Verification and accuracy check of simulations with PoPe and iPoPe

Thomas Cartier-Michaud, Philippe Ghendrih, Virginie Grandgirard, Eric Serre

Verification and accuracy check of simulations with PoPe and iPoPe

Thomas CARTIER-MICHAUD[(a,b)] Philippe GHENDRIH[(a)], Virginie GRANDGIRARD[(a)], Eric SERRE[(b)]

a CEA, IRFM, F-13108 Saint-Paul-lez-Durance, France.

b Aix Marseille Univ, CNRS, Centrale Marseille, M2P2, Marseille, France.

Email: philippe.ghendrih@cea.fr

September 8, 2022

Abstract

The theoretical background of the PoPe and iPoPe verification scheme is presented. Verification is performed using the simulation output of production runs. The computing overhead is estimated to be at most 10%. PoPe or iPoPe calculations can be done offline provided the necessary data is stored, for example additional time slices, or online where iPoPe is more effective. The computing overhead is mostly that of storing the necessary data. The numerical error is determined and split into a part proportional to the operators, which are combined to form the equations to be solved, thus modifying their control parameters, completed by a residual error orthogonal to these operators. The accuracy of the numerical solution is determined by this modification of the control parameters. The PoPe and iPoPe methods are illustrated in this paper with simulations of a simple mechanical system with chaotic trajectories evolving into a strange attractor with sensitivity to initial conditions. We show that the accuracy depends on the particular simulation both because the properties of the numerical solution depend on the values of the control parameter, and because the target accuracy will depend on the problem that is addressed. One shows that for a case close to bifurcations between different states, the accuracy is determined by the level of detail of the bifurcation phenomena one aims at describing. A unique verification index, the PoPe index, is proposed to characterise the accuracy, and consequently the verification, of each production run. The PoPe output allows one to step beyond verification and analyse for example the numerical scheme efficiency. For the chosen example at fixed PoPe index, therefore at fixed numerical error, one finds that the higher order integration scheme, comparing order 4 to order 2 Runge-Kutta time stepping, reduces the computation cost by a factor 4.

Contents

1 Introduction

As numerical simulations are becoming increasingly important, and as the resources dedicated to these simulations have grown to be quite significant, reliability of this novel material is a concern, and consequently verification of the numerical tools is mandatory. This issue is not new and has always been considered, but it is more challenging and time consuming as the numerical tools and problems to be addressed are more complex. Accuracy estimate of the numerical output must also be regarded as a need when discussing reliability (rather than solution verification that is very challenging). The PoPe method, standing for Projection on Proper elements, has been designed in particular to address this problem for any simulation, without modifying the code or numerical problem, and only relying on the code output. The purpose of this paper is a general presentation of this method, completing published papers addressing verification of plasma turbulence codes with PoPe [9, 8, 10, 7]. To illustrate the properties of PoPe, a simple problem that can readily be coded as been chosen. However, being non-linear and chaotic, hence generic of many problems that require simulations, it is also challenging for standard verification. The issue of solution verification, which remains very challenging, will also be addressed in the paper.

When available, analytic solutions of the problem at hand are most useful for verification purposes. However, with the growing complexity of the problems to be addressed, numerically built solutions are implemented. The best known method with numerically built solution is the so-called Method of Manufactured Solutions (usually referred to as MMS) [22, 20, 19]. It is regarded as state of the art method to address complex code verification and is now used for fusion plasma simulation tools [21, 26, 24]. The MMS method, however elegant, suffers from two main drawbacks. First, the MMS requires that one modifies the code to enforce that a chosen function is the effective solution and furthermore, for an appropriate assessment of the accuracy, one must chose a function that is representative of the simulations to be performed. Second, one is led to select a stable target solution, which can prove quite restrictive. The example chosen to illustrate this paper is chaotic, hence unstable, and does not fit this requirement. We first propose an alternative verification scheme, akin to the MMS, which does not depend on a particular target solution, and that we have called by analogy Method of Reverse Solution (MRS). The latter also requires less modifications of the code but still depends on the particular time slots retained in the verification procedure. The MMS is recalled in Appendix A.1 and the alternative MRS verification scheme is presented in Section 3.2.

A particularly useful verification procedure should be available for each production run of the simulation effort, and ideally would provide a figure of merit of the exactness and accuracy of any particular simulation. The PoPe method, Projection on Proper elements, has been developed to achieve this task and to investigate the performance of reduced models [9, 8]. It has been used to analyse the exactness and accuracy of existing simulations [10, 7]. Although PoPe and the simplified iPoPe method are based on data analysis, they follow a defined mathematical procedure because the errors to be computed and the way to compute them are defined. This differs from advanced big data analysis based on artificial intelligence routines [3] that can address a larger class of problems but lead to results with a different controllability and understanding. Both PoPe and iPoPe, beyond verification check and accuracy tests, provide a means to further analyse the numerical scheme, identify the operators that are responsible for most of the numerical error as well

as the operators that play a leading role on the behaviour of the particular simulation. The PoPe method is presented in Section 2 together with a simplified alternative method that we have named iPoPe for independent Projection on Proper elements. We consider the equations to be solved as a linear function of the control parameters, weighted by values of an ensemble of operators, the right hand side operators, and yielding values of one particular operator, the left hand side operator, usually the time derivative, distinct from those of the ensemble. The key idea is to build numerically the values of all operators using the simulation output and project the left hand side operator on the ensemble of operators. This yields a linear system determining the values of the control parameters consistent with the simulation output and that can be compared to those assumed for the particular simulation. The problem and simulations used to illustrate this paper are presented in Section 3.1. We have chosen a simple mechanical system, namely a compass driven by an alternating magnetic field and subject to viscous damping. The trajectory in the 2D phase space is chaotic and, for non-vanishing damping, exhibits an attractor, called strange attractor, with fractal dimension ranging between 0 and 2 depending on the values of the control parameters. Chaos being generic in non-linear systems, verification methods must be able to handle such dynamics. The MRS verification procedure is used as reference verification and accuracy check for the chosen strange attractor in Section 3.2. Two appendices, one dedicated to the MMS method, Appendix A, and one addressing the scaling of the error with the MRS method, Appendix B, complete this first part. The PoPe and iPoPe verification schemes applied to the case of the strange attractor are presented and evaluated in Section 4. Finally, Section 5 dedicated to discussion and conclusion closes the paper.

2 PoPe and iPoPe verification

2.1 PoPe analysis

The aim of the PoPe verification scheme is to analyse the exactness of a particular simulation using the output data of that particular simulation. The standard simulation overhead when using this method is mostly to save more data than one would normally consider for a production run. For example, saving additional time slices for high order finite difference calculation of the time derivative or data at all mesh point to compute the phase space derivatives when only a fraction is saved routinely. Most of the work is then postponed to the post-processing stage. The weight of this additional output can also be optimised as will be discussed in the following. The alternative, with verification on the fly, leads to a computing overhead but with the benefit of immediate verification and accuracy estimate of the simulation.

The problem addressed numerically can most of the time be written in the following mathematical generic form:

$$O_t^{(m)} - \sum_{k=1}^{K} O_k^{(m)} = 0 \tag{1}$$

where $O_k^{(m)}$ are the various operators that are added to yield $O_t^{(m)}$. The superscript (m) refers to the mathematical equation, while in the following the superscript (n) will re-

fer to the mathematical approximation to be solved numerically and (s) for the actual simulation realisation. In this form, the control parameters that specify the weight of the different physical processes involved in Eq.(1) are included in the definition of the operators. The reference weight of each operator in Eq.(1) is unity. A form with an explicit dependence on the control parameters does not change the PoPe analysis. The weights of some of the operators are then the values of the control parameters as in [8]. This is only a matter of presentation.

