the spectral elements of  $DF(\bar{\mathbf{Y}})$  and those of  $(DF(\bar{\mathbf{Y}}))^*$  by  $(\lambda_j^*, \mathbf{f}_j^*)$ , the following reduced system based on the invariant manifold approximation  $h_2$ ,

$$\begin{aligned}
 \dot{z}_j &= \lambda_j z_j + \sum_{k,\ell=1}^m \langle B(\mathbf{f}_k, \mathbf{f}_\ell), \mathbf{f}_j^* \rangle z_k z_\ell \\
 &+ \sum_{k=1}^m \sum_{\ell=m+1}^9 \left( \langle B(\mathbf{f}_\ell, \mathbf{f}_k), \mathbf{f}_j^* \rangle + \langle B(\mathbf{f}_k, \mathbf{f}_\ell), \mathbf{f}_j^* \rangle \right) z_k h_{2,\ell}(z_1, \dots, z_m) \\
 &+ \sum_{k,\ell=m+1}^9 \langle B(\mathbf{f}_\ell, \mathbf{f}_k), \mathbf{f}_j^* \rangle h_{2,k}(z_1, \dots, z_m) h_{2,\ell}(z_1, \dots, z_m), \quad j = 1, \dots, m,
 \end{aligned}
 \tag{5.12}$$

serves us as a benchmark. Here  $h_{2,n}$  ( $6 \leq n \leq 9$ ) is given by (2.48) in which  $G_k$  is replaced by  $B$  given by (5.5) and the  $(\beta_j, \mathbf{e}_j)$ 's replaced by the  $(\lambda_j, \mathbf{f}_j)$ 's.

From the solution  $\mathbf{z}(t) = (z_1(t), \dots, z_m(t))^{\text{tr}}$  of the reduced system (5.12), the following approximation of  $\mathbf{C}(t)$  is then obtained,

$$\mathbf{C}^{\text{IM}}(t) = \sum_{j=1}^m z_j(t) \mathbf{f}_j + \sum_{n=m+1}^9 h_{2,n}(z_1(t), \dots, z_m(t)) \mathbf{f}_n + \bar{\mathbf{Y}}.
 \tag{5.13}$$

For the numerical results presented hereafter, the reduced state space  $E_c$  is taken to be spanned by the first five eigenmodes, i.e. by setting  $m = 5$  in this section. To determine our optimal PM closure, we used the quadratic parameterization,  $\Phi_n(\tau, \cdot)$  given by (4.34), in order to

parameterize each of the modes  $e_n$  with  $6 \leq n \leq 9$ . For each  $6 \leq n \leq 9$ , each of this parameterization is optimized in the  $\tau$ -variable by minimizing the parameterization defect

$$Q_n(\tau, T; t_0) = \int_{t_0}^{t_0+T} |\Pi_n y(t) - \Phi_n(\tau, \beta, \Pi_c y(t))|^2 dt, \tag{5.14}$$

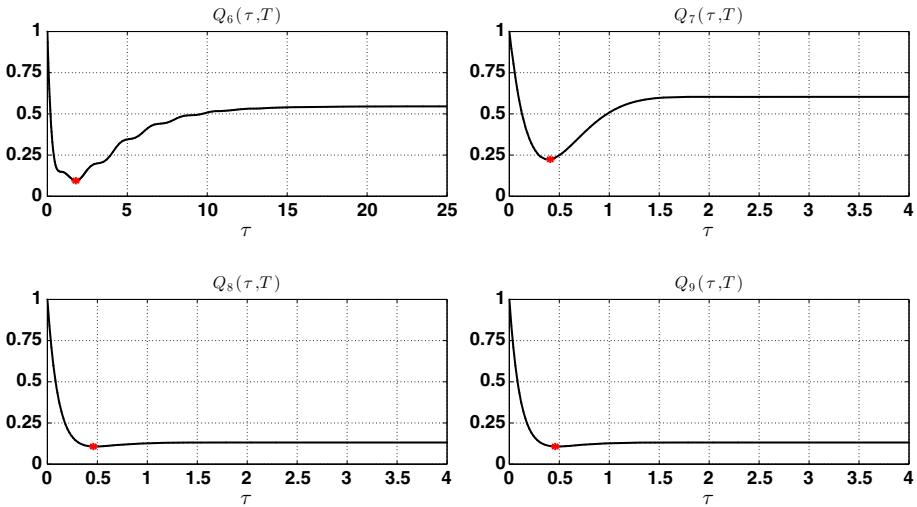
for some  $t_0$  chosen so that transient dynamics has been removed. Since the dynamics to emulate by a closure is here periodic, we selected  $T = 3T_p/4$ , where  $T_p$  ( $\approx 17.25$ ) corresponds to the period of the solution to the 9D RB system (5.1) in order to do not use all the available information about the periodic orbit. Other choices could have been made for the training interval such as  $T = T_p/2$ . Note that we observed that the choice of  $t_0$  plays a key role here. As discussed in Remark 8 below, depending on  $t_0$  the global minimizer  $\tau_n^*$  of  $Q_n$  here, does not provide necessarily the best parameterization within the  $\Phi_n$ -class, and one may have to rely on the parameterization correlation  $c(t)$  (see (3.6)) to discriminate between other local minimizers of  $Q_n$ . The results presented below corresponds to a time origin,  $t_0$ , for which the global minimizer of the  $Q_n$ 's lead to the best parameterization within the  $\Phi_n$ -class.

Despite the aforementioned  $t_0$ -dependence, for the sake of keeping the notations as concise as possible, the dependence on  $t_0$  will not be made apparent for the numerical results presented below. This being said, whatever the length  $T$  of the training interval, we have used the same training interval  $[t_0, t_0 + T]$  to estimate the mean state,  $\bar{C}$ , than used for evaluating the cost functionals  $Q_n$  in (5.14).

The mean state,  $\bar{C}$ , plays a key role in the determination of the closure as it determines the linear part  $L$  defined in (5.7), and thus the spectral elements  $(\beta_j, e_j)$  arising in the formulation of the parameterizations,  $\Phi_n(\tau, \cdot)$  (see (4.34)), and of the corresponding closure (5.10). Numerically, a fourth-order Runge-Kutta method is used to solve Eq. (5.9) with a time-step size taken to be  $\delta t = 5 \times 10^{-3}$  to determine a numerical approximation of  $y(t)$ . The minimization algorithm for the parameterization defect described in Appendix is used to find the minimizer  $\tau_n^*$  of  $Q_n(\tau, T)$ . In that respect, the trapezoid rule is used to approximate the integrals involved in (A.6).

The mapping  $\tau \mapsto Q_n(\tau, T)$  is shown in Fig. 7 from  $n = 6$  to  $n = 9$  and exhibits a non-convex behavior for each  $n$ , although this behavior is more pronounced for  $n = 6$  and  $n = 7$ . The minimizer  $\tau_n^*$  found by the algorithm of Appendix corresponds to the abscissa of the red dot shown in each of the panels. Among the parameterized modes, the minima of  $Q_n$  that are the most clearly distinguishable occur for the “adjacent” modes —  $e_6$  and  $e_7$ — located next to the cutoff dimension, i.e. for the modes whose real part of the corresponding eigenvalues is the closest (from below) to the real part of  $\beta_5$ . Nevertheless we emphasize that the “wavy” shape of the graph of  $Q_n(\tau, T)$  may experience noticeable changes when  $t_0$  varies. These changes may be manifested by the emergence of local minima that can modify substantially the global minimizer and thus affect the determination of the optimal PM; a sensitivity issue that can be fixed by the calculation of  $c(t)$  given by (3.6); see Remark 8.

Thus, the minimization of the  $Q_n$ 's possibly completed by the analysis of the parameterization correlation,  $c(t)$ , allows us to determine the optimal PM,  $\Phi_{\tau^*}^{(1)}$ , for Eq. (5.9) and  $E_c = \text{span}\{e_1, \dots, e_5\}$ . For our choice of  $t_0$ , the global minima of the  $Q_n$ 's provide the optimal PM. The values of the parameterization defects for this optimal PM are then given by,  $Q_6(\tau_6^*, T) = 9.5 \times 10^{-2}$ ,  $Q_7(\tau_7^*, T) = 2.2 \times 10^{-1}$  and  $Q_8(\tau_8^*, T) = Q_9(\tau_9^*, T) = 1.1 \times 10^{-1}$ . By comparison, for the invariant manifold approximation the parameterization defects (with  $h_{2,n}$  replacing  $\Phi_n$  in (5.14)) are given by  $Q_6(h_2) = 1.8 \times 10^{-1}$ ,  $Q_7(h_2) = 2.2$  and  $Q_8(h_2) = Q_9(h_2) = 8.2 \times 10^{-1}$ . Note that in both cases,  $Q_8 = Q_9$ , since here  $\beta_9 = \bar{\beta}_8$



**Fig. 7**  $Q_n(\tau, T)$  vs  $\tau$  for Eq. (5.1) for  $r = 14.1$  (period-doubling regime) and  $m = 5$ . For each parameterized mode shown here, the minimum is marked by a red dot

(and  $\lambda_9 = \overline{\lambda_8}$ ) and the corresponding parameterizations are just conjugate to each other; see Remark 6.

These values of the parameterization defects should be put in perspective with the energy budget for a better appreciation of the exercise of parameterization conducted here. Table 1 summarizes how the energy is distributed (in average) among the modes, over the training interval  $[0, T]$ . The distribution of energy is explained in part (but not only) by the spectral decomposition and ordering (2.12) adopted here from Sect. 2.1, i.e. the modes are ordered according to their linear rate of growth/decay. In our case, it turns out that Eq. (5.9) is a genuine forced-dissipative system in which the  $\beta_j$ 's have all their real parts negative. Thus the ordering is here from the least to the most stable ones; the least stable modes ( $e_1$  and  $e_2$ ) containing most of the energy.

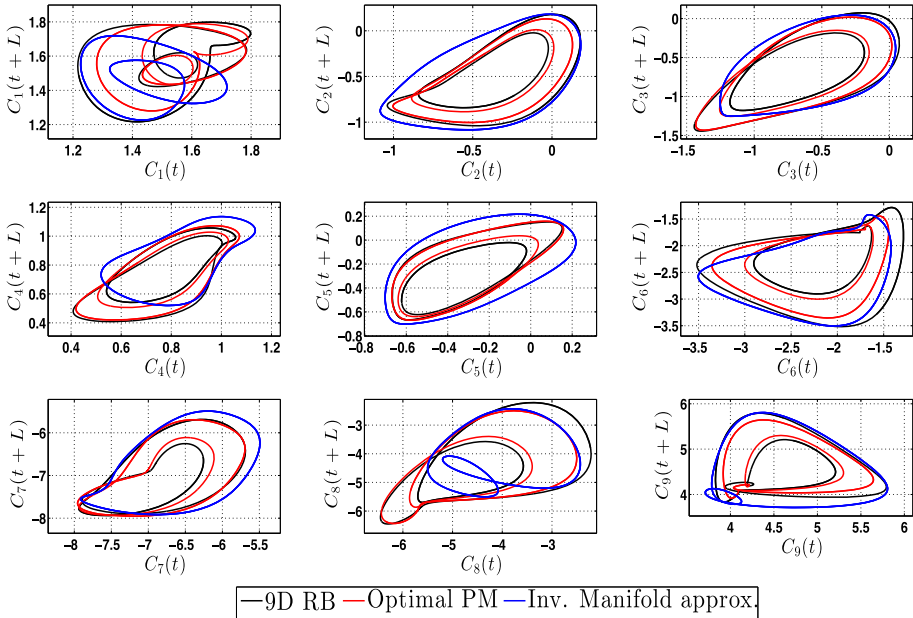
It is noteworthy that it is exactly (and only) for mode  $e_7$ —the mode that contains the smallest fraction of energy—that the parameterization defect  $Q_7(h_2)$  for  $h_2$  is above 1, leading to an over parameterization for this mode. Despite the small fraction of energy contained in a given mode, it is known that an over parameterization of such a mode can lead to an overall misperformance of the associated closure.

In contradistinction,  $Q_7(\tau_7^*, T)$  is of same order of magnitude than the  $Q_n$ 's for modes  $e_6$ ,  $e_8$  and  $e_9$ . As a result, the optimal PM,  $\Phi_{\tau^*}^{(1)}$ , provides comparatively, a much more efficient closure than when the parameterization  $h_2$  is used. Figure 8 shows for instance that in terms of attractor reconstruction, the approximation  $C^{IM}(t)$  given by (5.13) and obtained from the 5D reduced system (5.12) based on  $h_2$  (blue curve), fails—compared to its counterpart  $C^{PM}(t)$  obtained from the 5D optimal PM closure (5.10) (red curve)—in capturing, within the embedded phase space, the intricate behavior of the original model's periodic orbit (black curve).

A closer examination of the power spectral density (PSD) reveals that  $C^{IM}(t)$  fails in reproducing the dominant frequency and its subharmonics, whereas  $C^{PM}(t)$  captures them almost perfectly; compare panel (a) and (b) of Fig. 9. The length of simulation  $T_f$  for the original dynamics and the 5D optimal PM closure (5.10) used for the estimation of these

**Table 1** Averaged fraction of energy over  $[t_0, t_0 + T]$ : period-doubling regime

$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$	$e_9$
42.14%	42.14%	1.81%	3.87%	3.87%	4.27%	0.20%	0.86%	0.86%

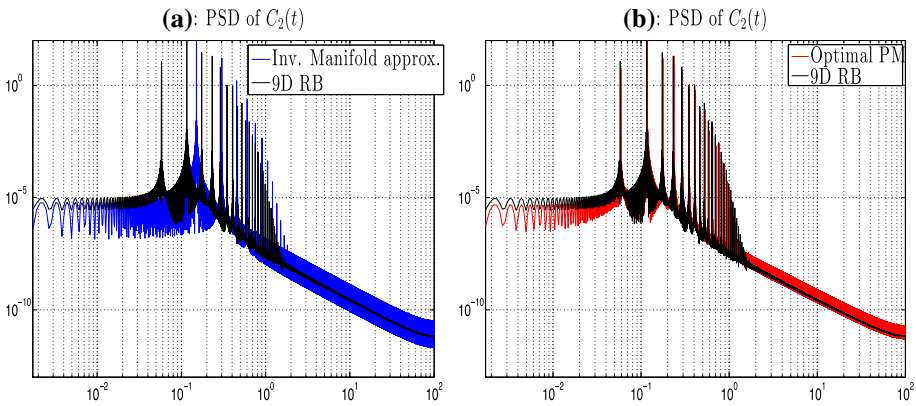


**Fig. 8** Attractor approximation for  $r = 14.1$  and  $m = 5$ . Here the attractor is projected onto the delay coordinates,  $(C_j(t), C_j(t + L))$  ( $1 \leq j \leq 9$ ), for the original 9D RB system (black curve). Here  $L = 1$ . The approximation  $C^{PM}$  given by (5.11) and obtained from the 5D optimal PM closure (5.10) is shown by the red curve. The approximation  $C^{IM}$  given by (5.13) and obtained from the 5D reduced system (5.12) based on the invariant manifold approximation  $h_2$ , is shown by the blue curve (Color figure online)

PSDs is  $T_f = 1000$ . Recall that for the latter, such results are obtained by optimizing the parameterization defects on a training interval of length  $T$  equals only to three fourth of the period  $T_p$  of the original dynamics, demonstrating thus good skills at least in the frequency domain. Similar skills than those shown in Fig. 9 for  $C_2(t)$ , hold for the other system’s components.

As progressing through the period-doubling cascade, the inability of the invariant manifold approximation,  $h_2$ , in reproducing the main features of the RB system’s solutions, is getting even worse, in particular right after the onset of chaos. The next section shows that the reduced systems (5.10), to the contrary, provide still low-dimensional efficient closures (when driven by the appropriate optimal PM) for such chaotic regimes.

**Remark 8** Depending on  $t_0$  (after removal of transient), the global minimizer  $\tau_n^*$  of  $Q_n$ , does not provide necessarily the best parameterization within the  $\Phi_n$ -class, and one may have to rely on the parameterization correlation  $c(t)$  (see (3.6)) to discriminate between other local minimizers of  $Q_n$ . We clarify here this statement which is relevant only for  $n = 6$  here; the global minima of  $Q_7, Q_8$ , and  $Q_9$  being in fact robust as  $t_0$  is varied.



**Fig. 9** PSD approximation for  $r = 14.1$  and  $m = 5$ . Here the PSDs are estimated for  $C_2(t)$  obtained from the original 9D RB system (black curve—panels **a** and **b**, for  $C_2^{PM}(t)$  obtained from the 5D optimal PM closure (5.10) (red curve—panel **b**), and for  $C_2^{IM}(t)$  obtained from the 5D reduced system (5.12) based on invariant manifold approximation (blue curve—panel **a**). A semi-log scale is used for panels **a** and **b** (Color figure online)

For the regime analyzed here, the “wavy” shape of the graph of  $Q_6(\tau, T)$  may experience noticeable changes when  $t_0$  varies. These changes may be manifested by the emergence of local minima that can modify substantially the location of the global minimizer and thus affect the determination of the optimal PM.

For instance the left panel of Fig. 10 shows  $Q_6(\tau, T)$  as obtained from another segment of the solution  $y(t)$  to (5.9) (in the period-doubling regime), that is for another  $t_0$  in (5.14) than used for Fig. 7. A simple visual comparison reveals that the global minimum shown for  $Q_6$  in Fig. 7 corresponds now to a local minimum (red asterisk), and a new global minimum closer to  $\tau = 0$  has appeared (green asterisk).

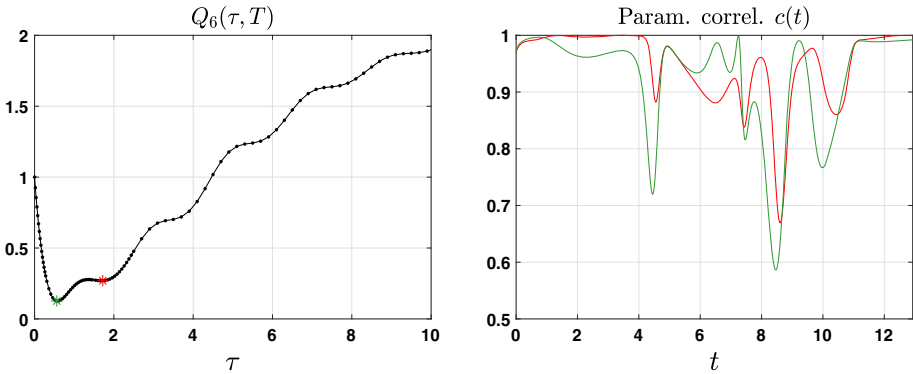
If one selects the corresponding global minimizer as  $\tau_6^*$ , the corresponding optimal closure captures only an excerpt of the dominant frequency and its harmonics (every other frequency more precisely), and the closure fails in reproducing the period-doubling. This issue can be easily fixed by the inspection of  $c(t)$  given by (3.6) over  $[0, T]$ . Indeed, by using the optimal PM for which  $\tau_6^*$  corresponds to the global minimum and the (sub)optimal PM for which  $\tau_6^*$  corresponds to the second local minimum, we obtain two curves for  $c(t)$ : one associated with the optimal parameterization (global minimum/green curve) and one associated with the suboptimal parameterization (local minimum/red curve).

The red curve is clearly closer to 1 than the green one (in average), indicating that  $\tau_6^*$  corresponding to the second local minimum (i.e. the suboptimal parameterization) should be in fact retained for determining the parameterization  $\Phi_n$ , as indeed the corresponding PM closure provides then similar modeling skills to those shown in Fig. 9.

This discrimination, made possible thanks to the parameterization correlation,  $c(t)$ , (prior to any simulation of (5.10)) teaches us the relevance of this non dimensional number to refine the determination of an optimal PM in practice, beyond this example and especially in presence of other local minima for a given  $Q_n$  as  $t_0$  is varied.

Other tests conducted in other parameter regimes indicate that such a situation requiring the discrimination via an inspection of  $c(t)$  and a selection of a suboptimal rather than optimal parameterization is rather the exception than the rule;<sup>8</sup> namely the parameterization

<sup>8</sup> For instance this issue is not encountered for the chaotic regime analyzed in Sect. 5.3.