The problems solved numerically often take the form of an evolution, the operator O_t then stands for a time derivative, hence the label t, governed by several effects characterised by the right hand side operators O_k. We present PoPe in this rather standard framework but the procedure holds to any problem of the form Eq.(1). It is important to underline that the PoPe method is very versatile and the choice and definition of the operators is not constrained. In a standard way one follows the way these operators are generated by the underlying physics, hence the choice of the time derivative for the operator O_t. However, this is by no means mandatory. For instance, for a system converging towards steady state with vanishing time derivative, one will want to avoid singularities and then use another operator instead of the time derivative to define O_t. Implementing the PoPe method, mostly in the post-processing stage, will clearly benefit from any insight into the processes that govern the simulation at hand.

In order to perform numerical simulations, Eq.(1) is transformed by discretising the operators.

$$O_t^{(n)} - \sum_{k=1}^{K} O_k^{(n)} = 0 \tag{2a}$$

This step introduces a first set of approximations and consequently of errors, that can be a priori determined. The two equations Eq.(1) and Eq.(2a) cannot hold together, although one can enforce that the two equations exhibit the same symmetries and thus the same conservation laws. When addressing the problem numerically, Eq.(2a) is to be solved so that Eq.(1) is only solved approximately. One can then rewrite this equation as:

$$O_t^{(m)} - \sum_{k=1}^{K} O_k^{(m)} = O_t^{(n)} - \sum_{k=1}^{K} O_k^{(n)} + E^{(n)} \tag{2b}$$

The system addressed numerically Eq.(2a) departs from the target mathematical system Eq.(1). The numerical simulation itself further contributes to the error build-up via the rounding errors, as well as possible errors in the implementation. The effective equation of a given simulation is then:

$$O_t^{(s)} - \sum_{k=1}^{K} O_k^{(s)} = 0 \tag{3a}$$

Compared to the previous forms of the equations, Eq.(1) and Eq.(2a), numerical noise governs the departure of the operators $O_t^{(s)}$ and $O_k^{(s)}$ from that implemented in the code,

$O_t^{(n)}$ and $O_k^{(n)}$. An error is therefore generated at this step and the form of Eq.(3a) to be addressed is therefore:

$$O_t^{(n)} - \sum_{k=1}^{K} O_k^{(n)} = O_t^{(s)} - \sum_{k=1}^{K} O_k^{(s)} + E^{(s)} \tag{3b}$$

We find therefore that the equation that is consistent with the output data departs from that considered initially due to errors with a known, potentially complicated error, $E^{(n)}$, and an error $E^{(s)}$, which is simulation dependent and not controlled, therefore unknown. The initial mathematical problem Eq.(1) has therefore been changed in the simulation process.

$$O_t^{(m)} - \sum_{k=1}^{K} O_k^{(m)} = E^{(n)} + E^{(s)} \tag{4}$$

We now consider a data driven approach and assume that the simulation has been performed so that the operators can be reconstructed using the output data. One then has the relationship:

$$O_t^{(r)} - \sum_{k=1}^{K} O_k^{(r)} = E^{(r)} \tag{5}$$

where the superscript (r) now identifies the reconstructed operators. In the latter equation the operators $O_t^{(r)}$ and $O_k^{(r)}$ are computed using the output data so that $E^{(r)}$ can also be computed and is therefore known for the specific simulation and according to the specific data saving process. Given Eq.(5), one can also write this equation as:

$$O_t^{(m)} - \sum_{k=1}^{K} O_k^{(m)} = E^{(r)} + \delta E^{(r)} \tag{6a}$$

$$\delta E^{(r)} = \delta O_t^{(r)} - \sum_{k=1}^{K} \delta O_k^{(r)} \tag{6b}$$

$$\delta O_t^{(r)} = O_t^{(m)} - O_t^{(r)} \tag{6c}$$

$$\delta O_k^{(r)} = O_k^{(m)} - O_k^{(r)} \tag{6d}$$

This system is not closed because the error $\delta E^{(r)}$ Eq.(6b) is not determined and depends on the departure between the reconstructed operators $O^{(r)}$ and the target mathematical operators $O^{(m)}$ as defined in Eq.(6c) and Eq.(6d). However, the possible closure $\delta E^{(r)} = 0$ can be considered whenever $||E^{(r)} + \delta E^{(r)}|| \approx ||E^{(r)}||$. This is made possible by choosing a reconstruction procedure in PoPe that is more accurate than that implemented in the code. For instance using an order 4 finite difference derivative in PoPe when an order 2 is used in the code.

$$||\delta O_t^{(r)}|| \ll ||O_t^{(m)} - O_t^{(n)}|| \tag{7a}$$

$$||\delta O_k^{(r)}|| \ll ||O_k^{(m)} - O_k^{(n)}|| \tag{7b}$$

and therefore $||\delta E^{(r)}|| \ll E^{(n)}$. We shall assume this relation to be fulfilled in the following, and, in the examples of this paper, we will give numerically based evidence that

the reconstruction scheme is consistent with this approximation. In the specific case where some parts of the discretisation scheme have been devised with highest accuracy, so that the reconstruction scheme can only achieve the same precision, one is led to assume $O^{(m)} \approx O^{(r)}$ for those parts of the discretisation scheme compared to the remaining ones, which are therefore assumed to generate all the error.

Since $E^{(r)}$ is determined by the output data, it is known for a series of points in phase space, at times t_i and at phase space locations X_i. The label i labels one point in the extended phase space combining time t and location X. It is to be underlined that the only constraint on the data, and therefore on the number of data-points i and their organisation in time and phase space, is to make possible a reconstruction procedure for the operators with better precision than the chosen discretisation procedure used for the simulations. In this framework Eq.(6a) is only defined for these selected data-points i.

$$O_{t,i}^{(m)} - \sum_{k=1}^{K} O_{k,i}^{(m)} = E_i^{(r)} \tag{8}$$

In the following the superscripts (m) and (r) are dropped to simplify the notation. The first step of the PoPe procedure is to build the error E_i for an ensemble of data points that are representative of the simulation that has been performed. This data verifies Eq.(8). In a second stage, the error E is projected on the operators driving the evolution of the system so that one can write:

$$E_i = \sum_{k=1}^{K} \delta c_k O_{k,i} + R_i \tag{9a}$$

The coefficients δc_k stemming from the projection do not depend on the data-point and only depend on the operator O_k. Such a projection procedure requires that the chosen operators are independent and not vanishingly small. Should either of these two cases occur, one must redefine the chosen operators accordingly. However, this drawback is compensated by the important insight one has gained on the physics of the system and on ways to address the numerical problem and its verification. Part of the error E_i is orthogonal to the set of operators O_k, which defines the residue R_i. Given Eq.(9a), one can rewrite Eq.(8) as

$$O_{t,i} - \sum_{k=1}^{K} \big(c_k + \delta c_k\big) O_{k,i} = R_i \tag{9b}$$

When choosing c_k to be the control parameter associated to the operator O_k, then δc_k is the absolute error made on that control parameter for the selected operator and chosen simulation. When one considers $c_k = 1$, as done in this paper, then δc_k is the relative error made on that control parameter. Irrespective of this choice one can write Eq.(9b) as:

$$E_i - \sum_{k=1}^{K} \delta c_k O_{k,i} = R_i \tag{10}$$

This linear equation depends on the K unknowns δc_k so that K data points are a priori sufficient to determine them when setting $R_i = 0$, which then defines the orthogonality.

One can then readily expect that for each set of K data points a different realisation of the K coefficient δc_k is computed. Three ways to address the possible statistics can be chosen. First, one can define the ensemble of coefficients δc_k for each available K-tuple of data-points and perform statistics on these realisations. Second, one can introduce the statistics directly in the calculation of the coefficients, for instance by computing the coefficients with a least square method using m-tuples of data-points with $m \geq K$. If N_{max} is the number of available data points, choosing $m = N_{max}$ then yields a unique value for each coefficient δc_k, $k \in [1, K]$. Third, when setting $K \leq m < N_{max}$, a mean square method can be used to define the projection and statistics can be performed on the results. It is to be noted that choosing a least square method with $m = K$, leads to a calculation that is quite similar to that proposed in the first item of this list. One can then recast the three possibilities that have just been described in terms of a specific choices of the m-tuples of data-points used in least square calculations.