**Fig. 10** Selection of suboptimal parameterization via parameterization correlation. The parameterization correlation  $c(t)$  are shown in the right panel for an interval of length  $T = 3T_p/4$  in the period-doubling regime. Here  $c(t)$  is computed from (3.6) with  $\Psi = \Phi_\tau^{(1)}$  for two choices of  $\tau$ . Choice 1:  $\tau_n = \tau_n^*$  for all the components (green curve). Choice 2:  $\tau_n = \tau_n^*$  except  $\tau_6$ , which is taken instead to be the local minimizer marked by the red asterisk on the left panel (red curve) (Color figure online)

**Table 2** Averaged fraction of energy over  $[0, T]$ : chaotic regime

$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$	$e_9$
37.59%	37.59%	8.23%	4.90%	4.90%	3.95%	0.31%	1.27%	1.27%

corresponding to a global minimizer of  $Q_n$ , provides in general the best closure results. Nevertheless we decided to communicate on this issue subordinated to the presence of local minima as it may be encountered for other systems.

### 5.3 Closure in a Chaotic Regime

We assess in this section the skills of the optimal PM closure (5.10) in a regime located right after the onset of chaos, after the system has gone through a period doubling cascade, i.e. for  $r = 14.22$ . We conduct also hereafter an analysis on the effect of the reduced dimension,  $m$ , of the reduced state space  $E_c$ . Still this reduced state space is spanned by few dominant eigenmodes of the linear part  $L$  of the perturbed system (5.6) about the mean state  $\bar{C}$  is given by (5.7), with now the latter estimated, after removal of transient dynamics, over the training interval of length  $T = T_p$ , with  $T_p$  denoting the period of the solution for  $r = 14.1$ ; see previous section.

Here again, the unresolved modes are parameterized by the quadratic manifold,  $\Phi_n(\tau, \cdot)$ , given by (4.34), optimized over the training interval  $[0, T]$  by minimizing the parameterization defect  $Q_n$  given by (5.14). The distribution of energy per mode for this regime is shown in Table 2. The distribution of energy is explained due to the ordering (2.12) adopted here from Sect. 2.1, i.e. by ordering the modes according to their linear rate of growth/decay; for this parameter regime again, from the least to the most stable modes. Since  $e_4$  and  $e_5$  come in pairs (i.e.  $\text{Re}(\beta_4) = \text{Re}(\beta_5)$ ), we analyze hereafter the cases  $m = 3$ ,  $m = 5$  and  $m = 6$ . Thus from Table 2, the energy to be parameterized corresponds to 16.6% of the total energy (over  $[0, T]$ ) for the case  $m = 3$ , to 6.8% for  $m = 5$ , and to 2.85% for  $m = 6$ .

**Table 3** Optimal parameterization defects for  $T = 25$ : chaotic regime

	$m = 3$	$m = 5$	$m = 6$
$Q_4(\tau_4^*, T)$	0.09		
$Q_5(\tau_5^*, T)$	0.09		
$Q_6(\tau_6^*, T)$	0.38	0.12	
$Q_7(\tau_7^*, T)$	0.22	0.2	0.04
$Q_8(\tau_8^*, T)$	0.05	0.09	0.02
$Q_9(\tau_9^*, T)$	0.05	0.09	0.02

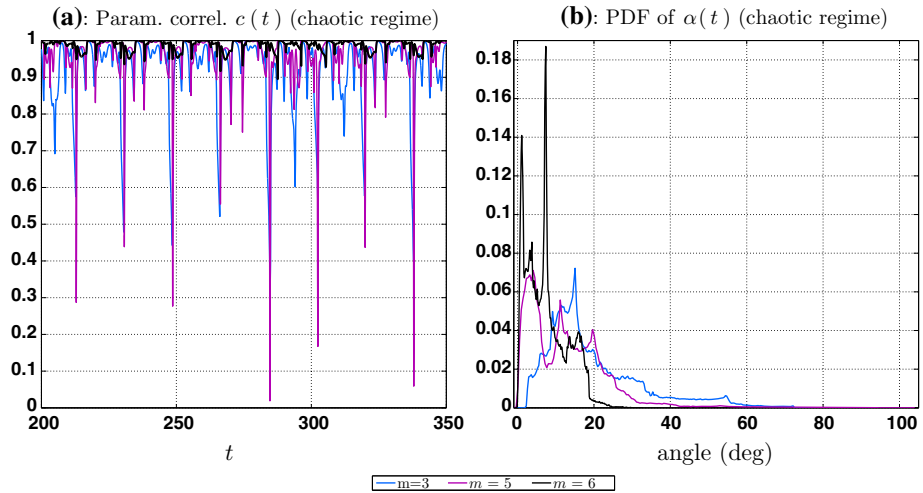
Given the solution  $y(t)$  of Eq. (5.9) over  $[0, T]$ , the minimal values  $Q_n(\tau_n^*, T)$  achieved by the optimal PM,  $\Phi_{\tau^*}^{(1)}$ , in terms of the reduced dimension  $m$  are shown in Table 3. Obviously, the case  $m = 6$  comes with the smaller parameterization defects, while the case  $m = 3$  presents for the modes  $e_6$  and  $e_7$ , values that although less than 1 are not on the same order of magnitude than the other values of  $Q_n$ .

The energy left after application of the optimal PM, represents  $0.04 \times 0.31 + 2 \times 0.02 \times 1.27 = 0.063\%$  of the total energy for the case  $m = 6$ , and represents  $0.765\%$  for the case  $m = 5$ , still below  $1\%$  of the total energy. To the contrary, an amount of energy representing  $5.42\%$  needs still to be parameterized after application of the optimal PM for the case  $m = 3$ . Compared with the fraction of energy left in the corresponding unresolved modes prior parameterization, an application of the optimal PM leads to an improvement by a factor approximately equal to 45 for  $m = 6$ , and equal to 9 and to 3 for respectively  $m = 5$  and  $m = 3$ . Without any surprise, the cutoff corresponding to the smallest amount of energy to be parameterized (i.e. when  $m = 6$ ) comes with the best improvement in terms of parameterization when the optimal PM is used. On the other hand, the cutoff corresponding to the biggest amount of energy (i.e. when  $m = 3$ ) comes with the poorest parameterization score in terms of energy that still needs to be parameterized after application of the optimal PM. Thus, one expects that an optimal PM closure should perform certainly better for  $m = 6$  than for  $m = 3$ , and must show some improvements compared to the optimal PM closure for  $m = 5$ .

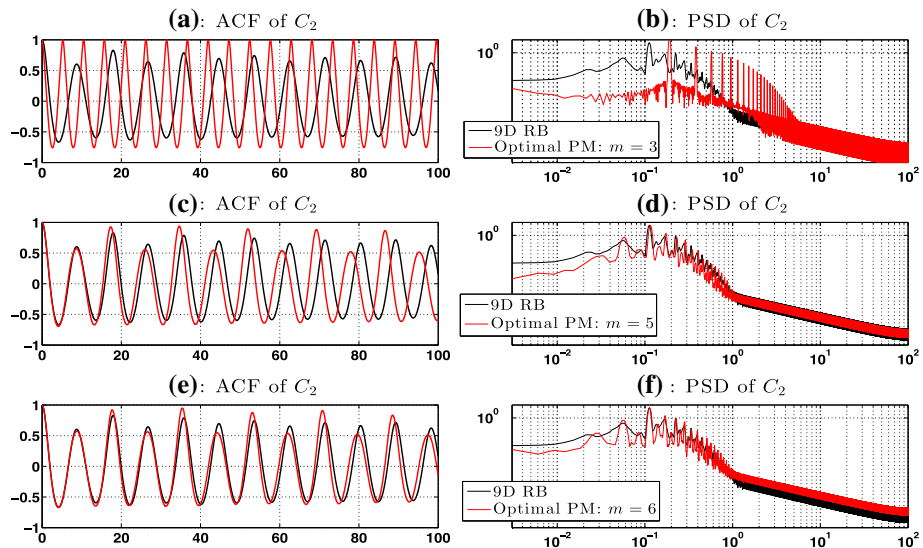
This energy budget analysis is comforted by the analysis of the parameterization correlation  $c(t)$  and of the probability density function (PDF) of the parameterization angle  $\alpha(t)$ . Here  $c(t)$  and  $\alpha(t)$  are respectively computed from (3.6) and (3.7), with  $\Psi = \Phi_{\tau^*}^{(1)}$ , the optimal PM as determined for each case,  $m = 3$ ,  $m = 5$ , and  $m = 6$ , from (4.36), for which the optimal vector  $\tau^*$  is obtained by minimization of (5.14) for the relevant  $n$ . As shown in panel (b) of Fig. 11, each of these PDFs is skewed towards zero. Nevertheless the PDF that is the most concentrated (i.e. with more mass) near zero corresponds to the case  $m = 6$  (black curve), then comes the PDF associated with the case  $m = 5$  (magenta curve), and finally the PDF for the case  $m = 3$  (blue curve).

These diagnostics are confirmed when looking at the ability of the corresponding optimal PM closures (5.10), in reproducing key statistics of the original model’s dynamics such as autocorrelation functions (ACFs) and PSDs. For the regime analyzed here ( $r = 14.22$ ), the time-variability of the chaotic dynamics is characterized by a broad band spectrum visible in each component’s PSD. The black curve in either right panels of Fig. 12, shows such a broad band spectrum for e.g. the PSD of  $C_2$  as estimated from integration of Eq. (5.1) after a simulation of length  $T_f = 1000$ . Other components display similar PSDs.



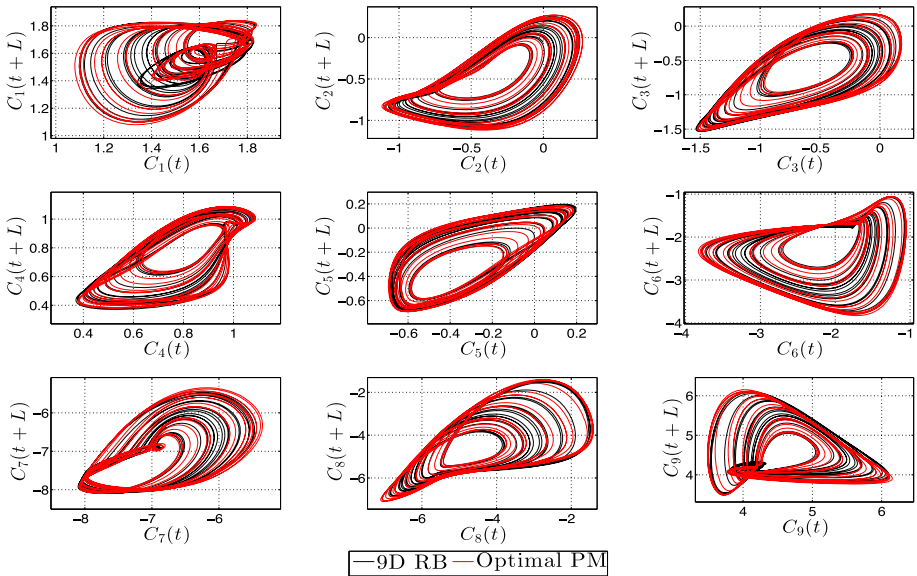


**Fig. 11** Effect of the reduced dimension  $m$ : diagnostic for  $r = 14.22$ . This effect is shown here on the parameterization correlation  $c(t)$  (panel **a**) and the PDF of the parameterization angle  $\alpha(t)$  (panel **b**) for the chaotic regime. Here  $c(t)$  and  $\alpha(t)$  are respectively computed from (3.6) and (3.7), with  $\Psi = \phi_{\tau^*}^{(1)}$ , the optimal PM



**Fig. 12** Effect of the reduced dimension  $m$ : simulation for  $r = 14.22$ . This effect is shown for the chaotic regime on the ability of the optimal PM closure (5.10) to reproduce the PSD and ACF, for the second component  $C_2$ . A semi-log scale is used for panels **b**, **d** and **f**

Figure 12 shows clearly, as anticipated by the energy budget analysis on a short interval  $[0, T]$  (with  $T = 17.25$ ) (and supported by the parameterization angle’s PDF analysis), that the 5D and 6D optimal PMs provide efficient closures, with a noticeable improvement for the ACF’s reproduction of  $C_2$  when the 6D optimal PM is used; see panel (e) of Fig. 12. Furthermore, Fig. 13 shows that the 6D optimal PM closure leads to an excellent approximation



**Fig. 13** Attractor approximation for  $r = 14.22$  and  $m = 6$ . Same as in Fig. 8 except  $r = 14.22$  (chaotic regime) and  $m = 6$ . Here also  $L = 1$

of the original model’s attractor, whereas the 5D optimal PM closure although reproducing correctly most of its features fails in reproducing certain solution’s large excursions in the embedded phase space (not shown). The 3D optimal PM fails however dramatically in the approximation of this attractor as it leads to a periodic orbit and fails thus to reproduce the time variability of the original model’s chaotic dynamics; see panels (a) and (b) of Fig. 12.

Based on these results, we may state that our parameterization formula of Sect. 4.3 (i.e.  $\Phi_{\tau^*}^{(1)}$  given by (4.36)) provides here, seemingly, a good approximation of the optimal PM as given by the abstract Theorem 4 when  $m = 5$  and  $m = 6$ . Our optimal PM as computed for the case  $m = 3$ , although leading to a periodic orbit, may still be a good approximation of the theoretical optimal parameterization (3.26) averaging out the unresolved variables, for the reduced state space,  $E_c = \text{span}\{e_1, e_2, e_3\}$ . It is indeed possible that the conditional expectation as defined in Theorem 5, gives a periodic solution for a given reduced state space. The theory of Sect. 3 does not exclude such a scenario.

To improve the results in the case  $m = 3$ , stochastic parameterizations may be then superimposed to our optimal PM in order to further reduce the parameterization defect. This topic is out of the scope of the present paper but will be pursued elsewhere; see Concluding Remarks in Sect. 7.

### 5.4 Heat Flux Analysis

We analyze here how the optimal LIA parameterization behaves in the physical domain, for the chaotic regime. We focus on the vertical heat flux, accomplished by the fluctuations around the time-averaged state that enables the system to sustain statistical equilibrium. Once a solution  $C(t)$  to Eq. (5.1) is computed, one can evaluate the following local heat flux

$$H(\mathbf{x}, t) = w(\mathbf{x}, t)\theta'(\mathbf{x}, t) - \partial_z \bar{\theta}(\mathbf{x}), \quad \mathbf{x} = (x, y, z), \tag{5.15}$$

**Table 4** Heat fluxes: relative error when “s” is replaced by optimal PM

	$m = 5$ (%)	$m = 6$ (%)
$\langle H \rangle$	15	4.5
$\langle H_{c_s} \rangle$	7.6	11.2
$\langle H_{s_s} \rangle$	64	21.9

where  $w$  denotes the vertical velocity, and  $\theta'$  denotes the anomaly of the temperature  $\theta$  with respect to the time-mean temperature  $\bar{\theta}$ . The vertical velocity  $w$  and temperature  $\theta$  are computed according to Eqns. (12) and (17) of [145].

Recall that our optimal PM is determined for the transformed variables, namely for Eq. (5.9). In particular our splitting between low and high modes is made within the system of coordinates in the  $y$ -variable. By transforming back into the original variables we can trace the contribution of the high and low modes (defined in the transformed variables) into the original system of coordinates. By doing so, the heat flux  $H(\mathbf{x}, t)$  decomposes as

$$H(\mathbf{x}, t) = H_{cc}(\mathbf{x}, t) + H_{cs}(\mathbf{x}, t) + H_{ss}(\mathbf{x}, t). \tag{5.16}$$

with

$$\begin{aligned} H_{cc}(\mathbf{x}, t) &= w_c(\mathbf{x}, t)\theta'_c(\mathbf{x}, t) - \partial_z \bar{\theta}_c(\mathbf{x}), \\ H_{ss}(\mathbf{x}, t) &= w_s(\mathbf{x}, t)\theta'_s(\mathbf{x}, t) - \partial_z \bar{\theta}_s(\mathbf{x}), \\ H_{cs}(\mathbf{x}, t) &= w_c(\mathbf{x}, t)\theta'_s(\mathbf{x}, t) + w_s(\mathbf{x}, t)\theta'_c(\mathbf{x}, t). \end{aligned} \tag{5.17}$$

When the high-mode contribution in (5.16) and (5.17) is replaced by the optimal LIA parameterization derived in the previous section (chaotic regime), errors in the “low-high” and “high-high” interactions to the heat flux are visible. Table 4 shows these relative errors in the  $L^2$ -norm in time, after space average  $\langle \cdot \rangle$ . Clearly these errors reduce as the dimension of the reduced state space (in the transformed variables) increases, but overall the reproduction of the time-variability of  $\langle H \rangle$  is satisfactory, especially when  $m = 6$ ; see Figs. 14 and 15. As a comparison when only the low modes are used to approximate the heat flux like in a Galerkin truncation, the heat flux errors are substantially larger; see Table 5. Without any surprise the improvement brought by the high-mode parameterization is more pronounced when  $m = 5$  than when  $m = 6$ . Taking volume- and time-average in (5.16), we observe that  $\overline{\langle H \rangle} = 54.6$ . Doing the same operation in which the  $s$ -variable is replaced by its high-mode approximation (as given by the optimal LIA) gives  $\overline{\langle H^{\text{app}} \rangle} = 61.4$  for  $m = 5$ , and  $\overline{\langle H^{\text{app}} \rangle} = 56.1$ , for  $m = 6$ .

## 6 Closing Kuramoto–Sivashinsky Turbulence and Fixing Backscatter Errors

In this section we show that the PM approach allows for deriving efficient closures for the Kuramoto-Sivashinsky (KS) turbulence, in strongly turbulent regimes. The closure results presented hereafter are obtained for cutoff scales placed well within the inertial range, keeping only the unstable modes in the reduced state space. The underlying optimal PMs obtained by our variational approach are far from slaving and allow for remedying the excessive backscatter transfer of energy to the low modes encountered by the LIA or the QSA parameterizations in their standard forms, when they are used at this cutoff wavelength.

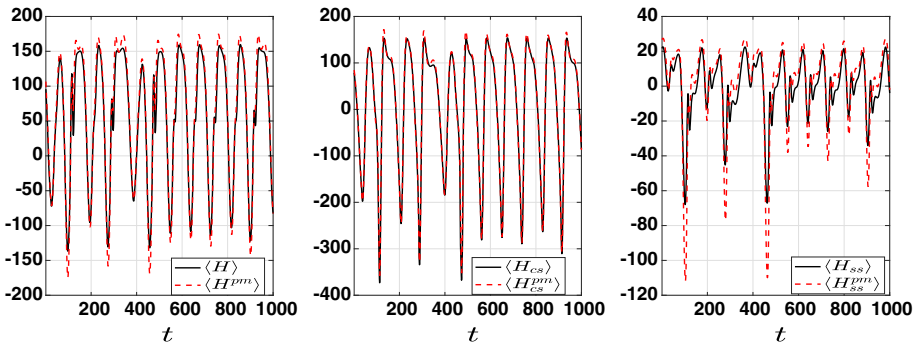


Fig. 14 Space-average heat fluxes for the chaotic regime. Here the reduced state space is five-dimensional ( $m = 5$ )

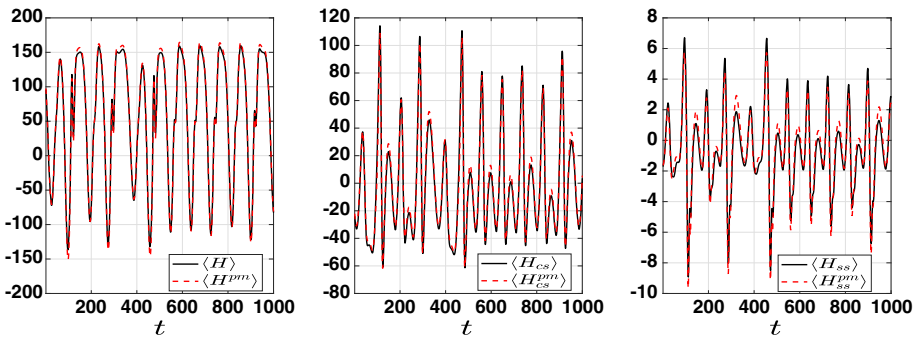


Fig. 15 Space-average heat fluxes for the chaotic regime. Here the reduced state space is six-dimensional ( $m = 6$ )

Table 5 Relative error

$$\mathcal{E}_c = | \langle H - H_{cc} \rangle |_{L^2} / | \langle H \rangle |_{L^2}$$

	$m = 5$	$m = 6$
$\mathcal{E}_c$	132%	35%

### 6.1 Preliminaries and Background

We consider the KS equation (KSE) [111,157] posed on the domain,  $\mathcal{D} = (0, L)$ , and subject to periodic boundary conditions:

$$\partial_t u = -v \partial_x^4 u - D \partial_x^2 u - \gamma u \partial_x u, \tag{6.1}$$

where  $v, D$  and  $\gamma$  are positive parameters. The KSE is commonly considered as a basic case study for spatio-temporal chaos.