- Coefficients δc_k are computed using the least square calculation for each available K-tuple of data-points, and statistics are performed given these realisations, case with $m = K$.
- Coefficients δc_k are computed using the least square calculation with m-tuple of data-points, with $m > K$. When $K < m < N_{max}$, statistics on the coefficients δc_k can be performed.
- The coefficients δc_k are computed using the least square calculation using all available data, hence with the N_{max}-tuple of data-points, $m = N_{max}$. A single value is generated for each coefficient.

2.2 PoPe projection defined with the least square method

Let us define d_i as:

$$d_i = E_i - \sum_{k=1}^{K} \delta c_k O_{k,i} \tag{11a}$$

and the distance d_m:

$$d_m^2 = \sum_{i=1}^{m} \tfrac{1}{2} d_i^2 \tag{11b}$$

The least square method then generates K-coupled linear equations defined by $\partial d^2 / \partial \delta c_k = 0$, namely by setting that d_m^2 is an extremum with respect to the variations of each δc_k. These coefficients are an optimum result for the particular choice of the m-tuple. The extremum equation obtained with respect to δc_k is then:

$$\sum_{k'=1}^{K} \delta c_{k'} \sum_{i=1}^{m} O_{k,i} O_{k',i} = \sum_{i=1}^{m} E_i O_{k,i} \tag{12}$$

When defining the scalar product $\langle F|G \rangle$ of the m-dimension vectors F and G by:

$$\langle F|G \rangle = \sum_{i=1}^{m} F_i G_i \tag{13}$$

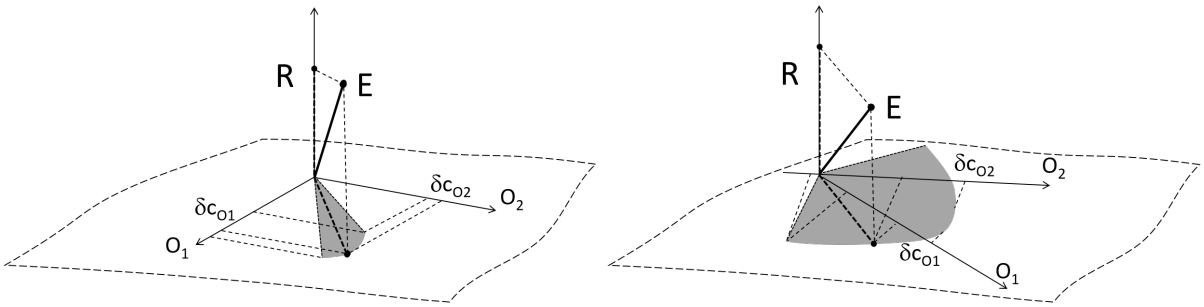

Figure 1: Projection of the error E, on the plane of the operators O_1 and O_2, yielding the coefficients δc_{O1} and δc_{O2} and defining the residue R in the direction orthogonal to this plane. The projection in the plane (O_1, O_2) of the fluctuations of the error are indicated by the grey region. Left hand side: sketch of the projection when O_1 and O_2 are near orthogonal, the variation of the coefficients δc_{O1} and δc_{O2} are reduced. Right hand side: sketch of the projection when O_1 and O_2 are nearly co-linear driving a larger uncertainty on δc_{O1} and δc_{O2}.

the extremum constraint takes the form of a projection:

$$\sum_{k'=1}^{K} \delta c_{k'} \langle O_k | O_{k'} \rangle = \langle O_k | E \rangle \tag{14a}$$

this result being completed by the orthogonality of the residue:

$$\langle O_k | R \rangle = \left\langle O_k \middle| \left(E - \sum_{k'=1}^{K} \delta c_{k'} O_{k'} \right) \right\rangle = 0 \tag{14b}$$

The least square method therefore defines a particular projection for a code output data. Other projections can be defined. For instance, one can specify a weight for each m-tuple enforcing in the result a class of m-tuples. For example, in the case of $K = 2$ one can define the weight as $\langle O_1|O_1\rangle \langle O_2|O_2\rangle - \langle O_1|O_2\rangle^2$ than ought to reduce the impact of co-linearity.

Regarding the dimension of the phase space used to evaluate the error, the PoPe method is reminiscent of the idea of generating a phase diagram using a time series of a single variable [16], here $O_{k,j}$ where k labels a particular vector and j the index in the time series for a chosen time delay. The vector $(O_{k,j}), j \in [j_1, j_1 + m]$ is then assumed to define a position at time j_1 in a phase space of dimension m. However, while a single time series is used for that reconstruction, for PoPe we consider the case with several signals generating different time series labelled here by $k \leq K$. Each operator is then identified to a function acting in a space of dimension m and requires a priori an infinite dimension bases of function to represent it. We also introduce another difference in building the vectors $(O_{k,i}), 1 \leq i \leq m$ by choosing of the same set of indices i for all operators but each set being chosen randomly. However, given the constraint Eq.(1), we assume that the trajectories are mostly embedded in the function-space of dimension K generated by the K operators. A presumed small contribution exists and is transverse to that plane, the residue R. Once a scalar product is defined, for instance that generated by the least square method, projections can be computed and one can follow a particular simulation

in the phase space generated according to this procedure, see Figure 1. To simplify the situation, we sketch the problem in $2D$, thus for two signals O_1 and O_2 generating the time series. Two cases are then observed. When the two operators are independent, 1 left hand side, the error E is projected in the plane (O_1, O_2), the coefficients δc_{O1} and δc_{O2} are well defined and their dispersion accounts for the numerical errors. However, when the two operators are nearly co-linear, Figure 1 right hand side, large variations of δc_{O1} and δc_{O2} can occur. Increasing the dimension m of the phase space tends to reduce the co-linearity, unless the operators O_1 and O_2 are ill chosen and actually co-linear (which would be however a useful information regarding the system). Note that in Figure 1, the numerical fluctuations are only indicated by their projection in the plane (O_1, O_2), the shaded grey regions, and that, for convenience of the representation, the operators are not shown to fluctuate. In practise, these fluctuations can govern transitions between left and right hand side relevant geometry. Minimising the impact of the latter situations of co-linearity by increasing the dimension m of the embedding space is performed at the cost of reducing the description of the statistics of the fluctuations, eventually narrowing the grey window to a single value.

2.3 iPoPe analysis

In order to solve Eq.(14a), one has to inverse a $K \times K$ matrix to obtain the coefficients δc_k, $1 \leq k \leq K$. In this process, all coefficients appear on the same footing. However, when the operators of the system do not have the same magnitude, a small error on the calculation of a large amplitude operator can have a large impact on an operator with comparatively smaller amplitude. There is a possibility of propagating the error from a particular operator on the coefficients of other operators. Furthermore, inverting a large matrix as required for the standard PoPe method can be cumbersome. However, when the matrix is diagonal elements each coefficient is computed independently. We generalise this property to define the iPoPe method, for independent Projection on Proper elements. This method addresses the projection operator after operator in a staged approach and is identical to the PoPe solution when the matrix is diagonal. Let us choose k as the first element of the projection, then one determines the iPoPe coefficient as:

$$\delta c_k^{(k)} \langle O_k | O_k \rangle = \langle O_k | E \rangle \tag{15a}$$

this result being completed by the calculation of the specific residue R_k orthogonal to O_k:

$$R_k = E - \delta c_k^{(k)} O_k \tag{15b}$$

Because of the various possible choices of k out of K, one is led to using more complicated notations. As for PoPe, the subscript of k of the coefficient $\delta c_k^{(k)}$ refers to the operator k and the error on its control parameter, while the superscript (k) refers to the order chosen to determine the coefficients.
Computing $\delta c_k^{(k)}$ is then absolutely straightforward:

$$\delta c_k^{(k)} = \langle O_k | E \rangle \, / \, \langle O_k | O_k \rangle \tag{15c}$$

The coefficient $\delta c_k^{(k)}$ that is obtained maximises the importance of the operator O_k in generating the error since one computes $\delta c_k^{(k)}$ as if all the error was stemming from that

operator. In a second stage, one can compute the coefficient $\delta c_{k'}^{(k,k')}$ and a new residue as follows:

$$\delta c_{k'}^{(k,k')} \langle O_{k'}|O_{k'}\rangle = \langle O_{k'}|R_k\rangle \tag{16a}$$

this result being completed by the calculation of the specific residue $R_{k,k'}$ orthogonal to $O_{k'}$:

$$R_{k,k'} = R_k - \delta c_{k'}^{(k,k')} O_{k'} \tag{16b}$$

Step by step one can iterate the procedure until all coefficients are determined, and the ultimate residue is computed. The values of a particular coefficient $\delta c_{\ell}^{(k,k',\ldots,\ell,\ldots)}$, the error made on the control parameter of operator O_ℓ, now depends on the order chosen for the calculation identified by the superscript $(k, k', \ldots, \ell, \ldots)$. The simplicity of iPoPe is balanced by the number $K!$ of different ways it can be applied. The total number of different possible values of any given coefficient N_{iPoPe} is not quite as big since computing a coefficient at a given stage only depend on the various combinations prior to that selected, on the left of ℓ in the sequence $(k, k', \ldots, \ell, \ldots)$ and not on those to the right.

$$N_{iPoPe} = \sum_{k=1}^{K} \frac{(K-1)!}{(K-k)!} \tag{17}$$

A systematic use of iPoPe considering all these combinations is prohibitive whenever K is large. The method would then only useful if a bias is introduced that defines an order in which the coefficients are determined. One can also consider a mix of iPoPe and PoPe in the procedure, for example giving a particular weight to a class of operators with iPoPe and treating the remnant on equal footing with PoPe.
The most useful iPoPe procedure is restricting iPoPe to only computing the first step for each operator $1 \leq k \leq K$, with Eq.(15). One then maximises the possible error measured by $\delta c_k^{(k)}$ for each operator. The benefit of this reduced iPoPe procedure is its simplicity, therefore reducing the CPU cost, together with the fact that it yields the largest possible error for each coefficient. This maximum error procedure can therefore be used to determine a figure of merit.

As a by-product of the PoPe or iPoPe verification method, one can investigate $\langle O_k|O_k\rangle$ the actual weight of the operator in the balance as well as its change in time or space. Comparing the magnitude of the operators $\langle O_k|O_k\rangle$ for different values of k, one then gains insight into the effective weight of each operator, and how it stands with respect to non-linear analysis principles, such as that of critical balance [1]. The error δc_k also indicates with what precision each particular operator is determined numerically. This understanding can then be used as a guideline to improve the code accuracy, knowing which operator has the largest weight and which exhibits the largest error.
The analysis of the error that is performed by PoPe and iPoPe indicates that the output data would not be discernible for the reference simulation with operator O_k, with weight 1, and a simulation with weight in the range 1, $1 + \delta c_k$ for the same operator O_k. The numerical solution therefore introduces an uncertainty on the effective value of the control parameters. In most cases, the control parameters that are used as input are known with error bars that are larger than that obtained by PoPe. However, in some cases

the error δc_k can be large enough to significantly modify the control parameter. It is then most important to know that the simulation output corresponds to a value of the control parameter that is different from what is assumed. Depending on the situation, one can perform a new run of the code changing the value of the control parameter to compensate the error. Alternatively, one must improve the numerical precision to address the particular problem with appropriate precision.
An important feature of PoPe and iPoPe is that the definition of the operators to be addressed in the verification procedure must be guided by the physics to be addressed by the simulation. As an example for a system that exhibits an evolution in time, the target operator $O_t^{(m)}$ in Eq.(1) is implicitly assumed to be the time derivative. This is appropriate in a turbulent or chaotic regime. Should the solution evolve towards a fixed point, then such a choice is only relevant in the early phase of the transient. To address a later stage of this transient, a different choice of $O_t^{(m)}$ will be more appropriate since the time derivative operator in the simulation is converging towards the null operator. It is also possible to define differently the operators, for example using $O_{k,k'} = O_k + O_{k'}$. The choice is guided by the physics insight into the simulation or a particular interest in analysing the numerical scheme.
These various possibilities underline the versatility of PoPe for verification purposes based on simulation output. Furthermore, one finds that the PoPe or iPoPe methods provide an in depth analysis of the chosen simulation, both as a tool to investigate the physics and that to identify possible shortfalls of the chosen numerical scheme.

3 Standard verification of strange attractor simulations

3.1 The strange attractor model

The model we consider to present the PoPe verification method is the simple model of a particle subject to two electrostatic waves with different pulsation and identical wave vector and amplitude. Alternatively, it can be understood as the model for a compass in a two component magnetic field, one fixed and the other rotating, both components having the same amplitude. The phase space motion is thus two dimensional ($2D$) with one dimension standing for the position x, either the position of the particle or the angle of the compass, and one standing for the momentum J, either the momentum of the particle or the angular momentum of the compass. The normalised evolution equations for dx/dt and dJ/dt are :

$$\frac{dx}{dt} = J \tag{18a}$$

$$\frac{dJ}{dt} = -2\pi\ B\Big(\sin\big(2\pi x\big) + \sin\big(2\pi(x-t)\big)\Big) - \nu\ J \tag{18b}$$

The parameter B -the normalised electric potential of the electrostatic waves or the amplitude of two components of the magnetic fields- is directly connected to the Chirikov overlap parameter [11] σ_{chir}. The characteristic island width δ_i is $\delta_i = 2\sqrt{B}$ and the chosen distance between the resonances is $\Delta = 1$ so that $\sigma_{chir} = 2\delta_i/\Delta = 4\sqrt{B}$. A fluid viscosity damping term $-\nu J$ governs the contraction of the phase space volume to zero.

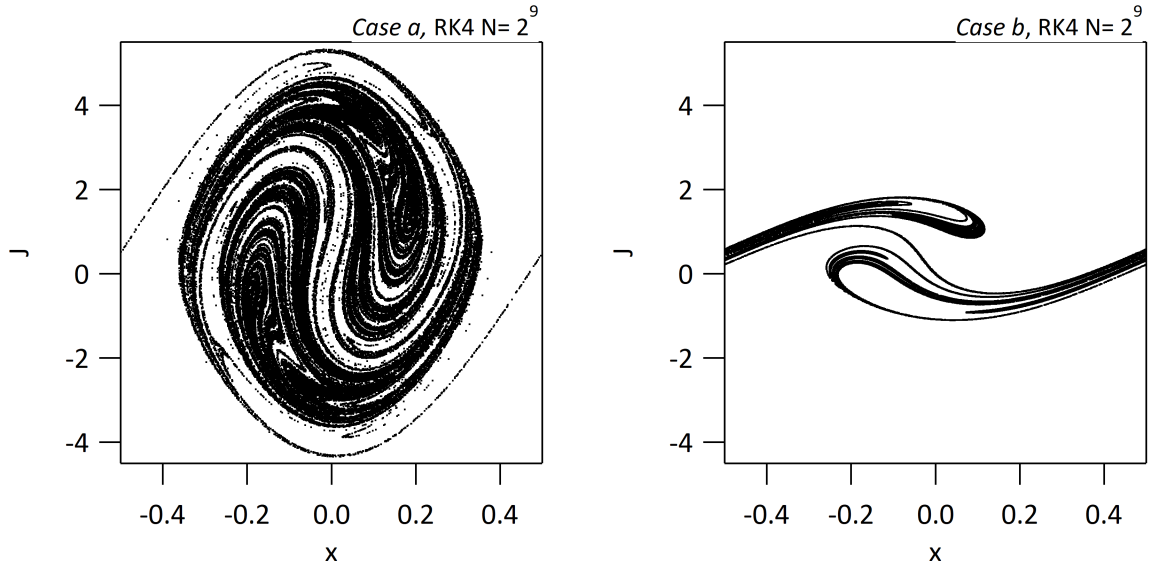

Figure 2: Poincaré section of the strange attractor generated by Eqs.(18) , *case a* with $\sigma_{chir} = 7$, hence $B \approx 3.0625$, and $\nu = 0.2$, left hand side, and *case b* with $\sigma_{chir} = 2.3$, hence $B \approx 0.330625$, and $\nu = 0.8$ right hand side. Simulation with order 4 Runge Kutta and 2^9 time steps per unit time.