Note that the KSE in its formulation (6.1) can be rescaled as posed on the interval  $(0, 2\pi)$ :

$$\partial_{\bar{t}} \bar{u} = -4 \partial_{\bar{x}}^4 \bar{u} - \alpha \left( \partial_{\bar{x}}^2 \bar{u} + \bar{u} \partial_{\bar{x}} \bar{u} \right), \tag{6.2}$$

by using the following scaling

$$L = \sqrt{\frac{v\alpha}{D}} \pi, \quad u = \frac{2D^{3/2}}{\gamma \sqrt{v\alpha}} \bar{u}, \quad x = \frac{\sqrt{v\alpha}}{2\sqrt{D}} \bar{x}, \quad t = \frac{v\alpha^2}{4D^2} \bar{t}. \tag{6.3}$$

Although mathematically equivalent, depending on the purpose one may prefer one formulation to the other for the closure exercises considered hereafter; see Remark 9.

We aim at closure of the KSE. Various purposes are pursued regarding what a low-dimensional closure should do and this may cause confusion when comparing methods. Among the purposes targeted in the literature concerning the closure/reduction problem of the KSE, are the following: (i) finite-time approximation error such as in AIM theory [52,131] or renormalization group (RG) methods [156], (ii) reproduction of local and global bifurcations [2,15,96,97], (iii) optimal prediction of resolved variables [158], and (iv) reproduction of long-term statistics such as the energy spectrum. We follow clearly this latter path, to which we add the question of reproduction by closure of patterns and their statistical features. For the KSE, only few works have addressed the closure in the latter sense. We refer to [124] for closure aimed at reproducing long-term statistics and to [158] for optimal prediction. In all these works, the regimes for which an efficient closure is sought correspond either to specific solutions or to weakly turbulent regimes associated with a few pairs of unstable modes: 2 pairs in [2], up to 4 pairs of unstable modes for [15,96,97], and 3 pairs in [124,158].

In this study, we aim at determining efficient closures for the reproduction of patterns and long-term statistics in two strongly turbulent regimes: one regime corresponding to 31 pairs (Regime A, Table 6) of unstable modes and another one corresponding to 90 pairs of unstable modes (Regime B, Table 7). Our approach relies on optimal PMs that allow for approximating the conditional expectation (Theorem 5) without assuming separation of scales and differ in that sense from averaging techniques and other RG methods.

The reproduction of the energy spectrum of KS solutions will be one of the core metrics to assess the quality of our parameterizations. For either formulation (6.1) or (6.2), a typical energy spectrum,  $E(k)$ , of a chaotic KS solution is shown as the black curve in panel (e) of Fig. 16. Four parts of this spectrum are distinguishable [174]: (i) The *large scale region* as  $k \rightarrow 0$  which is characterized by a plateau reminiscent of a thermodynamic regime with equipartition of energy; (ii) the *active scale region* that contains most of the energy, with a peak corresponding to a characteristic length  $l_p = L/(2\pi k_p)$  with  $k_p$  that corresponds to the wavenumber of the most linearly unstable mode; (iii) a power law decay with an exponent experimentally indistinguishable from 4 within this active region; and (iv) an exponential tail due to the strong dissipation at small scales. It is tempting to think of the region  $E(k) \sim k^{-4}$ , where production and dissipation are almost balanced ( $Dk^2 \approx \nu k^4$ ), as an “inertial range.” This latter aspect has been already discussed in the literature; see [141].

From a mathematical perspective, the KSE is a well-known example of PDE that possesses an inertial manifold, in the invariant space of odd functions [38,65], and in the general periodic case [149,165], but the current IM theory [180] predicts that the underlying slaving of the high modes to the low modes, holds when the cutoff wavenumber,  $k_c$ , is taken sufficiently far within the dissipative range, especially in “strongly” turbulent regimes that correspond to the presence of many unstable modes; see the Supplementary Material. Still, as the AIM theory underlines, satisfactory closure may be expected to be derived for  $k_c$  corresponding to scales larger than what predicts the IM theory. Nevertheless, as one seeks to further decrease  $k_c$  within the inertial range, standard AIMS fail typically in providing relevant closures and one needs to rely on no longer a fixed cutoff but instead a dynamic one so as to avoid energy accumulation on the cutoff level [50,54,56]. This situation has been already documented for the Navier–Stokes equations [137], but is less known for the KSE.

As pointed out below, such a failure by traditional (nonlinear) parameterizations for closing the KSE when  $k_c$  is placed low within the inertial range occurs e.g. for Regime A considered

hereafter and whose parameters<sup>9</sup> are listed in Table 6. For this regime, the KS flow is strongly turbulent (see Fig. 16b) and possesses 31 pairs of unstable modes. We selected  $k_c$  to be the wavenumber corresponding to the smallest scale present among the unstable modes, corresponding here to  $k_c = 31$  for Regime A, and making thus the reduced state space,  $E_c$ , to be spanned by the unstable modes. This choice of  $k_c$  places the cutoff wavelength within the aforementioned inertial range, as one can observe in Fig. 16d. The fraction of energy to parameterize is quite substantial for this cutoff as it represents 15.7% of the total energy. For this selection of  $k_c$ , the energy distribution nearby this cutoff scale is comparable to the energy  $E(k)$  contained in the large scales ( $k \sim 1$ ). Beyond  $k_c$ , the energy does not drop suddenly (due to its decay following a power law) and actually takes values on a same order of magnitude compared to  $E(1)$  for roughly  $k_c < k < 1.5k_c$  while only after  $k > k_1 = 2k_c$ , the energy  $E(k)$  drops faster (exponentially); see black curve Fig. 16e.

Thus to close the KSE at this cutoff scale, makes, a priori, the closure problem difficult because quite a few energetic modes need to be properly parameterized. Actually, as discussed in Sect. 6.2 below, this difficulty is manifested when using nonlinear parameterizations such as the standard QSA (4.40) that suffers from a backscattering transfer of energy particularly overwhelming for the large scales. In this case an over-parameterization of the neglected scales (i.e. an excessive parameterization of the unresolved energy) leads to an incorrect reproduction of the backscatter transfer of energy due to nonlinear interactions between the modes, especially those near the cutoff scale. We speak of an inverse error cascade, i.e. errors in the modeling of the parameterized scales that contaminate gradually the larger scales and spoil the closure skills for the resolved variables.

To illustrate such an inverse error cascade in a simple context, we invite the reader to consult the AB-system in the Supplementary Material; see Eq. (17) therein. For this system, let us assume that an error of size  $\epsilon \bar{B}$  is made on the parameterized variable  $\bar{B}$  at the steady state  $(\bar{A}, \bar{B})$  given by (18) in the Supplementary Material. This error propagates then to the resolved variable  $\bar{A}$  through nonlinear coupling as  $\bar{A}_{\text{app}} = \sqrt{(\nu_2 \bar{B}_{\text{app}} - \alpha \bar{B}_{\text{app}}^3) / \gamma_2}$  where  $\bar{B}_{\text{app}} = (1 \pm \epsilon) \bar{B}$ . The  $(L^2)$  error on the resolved variable becomes then  $|\bar{A}^2 - \bar{A}_{\text{app}}^2|$ : of order  $\epsilon$  when  $\epsilon$  is small, and of order  $\epsilon^3$  when  $\epsilon$  is large. This simple example shows that an error made on the parameterization may be amplified through the nonlinear interactions as it propagates to the resolved variables when the parameterization is not accurate. Such an inverse error cascade is even more pronounced as the number of nonlinear interaction terms gets large while the neglected scales contain a non-negligible amount of energy. In that respect, the parameter regimes considered here for the KSE are particularly demanding to avoid an incorrect reproduction of the backscatter transfer of energy to the large scale.

Our purpose is to show that the parametric QSA formulas (4.42)–(4.44) of Sect. 4.4, when optimized by solving the minimization problems (4.46), allow for fixing the backscatter transfer of energy issue encountered by the standard QSA (4.40). As shown hereafter, the amount of data required to determine the underlying optimal PMs (here given as optimal QSAs), is related to mixing properties such as encoded into decay of temporal correlations. Typically, the faster the decay of (temporal) correlations is, the less the amount of data (in the time direction) required, is. The PM approach and its apparatus provides furthermore new understanding about essential variables and their interactions for closure of the KSE.

<sup>9</sup> These parameters become  $\alpha = 4000$ ,  $\bar{\delta}t = 10^{-7}$  and  $N_x = 256$  when scaling (6.3) is applied; see Remark 9.

**Table 6** Regime A: Parameters for Eq. (6.1)

$\nu$	$D$	$L$	$\gamma$	$\delta t$	$N_x$
$2 \times 10^{-4}$	0.2	$2\pi$	1	$10^{-3}$	256

**Table 7** Regime B: Parameters for Eq. (6.2)

$\alpha$	$\delta t$	$N_x$
33,000	$10^{-9}$	2048

To apply the PM approach and the parameterization formulas of Sect. 4.4 to Eq. (6.1) we first recall the spectral elements of the operator  $A = -\nu\partial_x^4 - D\partial_x^2$ , under periodic boundary conditions. These are given by

$$\beta_k = -\frac{16\nu\pi^4 k^4}{L^4} + \frac{4D\pi^2 k^2}{L^2}, \tag{6.4}$$

for the eigenvalues, and

$$e_k^\ell(x) = \begin{cases} \sqrt{\frac{2}{L}} \cos\left(\frac{2\pi kx}{L}\right), & \text{if } \ell = 0 \\ \sqrt{\frac{2}{L}} \sin\left(\frac{2\pi kx}{L}\right), & \text{if } \ell = 1, \end{cases} \tag{6.5}$$

for the eigenmodes. Note that because the spatial average of our KS-solutions considered hereafter is zero (see (6.10)), we consider  $k \geq 1$  in what follows.

Adopting the convention of Sect. 2.1, and after having reordered the  $\beta_k$ 's in descending order, the reduced state space is

$$E_c = \text{span}\{e_{p(1)}^\ell, \dots, e_{p(m)}^\ell, \ell = 0, 1\}, \tag{6.6}$$

where  $p(j)$  denotes the wavenumber of the cosine/sine pair associated with the  $j^{\text{th}}$  largest eigenvalue. Note that due to the distribution of the  $\beta_k$ 's given by (6.4), this reordering matters only when  $m < m_u$  with  $m_u$  denoting the total number of pairs of unstable modes.

The projector  $\Pi_c$  onto  $E_c$  is then given by

$$\Pi_c u = \sum_{\ell=0}^1 \sum_{j=1}^m \langle u, e_{p(j)}^\ell \rangle e_{p(j)}^\ell. \tag{6.7}$$

Hereafter we will consider closure for  $m \geq m_u$ . In this case, the reduced state space is simply given by

$$E_c = \text{span}\{e_1^\ell, \dots, e_m^\ell, \ell = 0, 1\}. \tag{6.8}$$

Here the ambient space is taken to be the Hilbert space  $\mathcal{H} = L^2(0, L)$ , and  $\langle \cdot, \cdot \rangle$  denotes its natural inner product. Hereafter we denote by  $\Pi_s$  the orthogonal complement of  $\Pi_c$  in  $\mathcal{H}$ , i.e.  $\Pi_s = \text{Id}_{\mathcal{H}} - \Pi_c$ .

Another regime that will be dealt with in Sect. 6.3 below has its parameters listed in Table 7 for the KSE written under its formulation (6.2). This regime is even more turbulent than Regime A, as it exhibits 90 pairs of unstable modes. Either for Regime A or B, the benchmark KS solution for the closure exercises conducted hereafter, is obtained by transforming the KSE in Fourier space and by using a modification of the exponential time-differencing fourth-order Runge-Kutta (ETDRK4) method proposed in [99] in order to solve the resulting stiff

ODE system. The number of Fourier modes retained ( $N_x$ ) and time step used ( $\delta t$ ) for each regime, are listed in Tables 6 and 7, for Regimes A and B, respectively. We refer hereafter to a KS solution thus obtained as a Direct Numerical Solution (DNS). The ODE closure derived hereafter are integrated with an semi-implicit Euler scheme, in which the linear terms are treated implicitly while the nonlinear ones, explicitly. These closure systems are integrated with the same time step as listed in Tables 6 and 7, depending on the regime.

In all our numerical experiments that follow, the KSE is integrated from the following initial datum with zero-mean

$$u_0(x) = \cos(x)(1 + \sin(x)). \quad (6.9)$$

In such a case, since the spatial average is a conserved quantity for the KS solution  $u(x, t)$ , we have for all  $t$ ,

$$\int_0^L u(x, t) dx = 0. \quad (6.10)$$

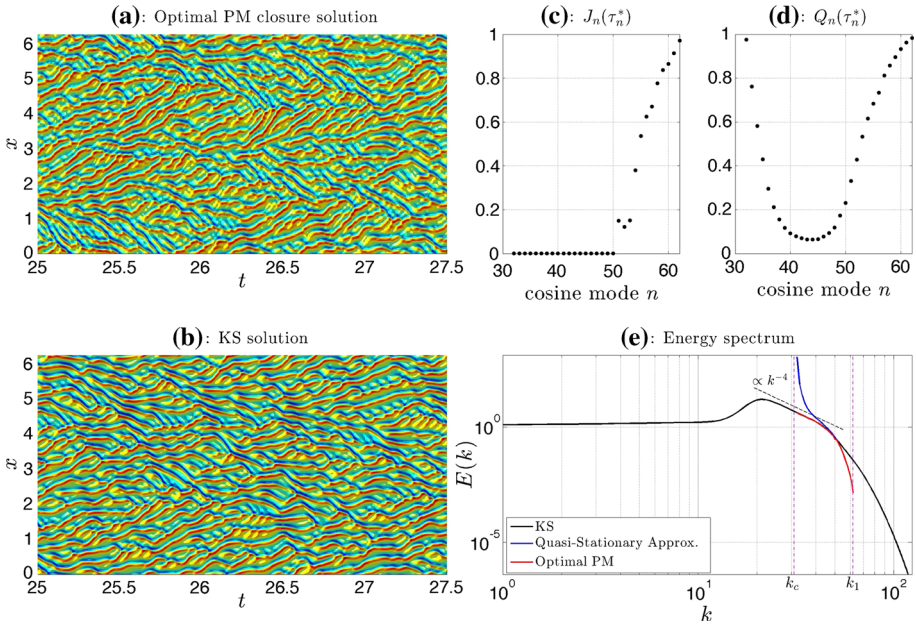
Note that compared with the original ETDRK4 proposed in [41], the modification in [99] consists of evaluating key coefficients as given by [99, Eq. (2.5)] using contour integrals rather than direct evaluation to avoid possible cancellation errors. The contours are taken to be circles of radius  $\delta t$  centered around each of the eigenvalues of the discretized linear operator, and the contour integrals are approximated using trapezoid rules with  $M$  equally spaced points on the circle. We have set  $M = 64$  for both parameter regimes considered. In our numerical calculations performed in Matlab (version R2018a), compared to the script given in [99, Fig. 7], the spatial discretization is taken to be  $x = L^* (0:Nx-1)'/Nx$  instead of  $x = L^* (1:Nx)'/Nx$  to suit the way the fast Fourier transform (FFT) is implemented in the Matlab built-in function `fft`.

**Remark 9** When the scaling (6.3) is performed, we find for Regime A that  $\alpha = 4000$  and  $\bar{t} = \theta t$  with  $\theta = 5 \times 10^{-5}$ . After transient is removed, to reach the same energy level,  $\|u\|_{L^2}^2$  than by integrating (6.1) (with the same solver), we have found that we can decrease the time-step compared to  $\delta t$  by a factor  $a = 10^4$ , that is  $\bar{\delta t} = 10^{-7}$ . Given an interval of length  $T$  in the original time variable  $t$ , it corresponds to  $\bar{T} = 5 \times 10^{-5} T$ , that is an amount of data in time that is given by  $\bar{N} = \bar{T}/\bar{\delta t} = 500T$  data points. Thus, since  $N = T/\delta t = 1000T$ , we have that  $\bar{N} = N/2$ . Although mathematically equivalent, we can thus store twice more data (while keeping  $N_x$  identical) by integrating numerically the formulation (6.2) than by integrating the formulation (6.1), integrating the dynamics up to the same time instant (taking into account the rescaling). Such observations have their interest to draw statistics from long time integration. For Regime A it turns out that the simulations performed hereafter were already sufficient to draw robust statistics with the formulation (6.1). We use however formulation (6.2) to simulate the turbulent Regime B with a higher number of unstable modes than for Regime A.

## 6.2 Fixing the Backscatter Transfer of Energy for KS Turbulence with Optimal PMs

It is known that when the cutoff wavelength is too low within the inertial range, the standard QSA (4.40) suffers typically from over-parameterization leading to an incorrect backscatter transfer of energy, i.e. errors in the modeling of the parameterized (small) scales that contaminate gradually the larger scales. In the case of Regime A, when  $k_c = 31$  (corresponding to  $E_c$  spanned by 31 pairs of unstable modes), the QSA leads to an over parameterization of  $E(k)$  by an amount of about 5800% (in average) over the wavenumbers  $32 \leq k \leq 36$ ;





**Fig. 16** Closure and parameterization results Regime A. Panel **a** shows the solution obtained from the optimal PM closure (6.23) with  $m = 31$ , while panel **b** shows the KS solution as obtained from DNS of Eq. (6.1). Here the optimal PM is obtained as QSA( $\tau^*$ ) with  $\tau^*$  obtained by optimization of the cost functional  $J_n$  given by (6.19) (with  $t = 1$  and  $T = 4$ ). The optimal values  $J_n(\tau_n^*)$  are shown in panels **c** for the parameterized cosine modes. The corresponding  $Q_n$ -values are shown in panel **d**, with  $Q_n$  given by (6.20). The resulting optimal QSA parameterizes the wavelength band,  $k_c < k < k_1 = 2k_c$ , as shown by the red curve in panel **e** on the energy spectrum  $E(k)$  (log–log scale). Here  $k_c$  is the wavenumber corresponding to the smallest scale present among the unstable modes, that is  $k_c = 31$ . The blue curve shows the dramatic failure of the standard quasi-stationary approximation (QSA) (4.40) for parameterizing this wavelength band, especially for  $k$  near  $k_c$

see blue curve in Fig. 16e. The nonlinear interactions between these modes and the unstable modes corresponding to  $k \leq k_c$  lead in this case to such an excessive backscatter transfer of energy, that a closure in which the unresolved modes are approximated by the QSA, blows up after few iterations no matter the numerical scheme used.

As pointed out in Sect. 4.4, the parametric QSA formulas (4.42)–(4.44) involve the same interaction coefficients, the  $B_{ij}^n$ 's given by (4.29) as for the standard QSA,  $K(\xi)$ . However the magnitudes of the nonlinear interactions, as encapsulated in the coefficients  $\delta_n(\tau)$ 's given by (4.43), is different from the coefficients  $-\beta_n^{-1}$  appearing in  $K(\xi)$ . The coefficients  $\delta_n(\tau)$ 's enable us here to counterbalance the excess of energy in the parameterization compared to a standard QSA. Furthermore, as explained below, these coefficients are optimized in the  $\tau$ -variable by solving the minimization problems (4.46) over short training periods of length comparable to a characteristic decorrelation time of the dynamics.