For convenience, we introduce the Hamiltonian H_0 of the non-dissipative evolution so that:

$$H_0 = \tfrac{1}{2}J^2 - B\Big(\cos\big(2\pi x\big) + \cos\big(2\pi(x-t)\big)\Big) \tag{19a}$$

$$\frac{dx}{dt} = \frac{\partial H_0}{\partial J} \qquad ; \qquad \frac{dJ}{dt} = -\frac{\partial H_0}{\partial x} - \nu\ J \tag{19b}$$

The trajectory of the system is presented in a standard fashion, in the so-called Poincaré sections, a stroboscope effect at time interval 1, which is the period of the driving force, figure (2). Two cases will be considered in this work:

- *case a* with control parameters $\sigma_{chir} = 7$, hence $B \approx 3.0625$, and $\nu = 0.2$, Figure 2 left hand side
- *case b* with control parameters $\sigma_{chir} = 2.3$, hence $B \approx 0.330625$, and $\nu = 0.8$, Figure 2 right hand side

The simulation of the strange attractor is chosen because it combines simplicity of the numerical integration and sensitivity to initial conditions. The latter makes verification slightly more challenging since any error, including numerical errors, governs an exponential separation between trajectories. The chosen numerical time stepping schemes are order 2 and order 4 Runge Kutta (RK2 and RK4 respectively). The sensitivity to initial conditions is governed by the Lyapunov exponent defined as the average along the trajectory defining the strange attractor of the largest eigenvector of the tangential map [2]. The latter is readily determined:

$$\frac{d\delta x}{dt} = \Big[\partial_J^2 H_0(x_t, J_t)\Big]\delta J \tag{20a}$$

$$\frac{d\delta J}{dt} = -\Big[\partial_x^2 H_0(x_t, J_t)\Big]\delta x - \nu\ \delta J \tag{20b}$$

where x_t, J_t is a phase space position belonging to the trajectory. The eigenvalues associated to the tangential map are therefore:

$$\lambda_t^{(\pm)} = -\frac{\nu}{2} \pm \Delta_t^{1/2} \tag{21a}$$

$$\Delta_t = \left(\frac{\nu}{2}\right)^2 - \partial_J^2 H_0(x_t, J_t)\partial_x^2 H_0(x_t, J_t) \tag{21b}$$

One can readily check that the phase space contraction of the strange attractor is governed by the viscosity ν since its volume shrinks exponentially in time according to $\exp(\left\langle \lambda_t^{(+)} + \lambda_t^{(-)} \right\rangle t) = \exp(-\nu t)$. The global property of the strange attractor is captured by the largest Lyapunov exponent Λ assuming $\Lambda > 0$. The latter measures the sensitivity to initial conditions and is determined numerically [2]. The eigenvalues are sometimes referred to as the local Lyapunov exponents, which underlines the connection between the actual Lyapunov exponents and the series of eigenvalues on a chaotic trajectory.

3.2 Method of Return Solution for the strange attractor

In this Section, we perform a verification of the simulations with the Method of return Solution (MRS) as reference for the PoPe and iPoPe methods. One compares the results obtained with two different integration schemes, of order 2 and 4 respectively, and when varying N the number of integration step per unit time, from 2^3 to 2^{12}. Properties of the chaotic attractors for each integration scheme and resolution are also determined and compared.

As discussed in Appendix A, the limitation of the Method of Manufactured Solution as implemented lies in the assumption that the generated fixed point is stable. The verification stage then allows one to check that the numerical response exhibits a fixed point and to determine with what precision the fixed point is recovered. Rather, than enforcing an arbitrary fixed point, the alternative Method of Return Solution (MRS) is based on a return to the initial condition: hence after N steps forward in time, the subsequent N steps are performed with the opposite time step [13]. Mathematically the system must therefore reverse to its initial position, which is therefore the fixed point. However, the numerical errors, partly amplified by the effect of the divergence of neighbouring trajectories, will distort the trajectory and a distance d_r is generated between the initial and final positions in phase space, see Figure 3. This distance is averaged over the points belonging to the strange attractor to yield a measure of the accuracy. The idea is therefore similar to the standard Method of Manufactured Solution except that the chosen reference solution is the initial condition via the backward steps in time. By analogy, we call this verification scheme the Method of Return Solution, or MRS. However, if the return steps are performed with the same algorithm, one cannot verify that the numerically solved equations are properly implemented[1] The MRS appears therefore better suited to evaluate the numerical scheme accuracy. Another limitation is for symplectic integrators that enforce time reversal of the numerical solution. The numerical error of such schemes is

[1]This issue is similar to that of the MMS where the source term ensuring that a given analytical expression is solution of the equations must not be computed with the code that is being verified. For the MRS, a possibility to achieve this verification stage would be to step backwards with a different and verified code.

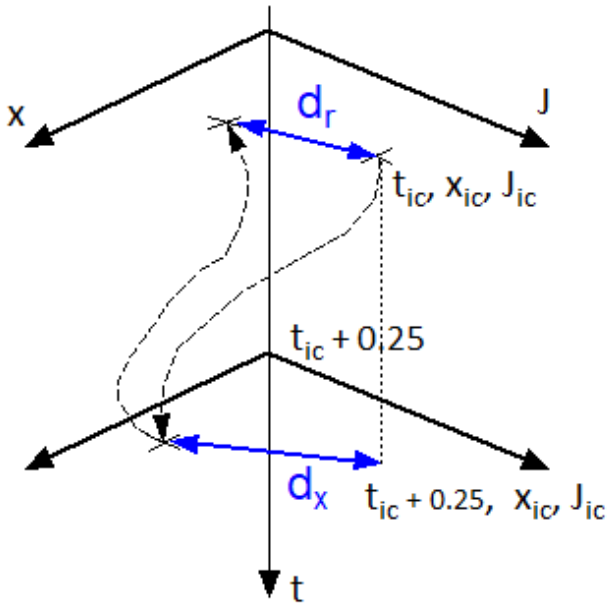

Figure 3: Verification with the Method of Return Solution, sketch of the method.
– initial condition (ic) belonging to the trajectory of the system t_{ic}, x_{ic}, J_{ic},
– trajectory stepped forward for $\Delta t = \frac{1}{4}$, reaches distance d_x from the initial condition,
– then, trajectory stepped backward for Δt, distance from initial condition d_r.

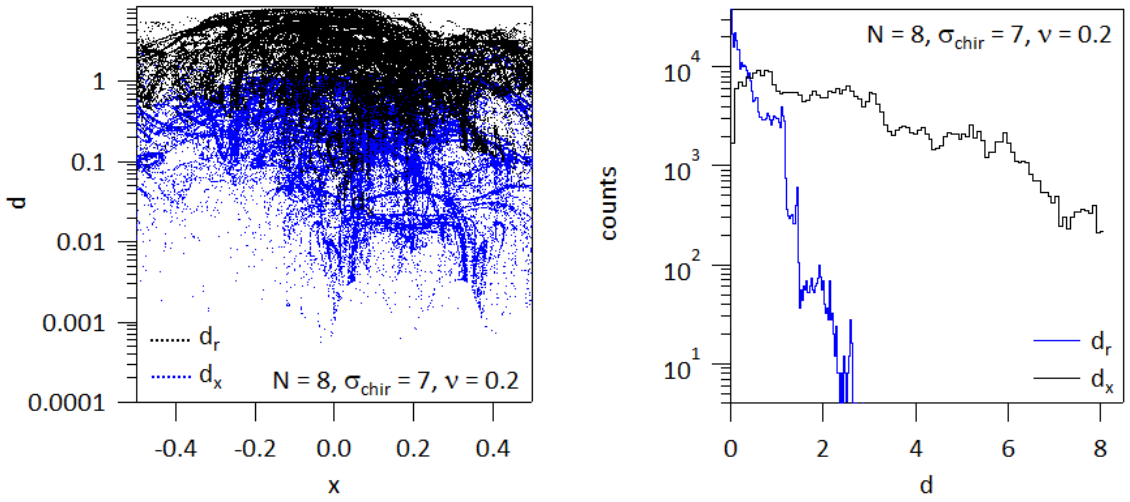

Figure 4: Verification with the Method of Return Solution for *case a*, $\sigma_{chir} = 7$ and $\nu = 0.2$. Left hand side: distance between phase space initial conditions and positions after $\Delta t = 0.25$, hence after $\Delta t/\delta t$ times steps, d_x black dots, and distance between initial condition and return point d_r blue dots versus the position of the initial condition x. Right hand side: same data, histograms of the distances, d_x black histogram, d_r blue histogram.

.

then "hidden" and cannot be estimated with the MRS. However, stepping back with the same symplectic scheme but smaller time step should provide a measure of the accuracy. This alternative for symplectic schemes has not yet been investigated. For the Runge Kutta time integrators we address in this paper, the chaotic nature of the trajectory plays a role in the distance d_r that is observed since any error is exponentially amplified. However, d_r will be an increasing function of the effective numerical error and it provides consequently a useful measure in the verification procedure, in particular to determine the order of the numerical scheme.

For the strange attractor both the second and fourth order Runge Kutta schemes are used varying the number of time steps per period from $N = 2^3 = 8$ to $N = 2^{12} = 4096$. As shown in Appendix B, one expects the error determined by MRS to scale like the order of the time stepping scheme plus one[2], hence a decrease of the error like N^{-3} for the order 2 Runge Kutta scheme, labelled RK2, respectively N^{-5} for the fourth order Runge Kutta scheme, labelled RK4.

We first consider *case a*, with large Chirikov parameter, $\sigma_{chir} = 7$ and $\nu = 0.2$, see Figure 2 left hand side, and comparing the RK4 and RK2 schemes. For a series of points belonging to the attractor, the evolution is stepped forward during a fourth of a period, $\Delta t = 0.25$, the distance from the initial condition d_x is then recorded, then the time stepping is reversed, and the trajectory therefore heads back towards the initial condition. The distance d_r between the initial and return points in phase space is then computed. For a large time step with $N = 2^3$ steps per period, $\delta t = 0.125$, one can compare the

[2]This result holds when the distance d_r is small enough to permit the expansion performed in Appendix B, otherwise the scaling is determined by the time stepping as reported in Ref. [13].

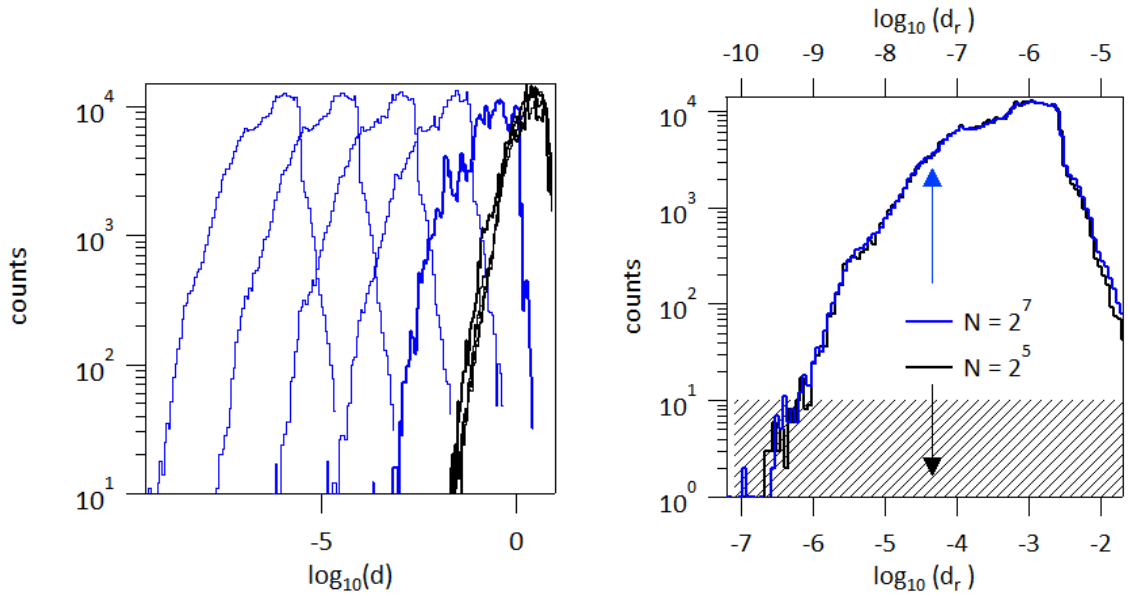

Figure 5: Verification with the Method of Return Solution for *case a*, $\sigma_{chir} = 7$ and $\nu = 0.2$. Left hand side: histograms of $\log_{10}(d)$ for various resolutions N, $N = 2^3, 2^4, 2^5, 2^6, 2^7$, black histograms of d_x black curves, histograms of d_r blue curves. Right hand side: similarity of the MRS error histograms of $\log_{10}(d_r)$ for $N = 2^5$ lower scale and $N = 2^7$ upper scale.

distribution of distances d_x and d_r, Figure 4 left hand side. These distances are plotted versus the position x of the initial condition, black dots for d_x and blue dots for d_r. At this low resolution one finds that d_r is large typically $\approx 0.1 d_x$ with $d_x \approx 2.$. The maximum distance reached after a fourth of a period is comparable to the "size" of the strange attractor, typically up to 5, Figure 2 left hand side. One can analyse the distribution of these distances, Figure 4, right hand side. The histograms of d_r, blue curve, and d_x, black curve indicates that the distribution of d_x is quite broad. Conversely, the measurement of the error d_r is characterised by a narrower histogram peaked on the smallest distance $d_r = 0$. The histograms of $\log_{10} d_x$ and $\log_{10} d_r$ yield more insight into the error. These histograms for different resolutions are compared on Figure 5 left hand side. The resolution is characterised by the number of steps N per unit time, the period of the potential, hence defining the time step $\delta t = 1/N$. On Figure 5 left hand side are compared the simulations for $N = 2^3$, $N = 2^4$, $N = 2^5$ and $N = 2^6$, the black curves that overlay correspond to the histogram of d_x while the various histograms in blue are those of d_r. The latter shift towards smaller distances as N is increased, while the former are typically unchanged. The histograms drawn with thick lines correspond to the resolution $N = 2^3$. One can remark that the shift towards the smaller values of the histograms of $\log_{10}(d_r)$ appears to be at a constant value for each increase of N by a factor 2. One thus finds that the distance d_x does not exhibit qualitative changes as the resolution is improved, while the measure of the MRS error exhibits a decrease with the resolution. The similarity between these various histograms of d_r is more clearly shown on Figure 5 right hand side, where the resolution $N = 2^5$, lower scale, is compared to $N = 2^7$ resolution upper scale. Note that the scales are identical but for a shift of $\log_{10}(10^{-3})$ from the lower to the upper scale. The shaded region corresponds to the number of counts smaller than 10. One can remark that the distribution of the distance d_r appears to be nearly

unchanged when N is varied. This distribution is broad and skewed: for $N = 2^7$, one finds $\langle d_r \rangle \approx -7.875$ and a standard deviation $\delta d_r \approx 0.66$ with skewness ≈ -0.66. For each value of the resolution N, these statistics are performed with 320064 different initial conditions chosen on the strange attractors computed with the different resolutions. The similarity of the distribution of the error for these different resolutions underlines the fact that the error governed by the time integration scheme is of the form $f(x_{ic}, J_{ic}, t_{ic})\delta_t^5$. Provided the change of phase space position x_{ic}, J_{ic} at time t_{ic} is statistically identical for each resolution, then the realisation of the function f will be identical, hence with the same shape of its distribution function, while the dependence on δ_t^5 will govern a shift of the form $-N\log_{10}(2)$.