In the case of the KSE, the parametric QSA (4.44),  $QSA(\tau)$ , takes the following form

$$\Psi_\tau(\xi) = \sum_{\ell=0}^1 \sum_{n=m+1}^{2m} \Psi_n^\ell(\tau_n, \beta, \xi) e_n^\ell, \tag{6.11}$$

with

$$\Psi_n^\ell(\tau_n^\ell, \boldsymbol{\beta}, \xi) = \sum_{i,j=1}^m \delta_n(\tau_n^\ell) \left( E_{ij}^{n,\ell} \xi_i^0 \xi_j^0 + C_{ij}^{n,\ell} \xi_i^0 \xi_j^1 + F_{ij}^{n,\ell} \xi_i^1 \xi_j^1 \right), \quad \xi \in E_c. \quad (6.12)$$

The index  $m$  in the upper bound of the sum is taken here to be equal to  $k_c = 31$ , which corresponds to the number of pairs of unstable modes for Regime A. The reduced state space  $E_c$  is thus  $2m$ -dimensional, taking into account  $\ell = 0, 1$ .

In (6.12),  $\delta_n(\tau_n^\ell)$  is given by (4.43) while

$$E_{ij}^{n,\ell} = \begin{cases} \langle B(\mathbf{e}_i^0, \mathbf{e}_j^0), \mathbf{e}_n^0 \rangle, & \text{if } \ell = 0 \\ \langle B(\mathbf{e}_i^0, \mathbf{e}_j^0), \mathbf{e}_n^1 \rangle, & \text{if } \ell = 1, \end{cases} \quad (6.13)$$

$$C_{ij}^{n,\ell} = \begin{cases} \langle B(\mathbf{e}_i^0, \mathbf{e}_j^1), \mathbf{e}_n^0 \rangle + \langle B(\mathbf{e}_j^1, \mathbf{e}_i^0), \mathbf{e}_n^0 \rangle & \text{if } \ell = 0 \\ \langle B(\mathbf{e}_i^0, \mathbf{e}_j^1), \mathbf{e}_n^1 \rangle + \langle B(\mathbf{e}_j^1, \mathbf{e}_i^0), \mathbf{e}_n^1 \rangle & \text{if } \ell = 1, \end{cases} \quad (6.14)$$

and

$$F_{ij}^{n,\ell} = \begin{cases} \langle B(\mathbf{e}_i^1, \mathbf{e}_j^1), \mathbf{e}_n^0 \rangle & \text{if } \ell = 0 \\ \langle B(\mathbf{e}_i^1, \mathbf{e}_j^1), \mathbf{e}_n^1 \rangle & \text{if } \ell = 1. \end{cases} \quad (6.15)$$

These coefficients correspond to the aforementioned interaction coefficients. They possess a simple analytic expression here given the nonlinearity and the trigonometric eigenfunctions. In particular, a majority of these coefficients are actually zero for  $m + 1 \leq n \leq 2m$ , leaving only a few of them non-zero.

More precisely, we have

$$\langle B(\mathbf{e}_i^0, \mathbf{e}_j^0), \mathbf{e}_n^0 \rangle = \langle B(\mathbf{e}_i^0, \mathbf{e}_j^1), \mathbf{e}_n^1 \rangle = \langle B(\mathbf{e}_i^1, \mathbf{e}_j^0), \mathbf{e}_n^1 \rangle = \langle B(\mathbf{e}_i^1, \mathbf{e}_j^1), \mathbf{e}_n^0 \rangle = 0, \quad \forall i, j, n, \quad (6.16)$$

$$\langle B(\mathbf{e}_i^0, \mathbf{e}_j^1), \mathbf{e}_n^0 \rangle = \langle B(\mathbf{e}_j^1, \mathbf{e}_i^0), \mathbf{e}_n^0 \rangle = \begin{cases} -\frac{\gamma\pi n}{\sqrt{2}L^{3/2}}, & \text{if } n = i + j, \\ \frac{\gamma\pi(i-j)}{\sqrt{2}L^{3/2}}, & \text{if } n = |i - j|, \\ 0, & \text{otherwise,} \end{cases} \quad (6.17)$$

and

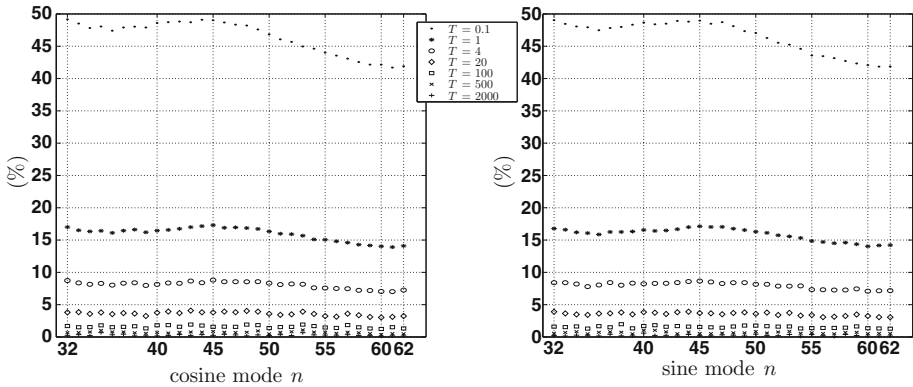
$$\langle B(\mathbf{e}_i^\ell, \mathbf{e}_j^\ell), \mathbf{e}_n^1 \rangle = \begin{cases} (-1)^\ell \frac{\gamma\pi n}{\sqrt{2}L^{3/2}}, & \text{if } n = i + j, \ell \in \{0, 1\}, \\ \frac{\gamma\pi n}{\sqrt{2}L^{3/2}}, & \text{if } n = |i - j|, \ell \in \{0, 1\}, \\ 0, & \text{otherwise.} \end{cases} \quad (6.18)$$

Note that formulas (6.16)-(6.18) show that the parameterization  $\Psi_n^\ell$  in (6.12) is sparse, for  $m + 1 \leq n \leq 2m$  and identically zero for  $n \geq 2m + 1$ .

The optimal QSA,  $\Psi_{\tau^*}$ , is obtained by solving the minimization problems (4.46). The corresponding normalized parameterization defect,

$$J_n(t, \tau) = \frac{\left| \int_t^{t+T} |\Pi_n u(s)|^2 ds - \int_t^{t+T} |\Psi_n(\tau, \boldsymbol{\beta}, u_c(s))|^2 ds \right|}{\int_t^{t+T} |\Pi_n u(s)|^2 ds}, \quad (6.19)$$

is shown in panel (c) of Fig. 16 for the  $\tau = \tau_n^*$ 's that correspond to the optimal values for the cosine modes, dropping here the dependence on  $\ell = 0$ . The results for the sine modes are almost identical, and are thus not shown. Here  $t$  is chosen after the transient behavior, as measured through the energy,  $\|u(t)\|_{L^2}$  of the DNS for Regime A. In our case, it corresponds to  $t = 1$ . The training length  $T$  is chosen to be  $T = 4$ .



**Fig. 17** Relative error of  $\frac{1}{T} \int_t^{t+T} [\Pi_n u(s)]^2 ds$  compared to  $E(n)$ . Here the energy contained in  $E(n)$  is estimated over  $4 \times 10^6$  snapshots, that is for  $T = 4000$

Note that unlike the case dealt with in Sect. 5.2, the cost functional  $J_n$  does not exhibit local minima (in contrast with Remark 8) and thus the dependence on  $t$  is secondary as far as one is concerned with optimal values:  $J_n(t, \tau_n^*)$  will be hereafter denoted by  $J_n(\tau_n^*)$ . Instead,  $\tau \mapsto J_n(\tau)$  exhibits, for  $n = 32$  through  $n = 50$ , sharp gradients near the origin that lead to  $\tau_n^*$ -values close to zero for these modes.

It is striking to observe that  $J_n(\tau_n^*)$  is almost identical to zero for  $n = 32$  up to  $n = 50$  (see Fig. 16c), resulting by an almost perfect parameterization of the energy contained into the corresponding modes; compare the red curve with the black curve in Fig. 16e. For instance, the corresponding optimal QSA comes with a (average) relative error of only 1.3% over the wavenumbers  $32 \leq k \leq 36$ , allowing in turn to fix the dramatic backscatter transfer of energy issue encountered by the standard QSA and even by standard Galerkin approximations with  $m > k_c$ ; see Remark 11 below.

This ability of the optimal QSA to accurately reproduce the amount of energy contained in the consecutive high modes located after the cutoff scale, is even more striking when one notes that QSA( $\tau$ ) is optimized by minimizing  $J_n$  on DNS data over a training length  $T = 4$  (corresponding to  $4 \times 10^3$  snapshots) whereas the energy spectrum  $E(k)$  shown in Fig. 16e is estimated over  $T = 4000$  ( $4 \times 10^6$  snapshots). The relative error  $r$  of  $\frac{1}{T} \int_t^{t+T} [\Pi_n u(s)]^2 ds$  compared to  $E(n)$  is shown as  $T$  evolves in Fig. 17 for the cosine and sine modes. For  $T = 4$  the average error is about 8%. Even if  $T = 1$  (corresponding to  $r \approx 16\%$ ) is selected to evaluate  $J_n$ , the resulting optimal QSA performs similarly than that optimized with  $T = 4$ , regarding the reproduction of the amount of energy contained in the high modes (not shown).

These observations show the usefulness of our variational approach: By optimizing the parameterization QSA( $\tau$ ) according to the cost functional  $J_n$ , one fixes the backscatter transfer of energy issue encountered by the standard QSA, while relying only on a short integration of the KSE. Furthermore, on a practical ground, it is worthwhile noting that one benefits greatly from the dynamically-based formulas QSA( $\tau$ ) (see (4.42)–(4.44)) to operate this optimization. As a comparison, a blind regression using homogeneous polynomials of degree 2 in the  $\xi$ -variable, would lead in this case to  $31 \times 15 \times 3 = 1395$  coefficients<sup>10</sup> to estimate for each high mode and by taking  $T = 1$  or  $T = 4$  ( $4 \times 10^3$  snapshots) the resulting regression problem would be either underdetermined or non-robust statistically. Instead,

<sup>10</sup> Obtained by counting the number of (distinct) monomials  $\xi_i^\ell \xi_j^{\ell'}$ , with  $i, j \in \{1, \dots, 31\}$ , and  $\ell, \ell' \in \{0, 1\}$ .

due to the parametric form of  $QSA(\tau)$ , only 2 scalar parameters ( $\tau_n^\ell$ ,  $\ell = 0, 1$ ) need to be determined, for each high mode.

As a complimentary diagnosis metric, we show in Fig. 16d, for the  $\tau_n^{*}$ 's obtained by minimizing (6.19), the values of the following parameterization defect,

$$Q_n(\tau_n^*) = \frac{\int_t^{t+T} |\Pi_n u(s) - \Psi_n(\tau_n^*, \beta, u_c(s))|^2 ds}{\int_t^{t+T} |\Pi_n u(s)|^2 ds}, \tag{6.20}$$

also for the cosine modes, and for  $t = 1$  and  $T = 4$ . Clearly for the modes whose wavenumbers are located right above the cutoff wavelength,  $k_c$ , the  $Q_n$ -values, although less than 1, are not as close to zero as for the  $J_n$ -values shown in Fig. 16c. Remark that since the mean values of the components of our KS-solution are zero, minimizing  $Q_n$  consists of minimizing the variance of the residual error, i.e.  $|u_n - f(\tau, u_c)|^2$ , for a given parameterization  $f(\tau, \cdot)$ . By construction, minimizing  $J_n$  consists instead of minimizing the residual error of the variance approximation, i.e.  $||u_n|^2 - |f(\tau, u_c)|^2|$ .

It is noteworthy that the  $Q_n$ -values in (6.20) differ slightly from the optimal ones that would be found by minimizing directly the  $Q_n$ 's in the  $\tau$ -variable, over the same training length. Nevertheless, the resulting differences in the corresponding minimizers matters as one would encounter an under-parameterization of about 50% (in average) for the modes near the cutoff wavelength ( $32 \leq n \leq 36$ ); see Remark 11 below.

To better understand the effect of the training length  $T$  (that determines the amount of data from DNS to be stored), we proceeded as follows. Given a training length  $T$ , the optimal QSA,  $\Psi_{\tau^*}$ , is determined by minimizing the corresponding cost functional  $J_n$  given by (6.19) (with  $t = 1$ ), providing thus the optimal parameters,  $\tau_n^{*}$ 's. Recalling that the interaction coefficients are zero for  $n \geq 2m + 1$  (see (6.16)–(6.18)), we analyzed then numerically the dependence on  $t$  and  $T$  of the following averaged parameterization defect

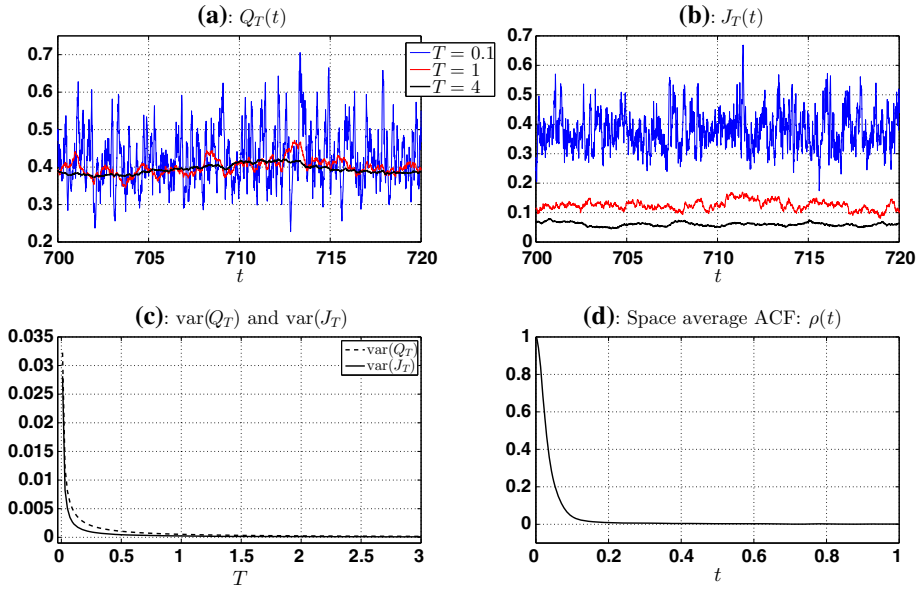
$$J_T(t, \Psi_{\tau^*}) = \frac{\sum_{n=m+1}^{2m} \left| \int_t^{t+T} [\Pi_n u(s)]^2 ds - \int_t^{t+T} [\Psi_n(\tau_n^*, \beta, u_c(s))]^2 ds \right|}{\sum_{n=m+1}^{2m} \int_t^{t+T} [\Pi_n u(s)]^2 ds}, \tag{6.21}$$

as well as of the parameterization defect  $Q_T(t, \Psi_{\tau^*})$  given by (3.4). To simplify the notations, we denote hereafter  $J_T(t, \Psi_{\tau^*})$  and  $Q_T(t, \Psi_{\tau^*})$  by  $J_T(t)$  and  $Q_T(t)$ , respectively. Panels (a) and (b) of Fig. 18 show the dependence on  $t$  of  $J_T(t)$  and  $Q_T(t)$ , respectively. This dependence is shown here for three values of  $T$ :  $T = 0.1$ ,  $T = 1$ , and  $T = 4$ . In each case,  $Q_T(t) < 1$  showing that  $\Psi_{\tau^*}$  is a PM, even for the short training length  $T = 0.1$ . Either for  $Q_T(t)$  or  $J_T(t)$  we observe that the amplitude of the oscillations in time is reduced as  $T$  is increased. This is further confirmed by inspecting the variance of  $Q_T$  and  $J_T$  as  $T$  is varied: both exhibit a fast convergence towards zero as  $T$  grows; see panel (c) of Fig. 18.

The decay towards zero of these variances can be put into perspective with the following space average temporal ACF,

$$\rho(t) = \frac{1}{2\pi T} \int_0^{2\pi} \int_0^T u(x, s)u(x, t + s) ds dx. \tag{6.22}$$

The latter quantity informs us on how the spatio-temporal field,  $u(x, t)$ , decorrelates in time, after averaging over  $x$ . This space average ACF is shown in panel (d) of Fig. 18. It exhibits decay of correlations on timescales comparable to those for the variances of  $Q_T$  and  $J_T$  supporting thus an earlier statement that the coefficients  $\delta_n(\tau)$ 's in (4.43) are optimized in the  $\tau$ -variable by solving the minimization problems (4.46) over short training periods of



**Fig. 18** Effects of the training period,  $T$ , on the parameterization defects  $J_T(t)$  and  $Q_T(t)$ . Here, we observe that: (i) as  $T$  is increasing,  $J_T(t)$  and  $Q_T(t)$  are converging towards a constant value (Panels **a** and **b**), (ii) the variance of  $J_T(t)$  (resp.  $Q_T(t)$ ),  $\text{var}(J_T)$  (resp.  $\text{var}(Q_T)$ ), decays to zero (Panel **c**), and (iii) the rate of decay of the latter is comparable to that of the space average ACF,  $\rho(t)$ , given by (6.22) (Panel **d**)

length comparable to a characteristic decorrelation time of the dynamics. For our closure results presented hereafter we selected  $T = 4$ .

Thus, after minimization in the  $\tau$ -variable of the cost functionals,  $J_n$ 's, given by (6.19), (with  $T = 4$  and after removal of transient,  $t = 1$ ), we use the resulting optimal (and sparse) PM,  $\text{QSA}(\tau^*)$  (i.e.  $\Psi_{\tau^*}$ ), with

$$\tau^* = \{\tau_{n,\ell}^*, : m + 1 \leq n \leq 2m, \ell = 0, 1\},$$

to construct the following optimal PM closure

$$\frac{dz_j^\ell}{dt} = \beta_j z_j^\ell + \left\langle B(z + \Psi_{\tau^*}(z), z + \Psi_{\tau^*}(z)), e_k^\ell \right\rangle, \quad 1 \leq j \leq m, \quad \ell \in \{0, 1\}, \quad (6.23)$$

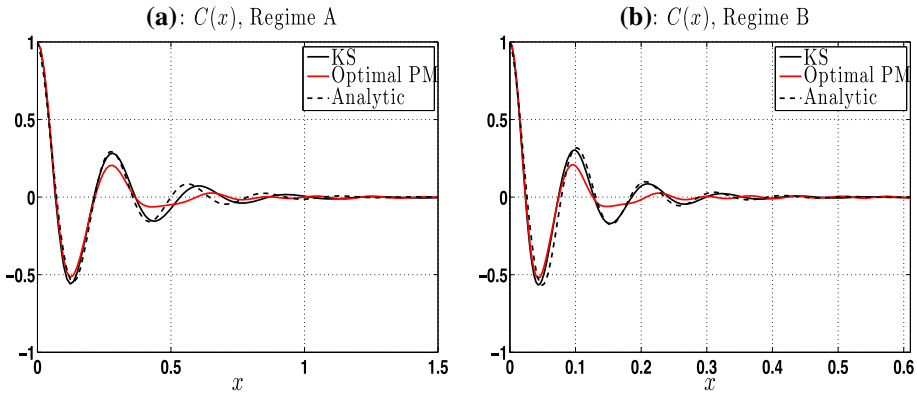
where  $z(x, t) = \sum_{\ell=0}^1 \sum_{j=1}^m z_j^\ell(t) e_j^\ell(x)$ , for  $m = 31$ , that, we recall, corresponds to the number of pairs of unstable modes.

Good closure skills are already visible with naked eyes, by simply comparing the solution patterns,  $u(x, t)$ , obtained by a full integration of Eq. (6.1) over  $N_x$  modes (i.e.  $u$  obtained by DNS), with the patterns exhibited by the optimal PM closure solution,

$$v(x, t) = z(x, t) + \Psi_{\tau^*}(z(x, t)), \quad (6.24)$$

obtained by resolving only  $m = 31$  pairs of reduced variables (i.e. by solving system (6.23)); compare panels (a) and (b) of Fig. 16.

To further assess the ability to reproduce the spatio-temporal dynamics by the optimal PM closure (6.23), we estimated the following time average spatial ACF



**Fig. 19** Time average spatial ACF,  $C(x)$ , for Regimes A and B. In both cases, the spatial ACF,  $C(x)$ , is estimated from (6.25) based on long simulations of the KSE and the optimal PM closure (6.23), with  $\tau^*$  minimizing the  $J_n$ 's given by (6.19). The simulation lengths correspond here, respectively, to  $N = 4 \times 10^6$  snapshots for Regime A, and to  $N = 2 \times 10^6$  snapshots for Regime B. These estimated ACFs are compared with the analytic formula for  $C(x)$  proposed in (6.26)

$$C(x) = \frac{1}{LT_f} \int_0^{T_f} \int_0^L u(x', t)u(x + x', t) dx' dt, \tag{6.25}$$

for  $u$  as obtained from DNS and its approximation  $v(x, t)$  given by (6.24), both integrated up to  $T_f = 4000$ , while we recall that the training length is  $T = 4$  to determine  $\Psi_{\tau^*}$ . The results are shown in panel (a) of Fig. 19. The correlation function  $C(x)$  captures both the underlying oscillatory, cellular spatial structure of the KS dynamics, and the rapid spatial decorrelation reflecting the spatial disorder in the spatio-temporal chaotic regime analyzed here. These features are thus well captured by the optimal PM closure (6.23).