One can then analyse the dependence of the error on the resolution N that determines the time step $\delta t = 2^{-N}$, Figure 6. For the reference *case a*, $\sigma_{chir} = 7$, $\nu = 0.2$, one checks that the error $\langle d_r \rangle$ scales like N^{-5} for the order four Runge Kutta scheme, blue open circles, and N^{-3} for the order two Runge Kutta scheme, black upside down triangles. The scaling appears to hold over the whole range of values of N, but for a small departure at $N \approx 2^3$. For completeness, the results for *case b*, $\sigma_{chir} = 2.3$, $\nu = 0.8$, are also plotted. These simulations are performed with the fourth order Runge Kutta time stepping. One recovers the appropriate slope associated to the order of the scheme, and, as can be expected, one can observe that the error levels-off when the error becomes comparable to machine precision. However, one finds that the error exhibits a quite different magnitude when comparing *case a* and *b*. This agrees with the fact that the sensitivity to initial conditions is characterised by a different Lyapunov exponent, which is larger in *case a* than in *case b*. This governs a larger exponential growth of the error in *case a* compared to *case b*. In the present examples, the difference in the MRS error is close to three orders of magnitude. The test for one regime of parameters does not allow one to predict the precision for another. Consequently, the accuracy test, combining verification and analysis of the effective precision, should be made for each particular regime addressed in the simulation effort.

When considering the phase portrait of the attractors, the eye inspection indicates that the accuracy issue is more demanding than simply assessing the precision of the numerical scheme. This is particularly noticeable with low resolution simulations, Figure 7. With $N = 2^3$ steps per unit time, the achieved phase portrait with both RK4, Figure 7 left hand side, and RK2, Figure 7 right hand side, depart significantly from that displayed on Figure 2 left hand sides obtained with the same control parameters but with $N = 2^9$ and RK4. Based on this eye inspection, these low-resolution results appear to be inaccurate and consequently misleading. Therefore, knowing the error and checking its scaling when changing the resolution is a verification of the numerical scheme but does not provide a clear measure of the accuracy. The way to proceed to a correct accuracy assessment appears to be unanswered but for the naive statement "the smaller the error, the better". This quest for minimum error is naive because: (i) it relies implicitly on infinite resources, (ii) it does not discuss the actual need in terms of precision, (iii) it cannot guaranty exactness for chaotic systems since the sensitivity to initial conditions implies that any error, however small, will be amplified to macro-scales. The alternative to the naive statement is to focus on numerical measurements that are relevant in terms of physics. Regarding the strange attractor, the largest Lyapunov exponent can be regarded as such a mea-

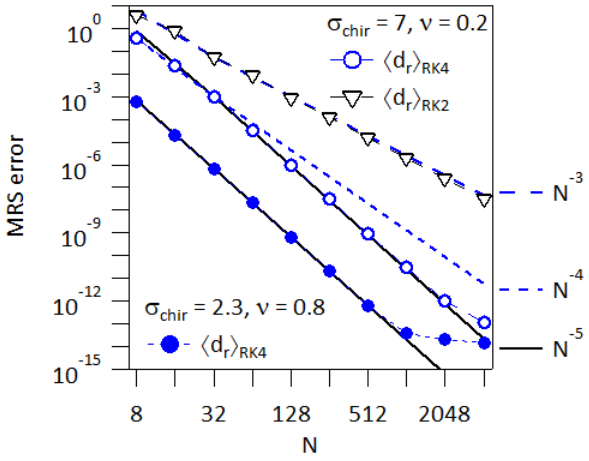

Figure 6: Investigation of the order of the numerical scheme with the Method of Return Solution (MRS). For *case a*, $\sigma_{chir} = 7$ and $\nu = 0.2$, comparison of the Runge Kutta schemes of order two (RK2), up-down open triangles, and four (RK4) open circles. For *case b*, $\sigma_{chir} = 2.3$ and $\nu = 0.8$, precision with the Runge Kutta schemes of order four, closed circles. The expected scaling exponents, N^{-5} for RK4 and N^{-3} for RK2, are recovered.

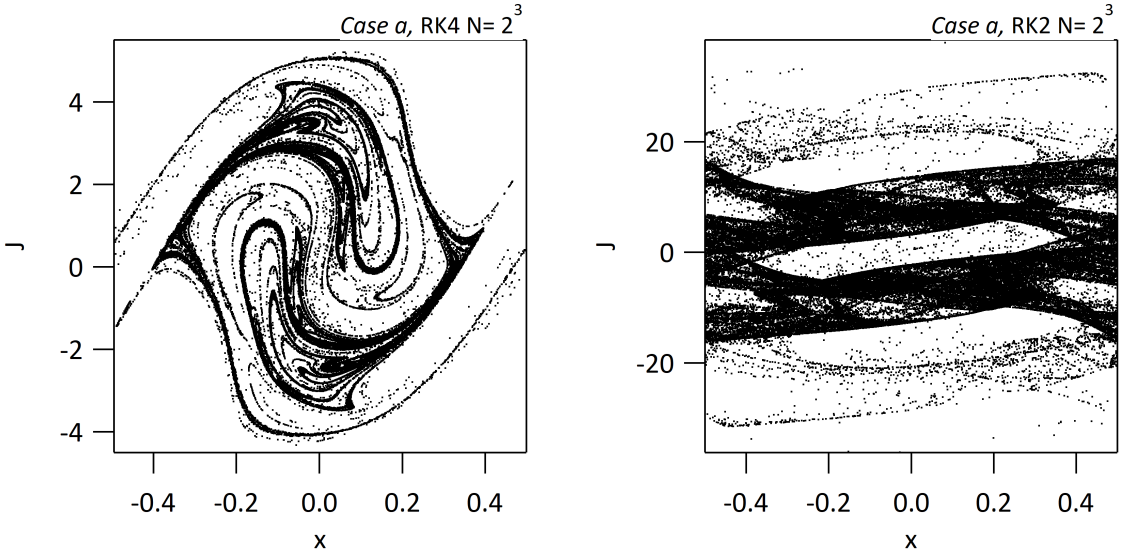

Figure 7: Strange attractor for *case a*, $\sigma_{chir} = 7$, $\nu = 0.2$ and time stepping with $N = 2^3$ steps per unit time. Left hand side: fourth order Runge Kutta integration scheme. Right hand side: second order Runge Kutta integration.