Following [174], we observed that the time average spatial ACF is well modeled for the DNS by the following analytic formula,

$$C(x) \approx \cos(k_p^{-1}x) \exp(-x/\lambda), \tag{6.26}$$

with  $k_p$  that corresponds to the wavelength associated with the peak in the energy spectrum  $E(k)$  shown in Fig. 16e, and  $\lambda$  to a correlation length for which spatial coupling becomes negligible beyond a few multiples of  $\lambda$ . For Regime A, we found  $k_p = 21$  and  $\lambda = 0.23$ . Only for large lags in the  $x$ -variable, the optimal PM fails to reproduce accurately this theoretical prediction.

**Remark 10** The QSA (4.40) may also be obtained as the limit of the parameterization

$$K_{\tau}(\xi) = -\tau(\text{Id} + \tau A\Pi_{\mathfrak{s}})^{-1}\Pi_{\mathfrak{s}}B(\xi, \xi), \tag{6.27}$$

obtained by using an implicit Euler method to approximate the high modes and by simplifying the nonlinear terms; see [63] and [67, Sec. 7.1]. In this case we have,

$$\lim_{\tau \rightarrow \infty} -\tau(\text{Id} + \tau A\Pi_{\mathfrak{s}})^{-1}\Pi_{\mathfrak{s}}B(\xi, \xi) = -A_{\mathfrak{s}}^{-1}\Pi_{\mathfrak{s}}B(\xi, \xi). \tag{6.28}$$

Note that in (6.27) unlike in [63], we consider the operator  $A$  to be the full linear operator and not only given by the 4th-order term. In its standard formulation, the parameterization  $K_{\tau}$  is not optimized and  $\tau$  is chosen to be  $\lambda_{m+1}^{-1}$ , where  $\lambda_m = 16\nu\pi^4 m^4/L^4$  denotes the eigenvalue of  $\nu\partial_x^4$ .

**Table 8** 1st and 2nd moments of  $\|u\|_{L^2}$ : relative error for regime A

	Energy contained in $E_s$ (%)	$\overline{\ u\ _{L^2}}$ (%)	$\text{std}(\ u\ _{L^2})$ (%)
QSA( $\tau^*$ )-closure (6.23), $\tau^*$ minimizing the $J_n$ 's	15.7	3.2	3.8
QSA( $\tau^*$ )-closure (6.23), $\tau^*$ minimizing the $Q_n$ 's	15.7	6.9	1.3
Galerkin ( $m = 49$ )	0.9	42.1	307
Galerkin ( $m = 53$ )	0.4	16.6	101
Galerkin ( $m = 58$ )	0.2	3.1	5.8
Galerkin ( $m = 61$ )	0.1	0.8	3.1

Taking  $A = \nu\partial_x^4 + D\partial_x^2$ , the analytic expression of the parameterization  $K_\tau$  is the same as for QSA( $\tau$ )(4.42), except that  $\delta_n(\tau)$  therein is replaced by  $\tau(1 - \beta_n\tau)^{-1}$ . Since  $0 \leq \tau(1 - \beta_n\tau)^{-1} < -\beta_n^{-1}$ , the range of this coefficient is the same as that of  $\delta_n(\tau)$  (see discussion at the end of Sect. 4.4), and the parameterization  $K_\tau$  once optimized by minimizing the cost functional  $J_n$  leads also to similar closure skills than those obtained by the optimal QSA.<sup>11</sup> We see thus here that the PM approach is not limited to the QSA-class nor the LIA-class introduced respectively in Sects. 4.4 and 4.3, but applies actually to any parametric family of nonlinear parameterizations.

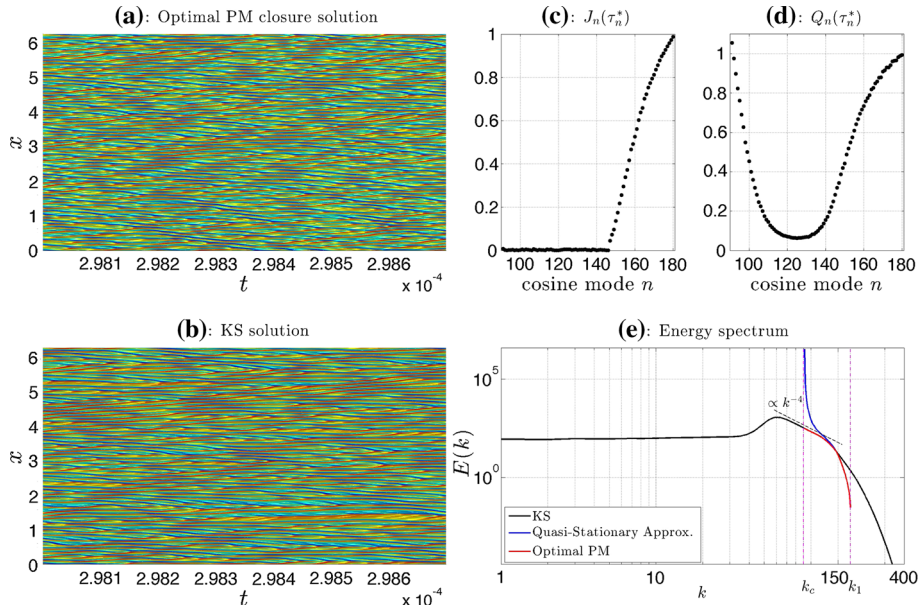
**Remark 11** We report briefly here on the closure skills obtained when QSA( $\tau$ ) is optimized by minimizing the  $Q_n$ 's instead of the  $J_n$ 's. The metrics used to assess these skills are  $\overline{\|u\|_{L^2}}$  (after transient removal) and its standard variation,  $\text{std}(\|u\|_{L^2})$ . The time averages are here estimated on an interval of length  $T = 100$  ( $10^5$  snapshots). We observe from Table 8 that the relative error of approximation for  $\overline{\|u\|_{L^2}}$  is increased while that for  $\text{std}(\|u\|_{L^2})$  is reduced, when the 62D closure (6.23) ( $m = 31$ ) is driven by the optimal QSA( $\tau^*$ ) with  $\tau^*$  minimizing the  $Q_n$ 's. Comparison with standard Galerkin approximations, show that only starting from a 118D Galerkin approximations ( $m=59$ ), one starts to improve, compared to the 62D closure,<sup>12</sup> the approximation of the mean value of  $\|u(t)\|_{L^2}$  (and comparable skills for  $\text{std}(\|u\|_{L^2})$ ) although a good reproduction of the KS patterns' qualitative features, is observed for lower dimension. However this latter aspect seems to be germane to the KSE. In general, indeed, an error in the reproduction of the right amount of energy come with failures in the reproduction of qualitative features as well, due to an incorrect reproduction of the backscatter transfer of energy. For instance, regarding the wind-driven circulation of the oceans [75], the jet extension and variability [53] are notoriously difficult to get parameterized due to eddy backscatter [7,8].

### 6.3 Closure Results in Presence of 90 Pairs of Unstable Modes

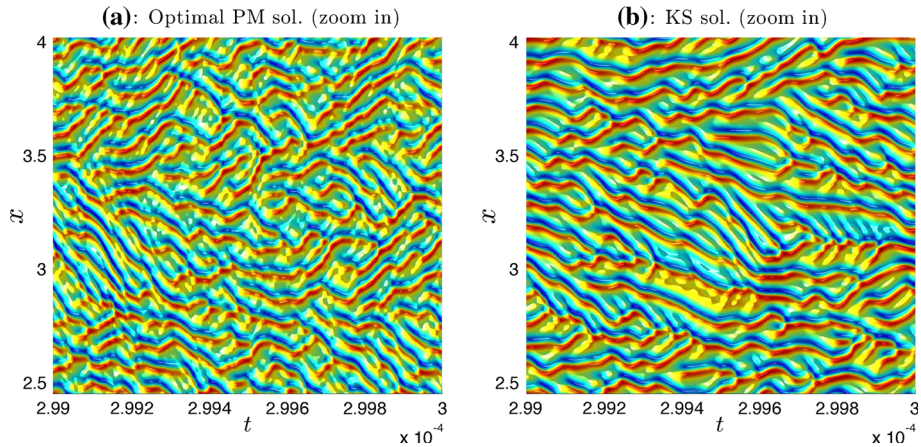
The ability of the optimal QSA to fix the backscatter transfer of energy issue, providing thus an efficient closure, is further tested by applying the PM approach to an even more turbulent regime, namely Regime B (see Table 7) that exhibits 90 pairs of unstable modes. Due to the

<sup>11</sup> Note that by taking  $A$  to be given by  $\nu\partial_x^4$  the resulting coefficients are bounded by  $\lambda_n^{-1}$ , and since  $\lambda_n^{-1} < -\beta_n^{-1}$  the optimized  $K_\tau$  is not a priori of comparable parameterization defects, and in fact leads to less efficient closures.

<sup>12</sup> Driven by the optimal QSA( $\tau^*$ ) with  $\tau^*$  minimizing the  $J_n$ 's.



**Fig. 20** Closure and parameterization results for Regime B. Same as Fig. 16 except that  $k_c = 90$ , since Regime counts 90 pairs of unstable modes. The energy spectrum  $E(k)$  in panel e is estimated over  $N = 2 \times 10^6$  snapshots whereas the optimal QSA is determined by minimizing the cost functional,  $J_n$ , exploiting the first  $2 \times 10^4$  snapshots (after removal of transient). Figure 21 shows blowup regions of panels a and b corresponding to  $2.5 \leq x \leq 4$

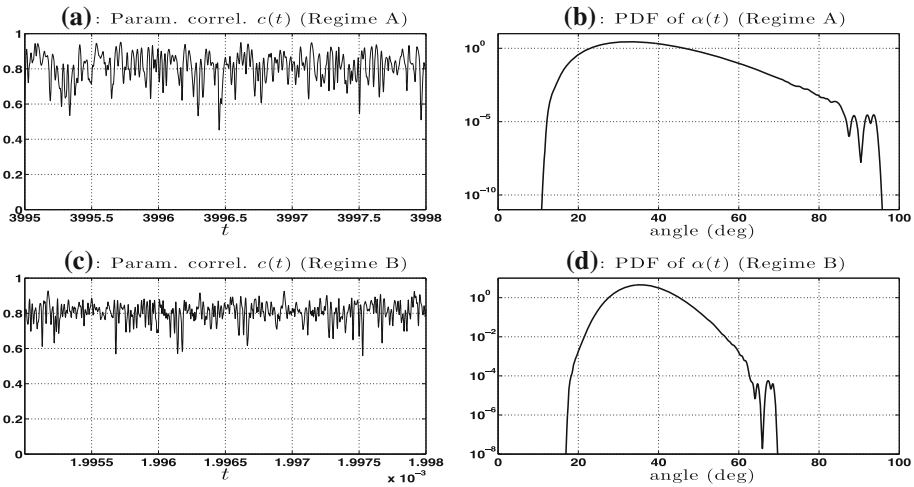


**Fig. 21** Closure results for Regime B: patterns. Blowup regions of panels a and b of Fig. 20 corresponding to  $2.5 \leq x \leq 4$

scaling (6.3) and the large value of  $\alpha$  (see Table 7) the time variable for Eq. (6.2) evolves on a much smaller timescale than for Eq. (6.1) and as a consequence we will often emphasize the number of snapshots that a given time instant represents rather than giving the (small) value of this time.

Here again we take the cutoff scale to be given by the smallest scale (higher wavenumber) contained among the unstable modes. Thus for Regime B,  $k_c = 90$ , and here also, 15.7% of the





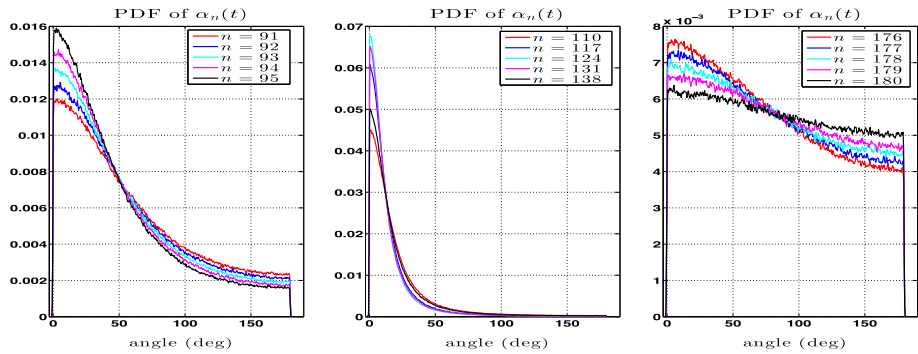
**Fig. 22** Parameterization correlation  $c(t)$ , and PDF of the parameterization angle  $\alpha(t)$ . Here these results are obtained for the optimal QSAs,  $\text{QSA}(\tau^*)$  used in Fig. 16 for Regime A, and in Fig. 20 for Regime B, that is with  $\tau^*$  minimizing the  $J_n$ 's with  $n \geq k_c = 31$  for Regime A, and  $n \geq k_c = 90$ , for Regime B. A semi-log scale is used for **b** and **d**

total amount of energy needs to be parameterized at this cutoff scale. For this more turbulent regime, the standard QSA fails even more dramatically than for Regime A and leads to an (ridiculous) over-parameterization of  $E(k)$  by an amount of about  $35 \times 10^3 \%$  (in average) over the range of wavenumbers  $91 \leq k \leq 121$ ; see blue curve in Fig. 20e. In contradistinction, the optimal QSA,  $\text{QSA}(\tau^*)$ , obtained by minimizing  $J_n$  given in (6.19) with  $T$  that corresponds to the first  $2 \times 10^4$  snapshots (after removal of transient),<sup>13</sup> leads to an average error of about 0.7% over the same range of wavelengths, fixing thus here again the backscatter transfer of energy to the large scales. As a consequence, good closure skills are obtained as shown in Fig. 20 for the reproduction of KS patterns, demonstrating furthermore the robustness of our approach to even more turbulent regimes. Note that  $Q_n$  is greater than 1 only for  $n = 91$  (see panel (d) of Fig. 20). This does not affect the overall quality of the  $\text{QSA}(\tau^*)$ -parameterization (optimized for the  $J_n$ 's) and we have still  $Q_T$  given by (3.4) that is strictly less than 1, here.

A finer inspection of the patterns is made possible by Fig. 21 which shows blowup regions of panels (a) and (b) of Fig. 20. Here, we observe that as time evolves the creation and annihilation of the humps displayed by the optimal PM closure solution is reminiscent with what can be observed for the KS solution. Statistically, the spatial correlations are also well reproduced for Regime B as shown in panel (b) of Fig. 19. Only the small-scale features of the optimal PM closure solution and the spatial coherence at long-range distance require improvements, and in that respect one might pursue some ideas proposed in Sect. 7 below.

These closure and parameterization skills are put into perspective by computing for each regime, the parameterization correlation,  $c(t)$ , (see (3.6)) and PDF of the corresponding parameterization angle,  $\alpha(t)$  (see (3.7)). As shown in panels (a) and (c) of Fig. 22,  $c(t)$  fluctuates away from 1, and  $\alpha(t)$  fluctuates over a broad range of values relatively far away

<sup>13</sup> Note that a blind regression would lead in this case to  $89 \times 45 \times 3 = 12015$  coefficients to estimate for each high mode; a number of coefficients comparable to the number of snapshots making thus the estimated coefficients by regression non-robust. Instead, one benefits here again greatly from the parametric (and dynamically-based) form of  $\text{QSA}(\tau)$  and only 2 scalar parameters ( $\tau_n^\ell$ ,  $\ell = 0, 1$ ) need to be determined, for each high mode.



**Fig. 23** PDFs of  $\alpha_n(t)$  given by (6.30). Here the PDFs are shown in linear scale

from zero. This situation is indicative that for both regimes, the optimal PM computed here is far from a slaving situation.

However, the distribution of  $\alpha(t)$  does not seem to be consistent with the good closure results shown here and the rule of thumb pointed out in Sect. 3.1.2. The reason behind this is the large number of modes parameterized (here 90 pairs) that makes the parameterization correlation less representative of the quality of a given parameterization than for low-dimensional systems. In the same vein that we have used modewise parameterization defects (the  $Q_n$ 's) instead of the global parameterization defect  $Q_T(t, \Psi_{\tau^*})$  given by (3.4), we inspect below a modewise version of  $c(t)$  to diagnose our parameterizations.

In that respect, for the bidimensional real vector  $f_n(t) = (f_n^0(t), f_n^1(t))$  with  $f_n^\ell(t) = \Psi_n^\ell(\tau_{n,\ell}^*, y_c(t))$ ,  $\ell = 0, 1$ , we introduce

$$c_n(t) = \frac{\langle f_n(t), y_n(t) \rangle}{\|f_n(t)\| \|y_n(t)\|}. \tag{6.29}$$

and the following parameterization angle,

$$\alpha_n(t) = \arccos(c_n(t)). \tag{6.30}$$

We computed  $c_n(t)$  and  $\alpha_n(t)$  for  $n = 91$  through  $n = 180$ . Figure 23 shows the results for the PDFs of  $\alpha_n(t)$ , as gathered into three groups: a group of parameterized modes adjacent to the cutoff scale, a group of modes (well) within the inertial range, and a group of modes corresponding to the smallest scales parameterized. Clearly the PDFs corresponding to the 2nd group of modes correspond to the best modewise parameterizations; compare middle panel of Fig. 23 with the two other panels of the same figure. Here, we observe for this group of modes PDFs that exhibit features discussed in Sect. 3.1.2. These PDFs are indeed skewed towards zero with the most frequent value of  $\alpha_n(t)$  also close to zero; cf. black curve in Fig. 2. These features are also shared by the PDFs of the adjacent modes to the cutoff scale (left panel of Fig. 23) with however a fat tail towards high values of  $\alpha_n(t)$ . The last group of modes corresponding to high wavenumbers (right panel of Fig. 23) corresponds to the less accurate modewise parameterizations as manifested by PDFs of  $\alpha_n(t)$  that although skewed are somewhat close to a uniform distribution.

These small-scale modes are weakly energetic, they contain less than 0.6 % of the total energy for  $n > 150$ , and here do not spoil the parameterization noticeably. However the fat tails of the PDFs corresponding to the adjacent parameterized modes is a determining factor responsible of pushing the (global) parameterization correlation,  $c(t)$  (given by (3.6)), away

**Table 9** 1st and 2nd moments of  $\|u\|_{L^2}$ : relative error for regime B

	$\ u\ _{L^2}$ (%)	$\text{std}(\ u\ _{L^2})$ (%)
QSA( $\tau^*$ )-closure, $\tau^*$ minimizing the $J_n$ 's	4	3.2
QSA( $\tau^*$ )-closure, $\tau^*$ minimizing the $Q_n$ 's	7.5	1.6
LIA( $\tau^*$ )-closure with $\tau^*$ minimizing the $J_n$ 's	8.9	1.7
LIA( $\tau^*$ )-closure with $\tau^*$ minimizing the $Q_n$ 's	10.2	0.3

from 1, as it can be observed by removing the contribution of these modes in the calculation of  $c(t)$  (not shown). On the other hand, these adjacent modes are important dynamically and cannot be removed for closure as they contain an amount of energy comparable to that of the modes right below the cutoff scale (i.e. for  $k < k_c$ ).

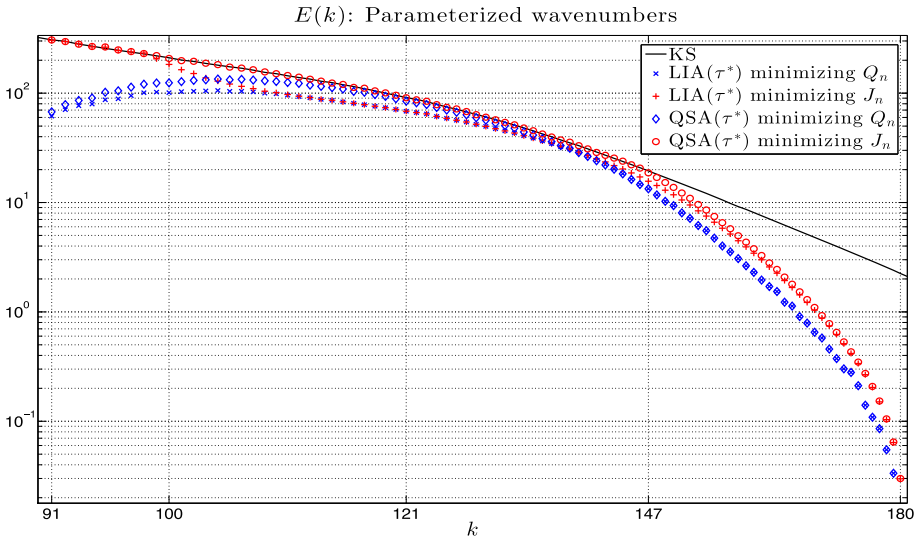
We conclude by reporting on how the choice of the cost functional and class of parameterization impacts the closure skills. The metrics used to assess these skills are those used for Table 8, namely  $\|u\|_{L^2}$  (after transient removal) and the standard variation,  $\text{std}(\|u\|_{L^2})$ . The time averages are here estimated on  $2 \times 10^4$  snapshots. As Table 9 shows, minimizing the  $Q_n$ 's instead of the  $J_n$ 's leads to a deterioration in the approximation of  $\|u\|_{L^2}$  but an improvement in the standard variation within a given class of parameterizations.

The portion of the energy spectrum  $E(k)$  parameterized—by the optimal LIA( $\tau^*$ ) or QSA( $\tau^*$ ) with  $\tau^*$  minimizing either the cost functionals  $J_n$ 's or  $Q_n$ 's—is shown in Fig. 24. As one can observe, the QSA( $\tau^*$ ) obtained by minimizing the  $J_n$ 's provides the best result and an almost perfect parameterization of the energy contained in the high modes over the range of wavenumbers,  $91 \leq k \leq 147$ , resulting thus into the good closure skills shown in Fig. 20 and panel (b) of Fig. 19. We emphasize that as for Regime A, these skills are obtained from an optimal PM designed from a training interval over which the statistics of  $|u_n|^2$  have not yet stabilized; cf. discussion relative to Fig. 17 for Regime A. When the  $Q_n$ 's are used to optimize either the LIA( $\tau$ )- or the QSA( $\tau$ )-parameterization, one observes an under-parameterization more pronounced near the cutoff scale  $k_c = 90$  and that vanishes as  $k$  is increased, before re-emerging beyond wavenumbers that contain a small fraction of the total energy  $E_{\text{tot}}$ ; for instance the scales beyond  $k = 147$ , contain only 0.6% of  $E_{\text{tot}}$ . Despite this under-parameterization, the optimal LIA( $\tau^*$ ) and QSA( $\tau^*$ ) with  $\tau^*$  minimizing the  $Q_n$ 's, provide also closure skills comparable to those shown in Fig. 20 and panel (b) of Fig. 19. The main differences are actually observed at the level of the approximation of  $\|u\|_{L^2}$  and  $\text{std}(\|u\|_{L^2})$ , as summarized in Table 9. We refer to the heuristic discussion at the end of Sect. 4.4 to better appreciate the nuances between the LIA- and QSA-classes of parameterizations in regards of these numerical results.

## 7 Concluding Remarks

Thus, the PM approach is not limited to a class of parametric parameterizations nor to a particular cost functional. As the closure exercise shows here in the context of KS turbulence, a good choice of the cost functional and class of parameterizations to optimize is nevertheless key to approximate certain features better than others. This is where the specificities of the problem at hand plays an important role<sup>14</sup> and where one may benefit from the flexibility

<sup>14</sup> In that respect, we may mention the variational normal mode initialization in Meteorology, pioneered by Daley [45], who combined the Machenhauer [127] non-linear normal-mode initialization within a variational



**Fig. 24** Approximations of  $E(k)$  for  $k_c < k \leq k_1$  for Regime A. Optimal LIA( $\tau^*$ ) and QSA( $\tau^*$ ) with  $\tau^*$  minimizing either the cost functionals  $J_n$ 's or  $Q_n$ 's. Recall that  $k_c = 90$  and  $k_1 = 2k_c$ . A log–log scale is used here

of the PM approach to optimize relevant parameterizations known by the practitioner, once the underlying formulas are made parametric, i.e. made as a function of a (collection) of (independent) scalar variable(s).

Rooted in the rigorous approximation theory of invariant manifolds (Part I), this articles provides a natural framework to extend the corresponding approximation formulas as non-linear parameterizations useful when slaving relations do not hold anymore, e.g., away from criticality (Part II). The framework opens up several possible directions for future research. We outline some of these directions below.

*1. Time-dependent parameterizing manifolds for non-autonomous systems* As for the autonomous case discussed here, formulas for time-dependent PMs may be rooted in the approximation theory of time-dependent invariant manifolds [143,144]. The leading order approximation,  $h_2$ , becomes now time-dependent and satisfies the following version of the homological equation (2.27) (with  $\mathcal{L}_A$  defined in (2.54)),

$$(\partial_t + \mathcal{L}_A)h = \Pi_\mathfrak{s} B(\xi, \xi) + \Pi_\mathfrak{s} F(t), \tag{7.1}$$

for a system of the form

$$\frac{dy}{dt} = Ay + B(y, y) + F(t), \quad y \in \mathbb{C}^N. \tag{7.2}$$

The backward–forward method to derive parametric formulas for PMs, extends to this non-autonomous setting and provides a parametric family of time-dependent manifold function,

---

Footnote 14 continued  
 procedure allowing for the adjustment of confidence weights arising in a fidelity functional  $I$ ; see also [168]. In these works, the manifold  $\mathcal{M}$  is fixed a priori and it is the point on  $\mathcal{M}$  nearest to the observation using the “metric” defined by  $I$ , that is sought.

$\Psi_\tau^{(1)}(t, \cdot)$ , that satisfies for instance in the case  $\Pi_c F = 0$ , the following modification of Eq. (4.6)

$$\left(\partial_t + \mathcal{L}_A\right)\Psi_\tau^{(1)}(t, \xi) = \Pi_s B(\xi, \xi) - e^{\tau A_s} \Pi_s B(e^{-\tau A_c} \xi, e^{-\tau A_c} \xi) + \Pi_s F(t) - e^{\tau A_s} \Pi_s F(t - \tau). \tag{7.3}$$

Due to the time-dependent coefficients to calculate in  $\Psi_\tau^{(1)}(t, \cdot)$ , the evaluation of the parameterization defect gets more involved than in the autonomous case. Nevertheless, the optimal value for the free parameter  $\tau$  may be still obtained by minimizing this defect, leading to an optimal PM, in the  $\Psi_\tau^{(1)}(t, \cdot)$ -class and thus to closures with time-dependent coefficients. The measure-theoretic framework of Sect. 3 may benefit here from the theory of SRB measures for non autonomous systems [178]. The formulas for the LIA and QSA parameterizations of Sects. 4.3 and 4.4 respectively, extend to this non-autonomous setting as well. The case of a stochastic forcing can be dealt with along the same lines, the backward–forward method providing in this case parametric formulas for PMs that come with non-Markovian coefficients depending on time-history of the noise (exogenous memory terms) [31].

2. *Combining PMs with stochastic parameterizations* To set the framework, we discuss stochastic improvements that can be made to the LIA class of Sect. 4.3, but the ideas apply to the QSA class of Sect. 4.4 as well. Given a cutoff dimension  $m$ , the optimal PM obtained by solving the minimization problems (4.35), for  $n \geq m + 1$ , is the best manifold—in the LIA class—that averages out the unresolved fluctuations lying in  $E_s$ . Once the optimal PM,  $\Phi_{\tau^*}^{(1)}$ , has been determined, we may still want to parameterize these fluctuations. These fluctuations are given by the residual  $\eta_t$  whose components are determined after having solved (4.35) for each  $n \geq m + 1$ . We have then

$$y_s(t) = \Phi_{\tau^*}^{(1)}(y_c(t)) + \eta_t. \tag{7.4}$$

From a closure viewpoint, we are thus left with the stochastic modeling of  $\eta_t$ . The next step consists of seeking for a stochastic parameterization  $\zeta_t$  of  $\eta_t$ . Here several approaches are possible; see [79] for a survey. The idea of incorporating a stochastic ingredient as a supplement to a nonlinear parameterization is not new and has been proposed in the context of two-dimensional turbulence [121], atmospheric turbulence [70] and more recently, oceanic turbulence [179].

Once a satisfactory stochastic parameterization  $\zeta_t$  has been determined, we arrive at the following closure for the resolved variable (in the case of bilinear system),

$$\frac{dz}{dt} = A_c z + \Pi_c B\left(z + \Phi_{\tau^*}^{(1)}(z) + \zeta_t, z + \Phi_{\tau^*}^{(1)}(z) + \zeta_t\right) + \Pi_c F. \tag{7.5}$$

Thinking of  $B$  as given by a nonlinear advective term, we see that the stochastic parameterization (7.4) brings new elements to the closure (7.5) such as stochastic advective terms compared to a closure that would be only based on the optimal PM. Other recent approaches have shown the relevance of such stochastic advective terms to derive stochastic formulations of classical representations of fluid flows as well as for emulating suitably the coarse-grained dynamics [3,39,91,146–148].

The selection of the best parameters (e.g. lags for an auto-regressive process) of a given stochastic parameterization aimed at emulating the residual,  $\eta_t$ , can here again be guided by the minimization of the parameterization defect  $Q_n$ ; the parameters of  $\zeta_t$  being determined so as to minimize further  $Q_n$  compared to when the optimal PM is used alone. Complementarily, the parameterization correlation,  $c(t)$ , for which  $\Psi = \Phi_{\tau^*}^{(1)} + \zeta_t$  in (3.6), can then be evaluated to further revise other ingredients in the stochastic parameterization, so that the probability

distribution of the corresponding correlation angle  $\alpha(t)$  gets skewed towards zero as much as possible. In other words, one should not only parameterize properly the statistical effects of the subgrid scales but also avoid to lose their phase relationships with the retained scales [132]. In that respect, the residual noise  $\eta_t$  in (7.4) is expected to depend on the state of the resolved variable  $\xi$ . The abstract formula (3.26) for the optimal PM suggests that subgrid-scale parameterization techniques with conditional Markov chains [44,78,116] constitute a consistent tool with our approach for the design of a stochastic parameterization  $\zeta_t$ .

**3. Beyond conditional expectation: Memory effects and noise** An alternative to the inclusion of stochastic ingredients as discussed above, relies on Theorem 5 as a starting point. The latter theorem shows that once an optimal PM is found, it provides the conditional expectation (in the case  $\eta = 0$ ). Nevertheless, as shown in Sect. 3.4, the conditional expectation alone, let us say  $\mathbf{R}$ , is sometimes insufficient to close fully the system. The Mori-Zwanzig formalism [134,181] of statistical physics, instructs us then that a complete closure exists under the form of the following *generalized Langevin equation (GLE)* [34,76,79,102],

$$\dot{x} = \mathbf{R}(x) + \int_0^t \mathbf{G}(t, s, x(s)) ds + \eta_t. \quad (\text{GLE})$$

Here, the integral term accounts for the nonlinear interactions between the resolved and unresolved variables that are not accounted for in  $\mathbf{R}$ ; it involves the past of the macroscopic variables and conveys *non-Markovian (i.e. memory) effects*. The term  $\eta_t$  accounts for effects of the unresolved variables which are uncorrelated with the resolved variables. This last term can be thus represented by a state-independent noise that may still involve correlations in time, i.e. of “red noise” type. It is well known that the analytical determination of the constitutive elements of the GLE is a difficult task in practice. By relying on Theorem 5 and formulas of Sect. 4, the PM approach can be seen as providing an efficient way to approximate the conditional expectation  $\mathbf{R}$  in (GLE). However, the practical determination of the memory and stochastic terms remains a challenge, especially for fluid flows [79,102]. Various approaches have been proposed to address this aspect that include for instance short-memory approximations [36], the  $t$ -model [82,159], formal expansions of the Koopman operator [175,176], NARMAX techniques [35,124], and the dynamic- $\tau$  model [135,136]. See also [89,106,107,133,142,179] for other reduced modeling/parameterization approaches that involve memory terms (and noise) in the context of homogeneous turbulence, shear dynamo and oceanic turbulence, respectively.

Once  $\mathbf{R}$  is approximated from an optimal PM, the practical determination of the memory and stochastic term could also benefit from the data-driven modeling techniques of [25], to model the residual,  $\dot{y}_c - \mathbf{R}(y_c)$ , where  $y_c$  denotes the low-mode projection of a fully resolved solution  $y$ . As illustrated and discussed in [105] for a wind-driven ocean gyres model, the data-driven techniques of [25] have been successfully applied to model the coarse-scale dynamics. To operate in practice, the data-driven techniques of [25] require observations of  $y(t)$  of length comparable also to a decorrelation time of the dynamics [25,103,104], as for the optimization of the dynamically-based PMs of Sect. 4.

**4. Combining modal reductions and the PM approach** In many applications such as arising in turbulence, the number of ODEs associated to a given discretization, is very large. This is where modes computed in the physical domain from DNS may be used to proceed to a first reduction (data compression) of the phase space. Among the most commonly employed modal decomposition techniques are the proper orthogonal decomposition (POD) [92], and its variants; see [161] and references therein. Of demonstrated relevance for the reduction of nonlinear PDEs are also the principal interaction patterns (PIPs) modes [86,112,113] that find a compromise between minimizing tendency error with maximizing explained variance

in the resolved modes; see [114,115] for applications to atmospheric models, and [43] for a very clear comparison between POD and PIP modes. In the last decade, related promising techniques such as the dynamic mode decomposition (DMD) [150,155,161,173] have also emerged; see [169] for a discussion on the relationships between PIPs, DMD, and the linear inverse modeling [139].

Also, the use of time-dependence in the basis elements—the so-called Dynamical Orthogonal (DO) modes [153,154]—have been considered, as in principle it allows for the representation of the transient character of the solution using much fewer modes. A dynamical orthogonality condition leads then to a closed set of equations that allows for the evolution of the mean field, the DO modes and the corresponding (stochastic) coefficients [61]. From the mean, the time-dependent patterns of the DO modes plus the distribution of the stochastic coefficients (at a certain time  $t$ ), an approximation to the probability density function of the state vector can be obtained [152,160,170]. In terms of computational performance, there is however a trade-off between fewer modes to consider on one hand, and more equations (including interactions between the modes) to solve, on the other.

For certain problems of turbulence, even after modal reduction, one may wish still to further reduce the dimension of the ODE approximation. Whatever the modes used to represent the dataset at hand, one should avoid to compute parameterizations by taking the reduced state space,  $E_c$ , to be spanned by only the first few modes. There are several reasons behind this caution. One reason is that these modes may mix the large and small spatial scales, making the distinction between  $E_c$  and  $E_s$  not obvious. Another reason, more technical, is that  $E_c$  and its complement  $E_s$  are no longer invariant subspaces for the linear part of the original PDE, which introduces linear interaction terms between the modes in  $E_c$  and  $E_s$  that have to be taken into account for the parameterization. Although one could still apply formally the backward-forward method of Sect. 4 to derive parametric families of parameterizations, a more reasonable approach consists of proceeding directly from the Galerkin ODE systems obtained by projecting the original PDE onto these modes. This way, we are indeed left with the theory and techniques presented in this article, and by determining the equations for the perturbed variable about a mean state and work within the eigenbasis of the linearized operator, we can then use the dynamically-based formulas of Sect. 4 to calculate and optimize the parameterizations.

**Acknowledgements** MDC wishes to acknowledge David Neelin for the stimulating discussions on the closure problem of convective processes in the tropical atmosphere. MDC and JCM are also thankful to Darryl Holm for his constructive comments at the beginning of this work. Finally, MDC and HL are greatly indebted to Shouhong Wang for the numerous and stimulating discussions about this work over the years, and it is a pleasure to express our gratitude to Shouhong for his constant encouragement. This work has been partially supported by the National Science Foundation grants DMS-1616981 (MDC) and DMS-1616450 (HL).

## Appendix: Parameterization Defect Minimization Algorithm

We present in this Appendix a simple gradient-descent method to solve efficiently the minimization problem (4.35) in order to determine the optimal  $\tau$ -value,  $\tau^*$ , for the parameterization,  $\Phi_n(\tau, \beta, \xi)$ , given by (4.34). As shown below, the method allows furthermore for making apparent the dependence of the parameterization defect on statistical moments (up to order 4) of the original system's solution.

To present the method, we first recast the parameterization defect associated with  $\Phi_n$ ,

$$\mathcal{Q}_n(\tau, T) = \frac{1}{T} \int_0^T |\Pi_n y(t) - \Phi_n(\tau, \beta, \Pi_c y(t))|^2 dt, \quad (\text{A.1})$$

into a matrix format. For this purpose, we arrange the coefficients  $D_{i,j}^n(\tau, \boldsymbol{\beta})B_{i,j}^n$  involved in the expression of  $\Phi_n(\tau, \boldsymbol{\beta}, \xi)$  into an  $m^2 \times 1$  vector  $\mathbf{d}(\tau)$  so that the indices  $(i, j)$ 's are arranged in lexicographical order; namely the  $k^{\text{th}}$  component of  $\mathbf{d}(\tau)$  is given by

$$d_k(\tau) = D_{i,j}^n(\tau, \boldsymbol{\beta})B_{i,j}^n, \quad k = 1, \dots, m^2, \tag{A.2}$$

where  $(i, j)$  is the unique low-mode pair of indices satisfying

$$(i - 1)m + j = k, \quad \text{with } i, j \in \{1, \dots, m\}. \tag{A.3}$$

More precisely, the index pair  $(i, j)$  in (A.2) is determined by:

$$\begin{cases} i = \frac{k - \text{mod}(k, m)}{m} + 1 \text{ and } j = \text{mod}(k, m), & \text{if } \text{mod}(k, m) \neq 0, \\ i = \frac{k}{m} \text{ and } j = m, & \text{otherwise.} \end{cases} \tag{A.4}$$

Similarly, we define an  $m^2 \times 1$  vector  $\boldsymbol{\gamma}(\tau)$ , whose components are given by

$$\gamma_k(\tau) = V_{i,j}^n(\tau, \boldsymbol{\beta})F_j(B_{i,j}^n + B_{j,i}^n), \quad k = 1, \dots, m^2. \tag{A.5}$$

Now, given the solution  $y(t)$  to the underlying  $N$ -dimensional ODE system (4.16) over  $[0, T]$ , we introduce

$$u_k(t) = \Pi_k y(t), \quad k = 1, \dots, m,$$

where  $\Pi_k$  denotes the projection onto the mode  $\mathbf{e}_k$ ; see (4.19).

We define next the column vectors  $\mathbf{Q}_1, \mathbf{Q}_2, \widehat{\mathbf{Q}}_2$  and  $\mathbf{Q}_3$  of size  $m^2 \times 1$  as well as the matrices  $\widetilde{\mathbf{Q}}_2, \widetilde{\mathbf{Q}}_3$  and  $\mathbf{Q}_4$  of size  $m^2 \times m^2$  as follows:

$$\begin{aligned} (\mathbf{Q}_1)_p &= \langle \bar{u}_{p_1} \rangle_T, \quad p = 1, \dots, m^2, \\ (\mathbf{Q}_2)_p &= \langle \bar{u}_{p_1} \bar{u}_{p_2} \rangle_T, \quad p = 1, \dots, m^2, \\ (\widehat{\mathbf{Q}}_2)_p &= \langle u_n \bar{u}_{p_1} \rangle_T, \quad p = 1, \dots, m^2, \\ (\mathbf{Q}_3)_p &= \langle u_n \bar{u}_{p_1} \bar{u}_{p_2} \rangle_T, \quad p = 1, \dots, m^2, \\ (\widetilde{\mathbf{Q}}_2)_{pq} &= \langle \bar{u}_{p_1} u_{q_1} \rangle_T, \quad p, q = 1, \dots, m^2, \\ (\widetilde{\mathbf{Q}}_3)_{pq} &= \langle \bar{u}_{p_1} u_{q_1} u_{q_2} \rangle_T, \quad p, q = 1, \dots, m^2, \\ (\mathbf{Q}_4)_{pq} &= \langle \bar{u}_{p_1} \bar{u}_{p_2} u_{q_1} u_{q_2} \rangle_T, \quad p, q = 1, \dots, m^2, \end{aligned} \tag{A.6}$$

where  $\bar{z}$  denotes the complex conjugate of  $z$  in  $\mathbb{C}$ ,  $\langle \cdot \rangle_T$  denotes the time average over  $[0, T]$ , and the low-mode index pair  $(p_1, p_2)$  (resp.  $(q_1, q_2)$ ) relates to  $p$  (resp.  $q$ ) according to (A.4), namely where  $p$  (resp.  $q$ ) plays the role of  $k$  and  $(p_1, p_2)$  (resp.  $(q_1, q_2)$ ) that of  $(i, j)$  in (A.4).

Besides, let us recall the constant terms given in the RHS of (4.33) for the parameterization,  $\Phi_n(\tau, \boldsymbol{\beta}, \xi)$ :

$$\alpha_n(\tau) = \sum_{i,j=1}^m U_{i,j}^n(\tau, \boldsymbol{\beta})B_{i,j}^n F_i F_j - \frac{1 - e^{\tau\beta_n}}{\beta_n} F_n. \tag{A.7}$$

Thus, we rewrite the parameterization defect  $\mathcal{Q}(\tau, T)$  recalled in (A.1) as follows:

$$\begin{aligned} \mathcal{Q}_n(\tau, T) &= \mathbf{d}(\tau)^* \mathbf{Q}_4 \mathbf{d}(\tau) - 2\text{Re}(\mathbf{Q}_3^* \mathbf{d}(\tau)) + 2\text{Re}(\boldsymbol{\gamma}(\tau)^* \widetilde{\mathbf{Q}}_3 \mathbf{d}(\tau)) + \boldsymbol{\gamma}(\tau)^* \widetilde{\mathbf{Q}}_2 \boldsymbol{\gamma}(\tau) \\ &\quad - 2\text{Re}(\widehat{\mathbf{Q}}_2^* \boldsymbol{\gamma}(\tau)) + 2\text{Re}(\bar{\alpha}_n(\tau) \mathbf{Q}_2^* \mathbf{d}(\tau)) + 2\text{Re}(\bar{\alpha}_n(\tau) \mathbf{Q}_1^* \boldsymbol{\gamma}(\tau)) \\ &\quad + \langle u_n \bar{u}_n \rangle_T - 2\text{Re}(\bar{\alpha}_n(\tau) \langle u_n \rangle_T) + \alpha_n(\tau) \bar{\alpha}_n(\tau), \end{aligned} \tag{A.8}$$



where  $M^*$  denotes the conjugate transpose of a given vector or matrix  $M$ .

Note also

$$\begin{aligned} \frac{d}{d\tau} Q_n(\tau, T) = & 2\text{Re}\left( d(\tau)^* Q_4 d'(\tau) - Q_3^* d'(\tau) + y'(\tau)^* \tilde{Q}_3 d(\tau) + y(\tau)^* \tilde{Q}_3 d'(\tau) \right. \\ & + y(\tau)^* \tilde{Q}_2 y'(\tau) - \tilde{Q}_2^* y'(\tau) + \bar{\alpha}'_n(\tau) Q_2^* d(\tau) + \bar{\alpha}_n(\tau) Q_2^* d'(\tau) \\ & \left. + \bar{\alpha}'_n(\tau) Q_1^* y(\tau) + \bar{\alpha}_n(\tau) Q_1^* y'(\tau) - \bar{\alpha}'_n(\tau) \langle u_n \rangle_T + \alpha'_n(\tau) \bar{\alpha}_n(\tau) \right). \end{aligned} \tag{A.9}$$

With the above expression of  $Q_n(\tau, T)$  and of its derivative, the minimization of  $Q_n(\tau, T)$  in the  $\tau$ -variable can now be performed efficiently by application of a gradient-descent method as described in Algorithm 1. Note that if the first moments up to the 4th order are known, then the determination of  $\tau^*$  by Algorithm 1 does not require any data from direct integration of the full system. There is a vast literature about moment closure techniques and we refer to [110] for a recent survey on the topic.

---

**Algorithm 1:** Find the optimal  $\tau$  for the minimization problem (4.35) using a gradient-descent

---

**Setup:** Let  $[0, T]$  be a training interval, and  $\delta t = T/K$  with  $K > 0$ . We assume that for each  $k$  in  $\{0, \dots, K - 1\}$ , a numerical solution of Eq. (4.16) is computed, which is denoted by  $y^k$ .

**Input:** It consists of collecting the following projections of the numerical solution

$$\bigcup_{k=0, \dots, K} (u_1^k, \dots, u_m^k; u_n^k),$$

where  $u_i^k = \langle y^k, e_i^* \rangle$  for  $i = 1, \dots, m$ , and  $u_n^k = \langle y^k, e_n^* \rangle$ , with  $e_j^*$ 's denoting the generalized eigenvectors associated with  $A$  in Eq. (4.16).

**Output:** The optimal  $\tau$ -value,  $\tau^*$ , that minimizes (A.1) is obtained as follow:

- 1 Set parameter values for  $\tau$ ,  $\delta\tau$  and  $\epsilon$ , which represent respectively the initial guess of  $\tau^*$ , the initial step size of  $\tau$ , and the convergence tolerance for the iteration. For instance,

```

tau = 0;           % initial guess
delta_tau = 0.1;  % initial step size of tau
epsilon = 10^-10; % convergence tolerance
    
```

- 2 Compute  $Q_1, Q_2, \tilde{Q}_2, Q_3, \tilde{Q}_3$ , and  $Q_4$  defined in (A.6) as well as  $\langle u_n \bar{u}_n \rangle_T$  and  $\langle u_n \rangle_T$  appearing in (A.8) by using a standard numerical quadrature.

- 3 Evaluate  $Q' = \frac{d}{d\tau} Q(\tau, T)$  by using (A.9);

**while**  $|Q'| > \epsilon$  **do**

```

    Set  $\tau_\delta = \tau - \text{sgn}(Q')\delta\tau$ ;
    Compute  $Q'_\delta = \frac{d}{d\tau} Q(\tau_\delta, T)$  by using (A.9);
    if  $|Q'_\delta| > \epsilon$  and  $\text{sgn}(Q'_\delta) \neq \text{sgn}(Q')$  then
        |  $\delta\tau = \delta\tau/2$ ;
    else
        |  $\tau = \tau_\delta$ ;
        |  $Q' = Q'_\delta$ ;
    end

```

**end**

---

## References

1. Alves, J.F., Bonatti, C., Viana, M.: SRB Measures for Partially Hyperbolic Systems Whose Central Direction is Mostly Expanding, *The Theory of Chaotic Attractors*, pp. 443–490. Springer, New York (2000)
2. Armbruster, D., Guckenheimer, J., Holmes, P.: Kuramoto-Sivashinsky dynamics on the center-unstable manifold. *SIAM J. Appl. Math.* **49**(3), 676–691 (1989)
3. Arnaudon, A., De Castro, A.L., Holm, D.D.: Noise and dissipation on coadjoint orbits. *J. Nonlinear Sci.* **28**(1), 91–145 (2018)
4. Arneodo, A., Coulet, P.H., Spiegel, E.A., Tresser, C.: Asymptotic chaos. *Physica D* **14**(3), 327–347 (1985)
5. Arnold, V.I.: *Geometrical Methods in the Theory of Ordinary Differential Equations*, 2nd edn. Springer, New York (1988)
6. Baer, F., Tribbia, J.J.: On complete filtering of gravity modes through nonlinear initialization. *Mon. Weather Rev.* **105**(12), 1536–1539 (1977)
7. Berloff, P.S.: On dynamically consistent eddy fluxes. *Dyn. Atmos. Oceans* **38**(3–4), 123–146 (2005)
8. Berloff, P.S.: Random-forcing model of the mesoscale oceanic eddies. *J. Fluid Mech.* **529**, 71–95 (2005)
9. Berloff, P.: Dynamically consistent parameterization of mesoscale eddies. Part I: Simple model. *Ocean Model.* **87**, 1–19 (2015)
10. Beyn, W.-J., Kleß, W.: Numerical Taylor expansions of invariant manifolds in large dynamical systems. *Numer. Math.* **80**(1), 1–38 (1998)
11. Bibikov, Y.N.: *Local Theory of Nonlinear Analytic Ordinary Differential Equations*, Lecture Notes in Mathematics, vol. 702. Springer, New York (1979)
12. Bonatti, C., Pumarino, A., Viana, M.: Lorenz attractors with arbitrary expanding dimension. In: *Equadiff 99: (In 2 Volumes)*. World Scientific, pp. 39–44 (2000)
13. Bowen, R., Ruelle, D.: *The Ergodic Theory of Axiom A Flows, The Theory of Chaotic Attractors*, pp. 55–76. Springer, New York (1975)
14. Brézis, H.: *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York (2010)
15. Brown, H.S., Kevrekidis, I.G., Jolly, M.S.: A minimal model for spatio-temporal patterns in thin film flow. In: Aris, R., Aronson, D.G., Swinney, H.L. (eds.) *Patterns and Dynamics in Reactive Media*, pp. 11–31. Springer, New York (1991)
16. Cabré, X., Fontich, E., de la Llave, R.: The parameterization method for invariant manifolds I: manifolds associated to non-resonant subspaces. *Indiana Univ. Math. J.* **52**, 283–328 (2003)
17. Cabré, X., Fontich, E., de la Llave, R.: The parameterization method for invariant manifolds II: regularity with respect to parameters. *Indiana Univ. Math. J.* **52**, 329–360 (2003)
18. Cabré, X., Fontich, E., De La Llave, R.: The parameterization method for invariant manifolds III: overview and applications. *J. Differ. Equ.* **218**(2), 444–515 (2005)
19. Carr, J.: *Applications of Centre Manifold Theory*, Applied Mathematical Sciences, vol. 35. Springer, New York (1981)
20. Chae, D.: On the ensemble average in the study of approximate inertial manifolds, II. *J. Math. Anal. Appl.* **164**(2), 337–349 (1992)
21. Chekroun, M.D., Lamb, J.S.W., Pangerl, C.J., Rasmussen, M.: A Girsanov approach to slow parameterizing manifolds in the presence of noise (2019). [arXiv:1903.08598](https://arxiv.org/abs/1903.08598)
22. Chekroun, M.D., Liu, H.: Post-processing finite-horizon parameterizing manifolds for optimal control of nonlinear parabolic PDEs. In: 2016 IEEE 55th Conference on Decision and Control (CDC), IEEE, pp. 1411–1416 (2016)
23. Chekroun, M.D., Tantet, A., Neelin, J.D., Dijkstra, H.A.: Ruelle-Pollicott resonances of stochastic systems in reduced state space. Part I: Theory, Submitted (2019)
24. Chekroun, M.D., Glatt-Holtz, N.E.: Invariant measures for dissipative dynamical systems: abstract results and applications. *Commun. Math. Phys.* **316**(3), 723–761 (2012)
25. Chekroun, M.D., Kondrashov, D.: Data-adaptive harmonic spectra and multilayer Stuart-Landau models. *Chaos* **27**(9), 093110 (2017)
26. Chekroun, M.D., Liu, H.: Finite-horizon parameterizing manifolds, and applications to suboptimal control of nonlinear parabolic PDEs. *Acta Appl. Math.* **135**(1), 81–144 (2015)
27. Chekroun, M.D., Ghil, M., Roux, J., Varadi, F.: Averaging of time-periodic systems without a small parameter. *Discret. Contin. Dyn. Syst.* **14**, 753–782 (2006)
28. Chekroun, M.D., Simonnet, E., Ghil, M.: Stochastic climate dynamics: random attractors and time-dependent invariant measures. *Physica D* **240**(21), 1685–1700 (2011)

29. Chekroun, M.D., Neelin, J.D., Kondrashov, D., McWilliams, J.C., Ghil, M.: Rough parameter dependence in climate models: the role of Ruelle-Pollicott resonances. *Proc. Natl. Acad. Sci. USA* **111**(5), 1684–1690 (2014)
30. Chekroun, M.D., Liu, H., Wang, S.: *Approximation of Stochastic Invariant Manifolds: Stochastic Manifolds for Nonlinear SPDEs I*, Springer Briefs in Mathematics. Springer, New York (2015)
31. Chekroun, M.D., Liu, H., Wang, S.: *Stochastic Parameterizing Manifolds and Non-Markovian Reduced Equations: Stochastic Manifolds for Nonlinear SPDEs II*, Springer Briefs in Mathematics. Springer, New York (2015)
32. Chekroun, M.D., Liu, H., McWilliams, J.C.: The emergence of fast oscillations in a reduced primitive equation model and its implications for closure theories. *Comput. Fluids* **151**, 3–22 (2017)
33. Chekroun, M.D., Ghil, M., Neelin, J.D.: Pullback attractor crisis in a delay differential ENSO model. In: Tsonis, A. (ed.) *Advances in Nonlinear Geosciences*, pp. 1–33. Springer, New York (2018)
34. Chorin, A.J., Hald, O.H.: *Stochastic Tools in Mathematics and Science, Surveys and Tutorials in the Applied Mathematical Sciences*, vol. 147. Springer, New York (2006)
35. Chorin, A.J., Lu, F.: Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proc. Natl. Acad. Sci. USA* **112**(32), 9804–9809 (2015)
36. Chorin, A.J., Hald, O.H., Kupferman, R.: Optimal prediction with memory. *Physica D* **166**(3), 239–257 (2002)
37. Collet, P., Eckmann, J.-P.: *Concepts and Results in Chaotic Dynamics: A Short Course*. Springer, New York (2007)
38. Constantin, P., Foias, C., Nicolaenko, B., Temam, R.: *Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations*, Applied Mathematical Sciences, vol. 70. Springer, New York (1989)
39. Cotter, C., Crisan, D., Holm, D.D., Pan, W., Shevchenko, I.: Numerically modeling stochastic lie transport in fluid dynamics. *Multiscale Model. Simul.* **17**(1), 192–232 (2019)
40. Couillet, P.H., Spiegel, E.: Amplitude equations for systems with competing instabilities. *SIAM J. Appl. Math.* **43**(4), 776–821 (1983)
41. Cox, S.M., Matthews, P.C.: Exponential time differencing for stiff systems. *J. Comput. Phys.* **176**(2), 430–455 (2002)
42. Crawford, J.D.: Introduction to bifurcation theory. *Rev. Mod. Phys.* **63**(4), 991 (1991)
43. Crommelin, D.T., Majda, A.J.: Strategies for model reduction: comparing different optimal bases. *J. Atmos. Sci.* **61**(17), 2206–2217 (2004)
44. Crommelin, D., Vanden-Eijnden, E.: Subgrid-scale parameterization with conditional markov chains. *J. Atmos. Sci.* **65**(8), 2661–2675 (2008)
45. Daley, R.: Variational non-linear normal mode initialization. *Tellus* **30**(3), 201–218 (1978)
46. Daley, R.: The development of efficient time integration schemes using model normal modes. *Mon. Weather Rev.* **108**(1), 100–110 (1980)
47. Daley, R.: *Atmospheric Data Analysis*. Cambridge University Press, Cambridge (1993)
48. Debussche, A., Marion, M.: On the construction of families of approximate inertial manifolds. *J. Differ. Equ.* **100**(1), 173–201 (1992)
49. Debussche, A., Temam, R.: Inertial manifolds and the slow manifolds in meteorology. *Differ. Integr. Equ.* **4**(5), 897–931 (1991)
50. Debussche, A., Dubois, T., Temam, R.: The nonlinear Galerkin method: a multiscale method applied to the simulation of homogeneous turbulent flows. *Theor. Comput. Fluid Dyn.* **7**(4), 279–315 (1995)
51. Dellacherie, C., Meyer, P.-A.: *Probabilities and Potential*, North-Holland Mathematics Studies, vol. 29. North-Holland Publishing Co., Amsterdam (1978)
52. Devulder, C., Marion, M., Titi, E.S.: On the rate of convergence of the nonlinear Galerkin methods. *Math. Comput.* **60**(202), 495–514 (1993)
53. Dijkstra, H.A., Ghil, M.: Low-frequency variability of the large-scale ocean circulation: a dynamical systems approach. *Rev. Geophys.* **43**(3), RG3002 (2005)
54. Dubois, T., Jauberteau, F.: A dynamic multilevel model for the simulation of the small structures in homogeneous isotropic turbulence. *J. Sci. Comput.* **13**(3), 323–367 (1998)
55. Dubois, T., Jauberteau, F., Marion, M., Temam, R.: Subgrid modelling and the interaction of small and large wavelengths in turbulent flows. *Comput. Phys. Commun.* **65**(1–3), 100–106 (1991)
56. Dubois, T., Jauberteau, F., Temam, R.: Incremental unknowns, multilevel methods and the numerical simulation of turbulence. *Comput. Methods Appl. Mech. Eng.* **159**(1–2), 123–189 (1998)
57. Eckmann, J.-P., Ruelle, D.: Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.* **57**, 617–656 (1985)
58. Eirola, T., von Pfaler, J.: Numerical Taylor expansions for invariant manifolds. *Numer. Math.* **99**(1), 25–46 (2004)

59. Elphick, E., Tirapegui, C., Brachet, M.E., Couillet, P., Iooss, G.: A simple global characterization for normal forms of singular vector fields. *Physica D* **29**(1–2), 95–127 (1987)
60. Faria, T.: Normal forms and bifurcations for delay differential equations. In: *Delay Differential Equations and Applications*, NATO Sci. Ser. II Math. Phys. Chem., vol. 205. Springer, Dordrecht, pp. 227–282 (2006)
61. Feppon, F., Lermusiaux, P.F.J.: A geometric approach to dynamical model order reduction. *SIAM J. Matrix Anal. Appl.* **39**(1), 510–538 (2018)
62. Foias, C., Sell, G.R., Temam, R.: Variétés inertielles des équations différentielles dissipatives. *C. R. Acad. Sci. Paris Série I* **301**(5), 139–142 (1985)
63. Foias, C., Jolly, M.S., Kevrekidis, I.G., Sell, G.R., Titi, E.S.: On the computation of inertial manifolds. *Phys. Lett. A* **131**(7–8), 433–436 (1988)
64. Foias, C., Manley, O., Temam, R.: Modeling of the interaction of small and large eddies in two-dimensional turbulent flows. *RAIRO Modél. Math. Anal. Numér.* **22**(1), 93–118 (1988)
65. Foias, C., Nicolaenko, B., Sell, G.R., Temam, R.: Inertial manifolds for the Kuramoto Sivashinsky equation and an estimate of their lowest dimension. *J. Math. Pure. Appl.* **67**, 197–226 (1988)
66. Foias, C., Sell, G.R., Temam, R.: Inertial manifolds for nonlinear evolutionary equations. *J. Differ. Equ.* **73**(2), 309–353 (1988)
67. Foias, C., Sell, G.R., Titi, E.S.: Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations. *J. Dyn. Diff. Equ.* **1**(2), 199–244 (1989)
68. Foias, C., Manley, O.P., Temam, R.: Approximate inertial manifolds and effective viscosity in turbulent flows. *Phys. Fluids A* **3**(5), 898–911 (1991)
69. Foias, C., Manley, O., Rosa, R., Temam, R.: *Navier-Stokes Equations and Turbulence*. Encyclopedia of Mathematics and Its Applications, vol. 83. Cambridge University Press, Cambridge (2001)
70. Frederiksen, J.S., Kepert, S.M.: Dynamical subgrid-scale parameterizations from direct numerical simulations. *J. Atmos. Sci.* **63**(11), 3006–3019 (2006)
71. Gallavotti, G., Cohen, E.G.D.: Dynamical ensembles in stationary states. *J. Stat. Phys.* **80**(5–6), 931–970 (1995)
72. García-Archilla, B., de Frutos, J.: Time integration of the non-linear Galerkin method. *IMA J. Numer. Anal.* **15**(2), 221–224 (1995)
73. Gent, P.R., McWilliams, J.C.: Intermediate model solutions to the Lorenz equations: strange attractors and other phenomena. *J. Atmos. Sci.* **39**(1), 3–13 (1982)
74. Ghil, M., Malanotte-Rizzoli, P.: *Data Assimilation in Meteorology and Oceanography*, Advances in Geophysics, vol. 33, pp. 141–266. Elsevier, Amsterdam (1991)
75. Ghil, M., Chekroun, M.D., Simonnet, E.: Climate dynamics and fluid mechanics: natural variability and related uncertainties. *Physica D* **237**(14–17), 2111–2126 (2008)
76. Givon, D., Kupferman, R., Stuart, A.: Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity* **17**(6), R55–R127 (2004)
77. Golubitsky, M., Schaeffer, D.G.: *Singularities and Groups in Bifurcation Theory*, vol. 1. Springer, New York (1985)
78. Gottwald, G.A., Peters, K., Davies, L.: A data-driven method for the stochastic parametrisation of subgrid-scale tropical convective area fraction. *Q. J. R. Meteor. Soc.* **142**(694), 349–359 (2016)
79. Gottwald, G.A., Crommelin, D.T., Franzke, C.L.E.: Stochastic climate theory. In: Franzke, C.L.E., O’Kane, T.J. (eds.) *Nonlinear and Stochastic Climate Dynamics*, pp. 209–240. Cambridge University Press, Cambridge (2017)
80. Graham, M.D., Steen, P.H., Titi, E.S.: Computational efficiency and approximate inertial manifolds for a Bénard convection system. *J. Nonlinear Sci.* **3**(1), 153–167 (1993)
81. Guckenheimer, J., Holmes, P.: *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Applied Mathematical Sciences, vol. 42. Springer, New York (1990)
82. Hald, O.H., Stinis, P.: Optimal prediction and the rate of decay for solutions of the euler equations in two and three dimensions. *Proc. Natl. Acad. Sci. USA* **104**(16), 6527–6532 (2007)
83. Haragus, M., Iooss, G.: *Local Bifurcations, Center Manifolds, and Normal Forms in Infinite-Dimensional Dynamical Systems*, Universitext. Springer, London (2011)
84. Haro, À.: Automatic differentiation methods in computational dynamical systems: invariant manifolds and normal forms of vector fields at fixed points, IMA Note (2008)
85. Haro, À., Canadell, M., Figueras, J.-L., Luque, A., Mondelo, J.-M.: *The Parameterization Method for Invariant Manifolds: From Rigorous Results to Effective Computations*, vol. 195. Springer, Berlin (2016)
86. Hasselmann, K.: PIPs and POPs: the reduction of complex dynamical systems using principal interaction and oscillation patterns. *J. Geophys. Res.* **93**(D9), 11015–11021 (1988)
87. Haugen, J.E., Machenhauer, B.: A spectral limited-area model formulation with time-dependent boundary conditions applied to the shallow-water equations. *Mon. Weather Rev.* **121**(9), 2618–2630 (1993)


88. Henry, D.: Geometric Theory of Semilinear Parabolic Equations, Lecture Notes in Mathematics, vol. 840. Springer, Berlin (1981)
89. Herring, J.R., Kraichnan, R.H.: Comparison of some approximations for isotropic turbulence. In: Ehlers, J., Hepp, K., Weidenmuller, H.A. (eds.) Statistical Models and Turbulence, Lecture Notes in Physics. Springer, pp. 148–194 (1972)
90. Heywood, J.G., Rannacher, R.: On the question of turbulence modeling by approximate inertial manifolds and the nonlinear Galerkin method. *SIAM J. Numer. Anal.* **30**(6), 1603–1621 (1993)
91. Holm, D.D.: Variational principles for stochastic fluid dynamics. *Proc. R. Soc. A* **471**(2176), 20140963 (2015)
92. Holmes, P., Lumley, J.L., Berkooz, G., Rowley, C.W.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry, 2nd edn. Cambridge University Press, Cambridge (2012)
93. Hopf, E.: A mathematical example displaying features of turbulence. *Commun. Pure Appl. Math.* **1**(4), 303–322 (1948)
94. Jansen, M.F., Held, I.M.: Parameterizing subgrid-scale eddy effects using energetically consistent backscatter. *Ocean Model.* **80**, 36–48 (2014)
95. Jolly, M.S.: Bifurcation computations on an approximate inertial manifold for the 2D Navier-Stokes equations. *Physica D* **63**(1–2), 8–20 (1993)
96. Jolly, M.S., Kevrekidis, I.G., Titi, E.S.: Approximate inertial manifolds for the Kuramoto-Sivashinsky equation: analysis and computations. *Physica D* **44**(1), 38–60 (1990)
97. Jolly, M.S., Kevrekidis, I.G., Titi, E.S.: Preserving dissipation in approximate inertial forms for the Kuramoto-Sivashinsky equation. *J. Dyn. Differ. Equ.* **3**(2), 179–197 (1991)
98. Jones, D.A., Titi, E.S.: A remark on quasi-stationary approximate inertial manifolds for the Navier-Stokes equations. *SIAM J. Math. Anal.* **25**(3), 894–914 (1994)
99. Kassam, A., Trefethen, L.N.: Fourth-order time-stepping for stiff PDEs. *SIAM J. Sci. Comput.* **26**(4), 1214–1233 (2005)
100. Kifer, Y.: Averaging and climate models. In: Imkeller, P., von Storch, J.-S. (eds.) Stochastic Climate Models, pp. 171–188. Springer, New York (2001)
101. Kifer, Y.: Another proof of the averaging principle for fully coupled dynamical systems with hyperbolic fast motions. *Discret. Contin. Dyn. Syst.* **13**(5), 1187–1201 (2005)
102. Kondrashov, D., Chekroun, M.D., Ghil, M.: Data-driven non-Markovian closure models. *Physica D* **297**, 33–55 (2015)
103. Kondrashov, D., Chekroun, M.D., Ghil, M.: Data-adaptive harmonic decomposition and prediction of Arctic sea ice extent. *Dyn. Stat. Clim. Syst.* **3**(1), 1–23 (2018)
104. Kondrashov, D., Chekroun, M.D., Yuan, X., Ghil, M.: Data-adaptive harmonic decomposition and stochastic modeling of Arctic sea ice. In: Tsonis, A. (ed.) Advances in Nonlinear Geosciences, pp. 179–205. Springer, New York (2018)
105. Kondrashov, D., Chekroun, M.D., Berloff, P.: Multiscale Stuart-Landau emulators: application to wind-driven ocean gyres. *Fluids* **3**, 21 (2018)
106. Kraichnan, R.H.: The structure of isotropic turbulence at very high Reynolds numbers. *J. Fluid Mech.* **5**(4), 497–543 (1959)
107. Kraichnan, R.H.: Approximations for steady-state isotropic turbulence. *Phys. Fluids* **7**(8), 1163–1168 (1964)
108. Kraichnan, R.H.: Eddy viscosity in two and three dimensions. *J. Atmos. Sci.* **33**(8), 1521–1536 (1976)
109. Kuehn, C.: Multiple Time Scale Dynamics, vol. 191. Springer, New York (2015)
110. Kuehn, C.: Moment closure—a brief review. In: Scholl, E., Klapp, S.H.L., Hovel, P. (eds.) Control of Self-organizing Nonlinear Systems, pp. 253–271. Springer, New York (2016)
111. Kuramoto, Y., Tsuzuki, T.: Persistent propagation of concentration waves in dissipative media far from thermal equilibrium. *Prog. Theor. Phys.* **55**(2), 356–369 (1976)
112. Kwasniok, F.: The reduction of complex dynamical systems using principal interaction patterns. *Physica D* **92**(1–2), 28–60 (1996)
113. Kwasniok, F.: Optimal Galerkin approximations of partial differential equations using principal interaction patterns. *Phys. Rev. E* **55**(5), 5365 (1997)
114. Kwasniok, F.: Empirical low-order models of barotropic flow. *J. Atmos. Sci.* **61**(2), 235–245 (2004)
115. Kwasniok, F.: Reduced atmospheric models using dynamically motivated basis functions. *J. Atmos. Sci.* **64**(10), 3452–3474 (2007)
116. Kwasniok, F.: Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling. *Philos. Trans. R. Soc. A* **370**(1962), 1061–1086 (2012)
117. Landau, L.D., Lifshits, E.M.: Fluid Mechanics. Pergamon Press, Oxford (1959)
118. Langford, W.F.: Periodic and steady-state mode interactions lead to tori. *SIAM J. Appl. Math.* **37**(1), 22–48 (1979)

119. Lebedez, D., Siehr, J., Unger, J.: A variational principle for computing slow invariant manifolds in dissipative dynamical systems. *SIAM J. Sci. Comput.* **33**(2), 703–720 (2011)
120. Leith, C.E.: Nonlinear normal mode initialization and quasi-geostrophic theory. *J. Atmos. Sci.* **37**(5), 958–968 (1980)
121. Leith, C.E.: Stochastic backscatter in a subgrid-scale model: plane shear mixing layer. *Phys. Fluids A* **2**(3), 297–299 (1990)
122. Lorenz, E.N.: Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**(2), 130–141 (1963)
123. Lorenz, E.N.: Attractor sets and quasi-geostrophic equilibrium. *J. Atmos. Sci.* **37**(8), 1685–1699 (1980)
124. Lu, F., Lin, K.K., Chorin, A.J.: Data-based stochastic model reduction for the Kuramoto-Sivashinsky equation. *Physica D* **340**, 46–57 (2017)
125. Ma, T., Wang, S.: *Bifurcation Theory and Applications*, World Scientific Series on Nonlinear Science. Series A: Monographs and Treatises, vol. 53. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ (2005)
126. Ma, T., Wang, S.: *Phase Transition Dynamics*. Springer, New York (2014)
127. Machenhauer, B.: On the dynamics of gravity oscillations in a shallow water model with applications to normal mode initialization. *Beitr. Phys. Atmos.* **50**, 253–271 (1977)
128. Majda, A.J., Timofeyev, I., Vanden-Eijnden, E.: A mathematical framework for stochastic climate models. *Commun. Pure Appl. Math.* **54**(8), 891–974 (2001)
129. Majda, A.J., Timofeyev, I., Vanden-Eijnden, E.: Systematic strategies for stochastic mode reduction in climate. *J. Atmos. Sci.* **60**(14), 1705–1722 (2003)
130. Mallet-Paret, J., Sell, G.R.: Inertial manifolds for reaction diffusion equations in higher space dimensions. *J. Am. Math. Soc.* **1**(4), 805–866 (1988)
131. Marion, M., Temam, R.: Nonlinear Galerkin methods. *SIAM J. Numer. Anal.* **26**(5), 1139–1157 (1989)
132. McComb, W.D., Hunter, A., Johnston, C.: Conditional mode-elimination and the subgrid-modeling problem for isotropic turbulence. *Phys. Fluids* **13**(7), 2030–2044 (2001)
133. McWilliams, J.C.: The elemental shear dynamo. *J. Fluid Mech.* **699**, 414–452 (2012)
134. Mori, H.: Transport, collective motion, and brownian motion. *Prog. Theor. Phys.* **33**(3), 423–455 (1965)
135. Parish, E., Duraisamy, K.: Reduced order modeling of turbulent flows using statistical coarse-graining. In: 46th AIAA Fluid Dynamics Conference, p. 3640 (2016)
136. Parish, E.J., Duraisamy, K.: A dynamic subgrid scale model for Large Eddy Simulations based on the Mori-Zwanzig formalism. *J. Comput. Phys.* **349**, 154–175 (2017)
137. Pascal, F., Basdevant, C.: Nonlinear Galerkin method and subgrid-scale model for two-dimensional turbulent flows. *Theor. Comput. Fluid Dyn.* **3**(5), 267–284 (1992)
138. Pavliotis, G., Stuart, A.: *Multiscale Methods: Averaging and Homogenization*, vol. 53. Springer, New York (2008)
139. Penland, C., Magorian, T.: Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *J. Clim.* **6**, 1067–1076 (1993)
140. Piomelli, U., Cabot, W.H., Moin, P., Lee, S.: Subgrid-scale backscatter in turbulent and transitional flows. *Phys. Fluids A* **3**(7), 1766–1771 (1991)
141. Pomeau, Y., Pumir, A., Pelce, P.: Intrinsic stochasticity with many degrees of freedom. *J. Stat. Phys.* **37**(1–2), 39–49 (1984)
142. Porta Mana, P., Zanna, L.: Toward a stochastic parameterization of ocean mesoscale eddies. *Ocean Model.* **79**, 1–20 (2014)
143. Pötsche, C., Rasmussen, M.: Taylor approximation of integral manifolds. *J. Dyn. Diff. Equ.* **18**(2), 427–460 (2006)
144. Pötsche, C., Rasmussen, M.: Computation of nonautonomous invariant and inertial manifolds. *Numer. Math.* **112**(3), 449–483 (2009)
145. Reiterer, P., Lainscsek, C., Schürer, F., Letellier, C., Maquet, J.: A nine-dimensional Lorenz system to study high-dimensional chaos. *J. Phys. A* **31**, 7121–7139 (1998)
146. Resseguier, V., Mémin, E., Chapron, B.: Geophysical flows under location uncertainty, Part I: random transport and general models. *Geophys. Astrophys. Fluid Dyn.* **111**(3), 149–176 (2017)
147. Resseguier, V., Mémin, E., Chapron, B.: Geophysical flows under location uncertainty, Part II: Quasi-geostrophy and efficient ensemble spreading. *Geophys. Astrophys. Fluid Dyn.* **111**(3), 177–208 (2017)
148. Resseguier, V., Mémin, E., Chapron, B.: Geophysical flows under location uncertainty, Part III: SQG and frontal dynamics under strong turbulence conditions. *Geophys. Astrophys. Fluid Dyn.* **111**(3), 209–227 (2017)
149. Robinson, J.C.: Inertial manifolds for the Kuramoto-Sivashinsky equation. *Phys. Lett. A* **184**(2), 190–193 (1994)
150. Rowley, C.W., Mezić, I., Bagheri, S., Schlatter, P., Henningson, D.S.: Spectral analysis of nonlinear flows. *J. Fluid Mech.* **641**, 115–127 (2009)

151. Ruelle, D., Takens, F.: On the nature of turbulence. *Commun. Math. Phys.* **20**(3), 167–192 (1971)
152. Sapsis, T.P., Dijkstra, H.A.: Interaction of additive noise and nonlinear dynamics in the double-gyre wind-driven ocean circulation. *J. Phys. Oceanogr.* **43**(2), 366–381 (2013)
153. Sapsis, T.P., Lermusiaux, P.F.J.: Dynamically orthogonal field equations for continuous stochastic dynamical systems. *Physica D* **238**, 2347–2360 (2009)
154. Sapsis, T.P., Lermusiaux, P.F.J.: Dynamical criteria for the evolution of the stochastic dimensionality in flows with uncertainty. *Physica D* **241**, 60–76 (2012)
155. Schmid, P.J.: Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.* **656**, 5–28 (2010)
156. Schmuck, M., Pradas, M., Pavliotis, G.A., Kalliadasis, S.: A new mode reduction strategy for the generalized Kuramoto-Sivashinsky equation. *IMA J. Appl. Math.* **80**(2), 273–301 (2015)
157. Sivashinsky, G.I.: Nonlinear analysis of hydrodynamic instability in laminar flames-I. Derivation of basic equations. *Acta Astronaut* **4**(11–12), 1177–1206 (1977)
158. Stinis, P.: Stochastic optimal prediction for the Kuramoto-Sivashinsky equation. *Multiscale Model. Simul.* **2**(4), 580–612 (2004)
159. Stinis, P.: Higher-order Mori-Zwanzig models for the Euler equations. *Multiscale Model. Simul.* **6**(3), 741–760 (2007)
160. Subramani, D.N.: Probabilistic regional ocean predictions: stochastic fields and optimal planning, Ph.D. thesis, Massachusetts Institute of Technology (2018)
161. Taira, K., Brunton, S.L., Dawson, S.T.M., Rowley, C.W., Colonius, T., McKeon, B.J., Schmidt, O.T., Gordeyev, S., Theofilis, V., Ukeiley, L.S.: Modal analysis of fluid flows: an overview. *AIAA J.* **55**, 4013–4041 (2017)
162. Temam, R.: Variétés inertielles approximatives pour les équations de Navier-Stokes bidimensionnelles. *C. R. Acad. Sci. Paris Série II* **306**, 349–402 (1988)
163. Temam, R.: Attractors for the Navier-Stokes equations, localization and approximation. *J. Fac. Sci. Univ. Tokyo. Soc. IA Math.* **36**, 629–647 (1989)
164. Temam, R.: *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Applied Mathematical Sciences. Springer, New York (1997)
165. Temam, R., Wang, X.: Estimates on the lowest dimension of inertial manifolds for the Kuramoto-Sivashinsky equation in the general case. *Differ. Integr. Equ.* **7**(3–4), 1095–1108 (1994)
166. Titi, E.S.: On approximate inertial manifolds to the Navier-Stokes equations. *J. Math. Anal. Appl.* **149**(2), 540–557 (1990)
167. Tribbia, J.J.: Nonlinear initialization on an equatorial beta-plane. *Mon. Weather Rev.* **107**(6), 704–713 (1979)
168. Tribbia, J.J.: On variational normal mode initialization. *Mon. Weather Rev.* **110**(6), 455–470 (1982)
169. Tu, J.H., Rowley, C.W., Luchtenburg, D.M., Brunton, S.L., Kutz, J.N.: On dynamic mode decomposition: theory and applications. *J. Comput. Dyn.* **1**, 391–421 (2014)
170. Ueckermann, M.P., Lermusiaux, P.F.J., Sapsis, T.P.: Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. *J. Comput. Phys.* **233**, 272–294 (2013)
171. Vanden-Eijnden, E.: Numerical techniques for multi-scale dynamical systems with stochastic effects. *Commun. Math. Sci.* **1**(2), 385–391 (2003)
172. Vanderbauwhede, A.: *Centre Manifolds, Normal Forms and Elementary Bifurcations*, Dynamics Reported, pp. 89–169. Springer, New York (1989)
173. Williams, M.O., Kevrekidis, I.G., Rowley, C.W.: A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. *J. Nonlinear Sci.* **25**(6), 1307–1346 (2015)
174. Wittenberg, R.W., Holmes, P.: Scale and space localization in the Kuramoto-Sivashinsky equation. *Chaos* **9**(2), 452–465 (1999)
175. Wouters, J., Lucarini, V.: Disentangling multi-level systems: averaging, correlations and memory. *J. Stat. Mech.* **2012**, P03003 (2012)
176. Wouters, J., Lucarini, V.: Multi-level dynamical systems: connecting the Ruelle response theory and the Mori-Zwanzig approach. *J. Stat. Phys.* **151**(5), 850–860 (2013)
177. Young, L.-S.: What are SRB measures, and which dynamical systems have them? *J. Stat. Phys.* **108**(5), 733–754 (2002)
178. Young, L.-S.: Generalizations of SRB measures to nonautonomous, random, and infinite dimensional systems. *J. Stat. Phys.* **166**, 494–515 (2016)
179. Zanna, L., Porta Mana, P., Anstey, J., David, T., Bolton, T.: Scale-aware deterministic and stochastic parametrizations of eddy-mean flow interaction. *Ocean Model.* **111**, 66–80 (2017)
180. Zelik, S.: Inertial manifolds and finite-dimensional reduction for dissipative PDEs. *Proc. R. Soc. Edinburgh Sect. A* **144**(6), 1245–1327 (2014)
181. Zwanzig, R.: *Nonequilibrium Statistical Mechanics*. Oxford University Press, Oxford (2001)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

**Mickaël D. Chekroun**<sup>1,2</sup>  · **Honghu Liu**<sup>3</sup> · **James C. McWilliams**<sup>2</sup>

✉ Mickaël D. Chekroun  
mchekroun@atmos.ucla.edu

Honghu Liu  
hhliu@vt.edu

James C. McWilliams  
jcm@atmos.ucla.edu

<sup>1</sup> Department of Earth and Planetary Sciences, Weizmann Institute, 76100 Rehovot, Israel

<sup>2</sup> Department of Atmospheric and Oceanic Sciences and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, USA

<sup>3</sup> Department of Mathematics, Virginia Tech, Blacksburg, VA 24061, USA