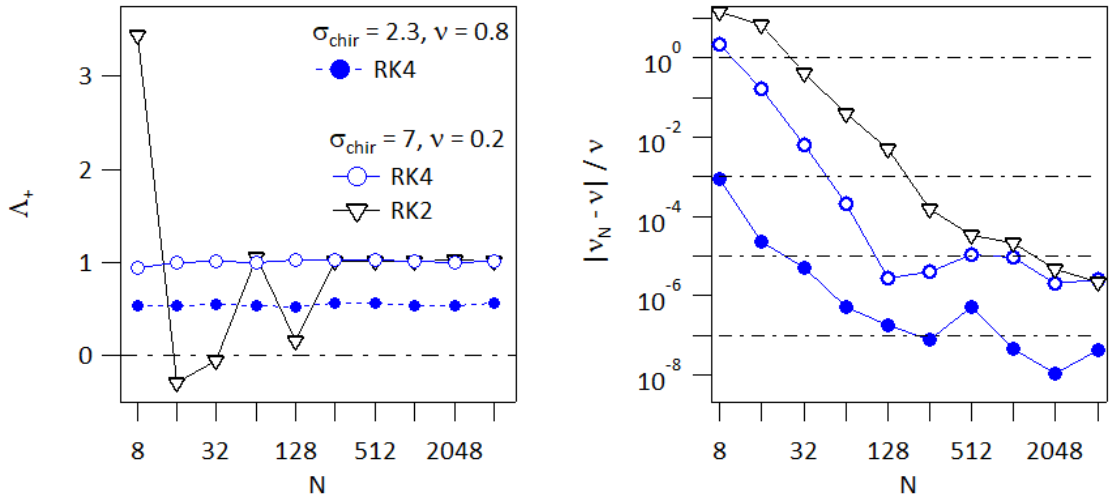

Figure 8: Accuracy investigation for the cases a, $\sigma_{chir} = 7$, $\nu = 0.2$, open symbols, and b, $\sigma_{chir} = 2.3$, $\nu = 0.8$ closed symbol with the RK4 and RK2 integration schemes, circles, respectively head down triangles. Left hand side, calculation of the largest Lyapunov exponent Λ_+. Right hand side, relative error on the value of ν determined by the calculation of the rate of decrease of the phase space volume.

surement, Figure 8, left hand side. One can observe that the results obtained with the RK4 scheme are characterised by a nearly constant value of the Lyapunov exponent with not distinct trend when increasing the precision. For the simulation conditions $\sigma_{chir} = 7$, hence $B \approx 3.0625$, and $\nu = 0.2$, the only significant change with RK4 is that between the simulation with $N = 2^3$, with $\Lambda_+ \approx 0.95$ and the other simulations with larger values of N where $\Lambda_+ \approx 1.0$ is observed. However, when considering the results obtained with the RK2 integration scheme, one finds a large variation until $N \geq 2^8$. The phase portraits confirm the negative values of the Lyapunov exponent. They exhibit fixed points with transient trajectories spiralling in towards them. Based on the largest Lyapunov exponent one can argue that one must consider $N \geq 2^8$ for the RK2 integration scheme, while $N \geq 2^4$ would suffice with the RK4 integration scheme. However, the examination of the Lyapunov exponent for the latter does not provide a clear measure to discriminate the accuracy. Still, considering these two critical number of steps, and since the cost of the RK4 scheme compared to RK2 is typically a factor 2 in the number of operations to be done, one finds a net gain of a factor 8 in computing resources by implementing the RK4 scheme rather than RK2 for this problem.

An alternative measure to evaluate the results is to determine the exponent that characterises the shrinking of the phase space volume, therefore for a given precision N: $\Lambda_+ + \Lambda_- = -\nu_N$. The benefit is that one expects the exponents ν_N to converge towards ν when N is increased. The relative error $|\nu - \nu_N|/\nu$ thus appears to be a more precise measure to evaluate the exactness of the numerical scheme. However, determining numerically the exponent ν_N adds a cost in computing resources of about 50% and yields an output that is known a priori but for the error in computing it. Another caveat is that this error can also be specific of the calculation of ν_N and consequently not relevant to assess the correctness of the evaluation of the Lyapunov exponent. For the three cases that have

been analysed, one finds that the relative error $|\nu - \nu_N|/\nu$ decreases as expected when N is increased. This gain in accuracy appears to level-off at a value of the order of 10^{-5} for *case a* $\sigma_{chir} = 7$, $\nu = 0.2$ and 10^{-7} for the *case b* $\sigma_{chir} = 2.3$, $\nu = 0.8$. The accuracy of this measure is again dependent on the problem of interest. The fact that the relative error appears to level-off also provides a possible rule to determine the reference precision as the effective lower bound as well as the optimum value of N where the roll over occurs. For the RK4 scheme one finds typically $N \approx 2^7$, while for the RK2 scheme $N \gtrsim 2^9$ seems appropriate. This criterion to evaluate the exactness still indicates that using the RK4 scheme compared to RK2 yields a net gain of a factor 2 in computing resources. The analysis of the relative error on the calculation of ν indicates that the lowest resolution yields an error exceeding unity, which is clearly too big. Comparing the relative error to the calculation of the Lyapunov exponent, one can determine the empirical rule that the relative error on the calculation of ν should be smaller than 10^{-3}.

The full analysis with the Method of Return Solution provides a verification of the numerical scheme and yields case dependent rules to assess the exactness of the simulation. Such an analysis must be performed and results checked for each class of simulations of interest. However, accuracy can be investigated a priori for a particular case that determines key trends: verification of the order of the numerical scheme and trade-off between error and computational cost. Then for any specific simulation, the MRS method can be used at any restart condition of a particular simulation. The actual accuracy for the chosen simulation can then be checked. This indicates that verification at the stage of production runs is relevant. First, because research oriented codes most often evolve continuously and verification that the equations effectively implemented in the code are the equations of interest cannot be done once for all. Second, because the choice of the control parameters has an impact on the "numerical stress" of a chosen scheme. A particular simulation with the MRS, or a particular solution for the MMS, provide the trends for the error but not a universal accuracy check. PoPe and iPoPe are designed to circumvent this issue by yielding an accuracy check for any production run.

4 PoPe verification for the strange attractor

4.1 PoPe error analysis for the strange attractor

The PoPe verification is based on data mining using the output of production runs. From the saved data, it is possible to reconstruct the values of the different operators that drive the problem at hand. For the strange attractor, the series of values of x_i, J_i and t_i, where the index i identifies the number in a time series, hence $x_i = x(t_i)$ and $J_i = J(t_i)$ are used for the verification. Provided the time series are saved with the same time step has that used by the numerical scheme, one can proceed to verification. Rather than using Eq.(18), which is actually implemented in the code, we consider the equivalent second order equation:

$$\frac{d^2x}{dt^2} \;=\; -2\pi\; B\Big(\sin\big(2\pi x\big) + \sin\big(2\pi(x-t)\big)\Big) - \nu\;\frac{dx}{dt} \tag{22}$$

One can note that in the verification procedure chosen here an equivalent but different mathematical setting of the problem is addressed. Computing the various operators of

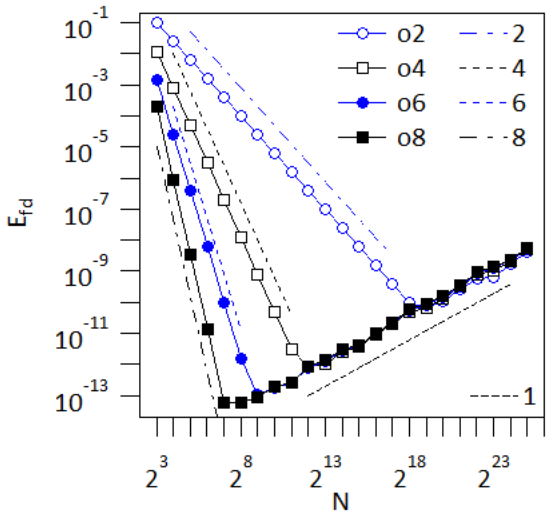

Figure 9: Error E_{fd} obtained by comparing the derivative of $\sin(t)$ to $\cos(t)$ obtained with finite difference, order 2 blue open circles and order 4 black open squares, order 6 blue full circles and order 8, full black squares. The error is plotted versus N the number of time steps per unit time. The theoretical decay rates are also indicated by the dash dot and dashed lines. The dashed black line with positive slope N^1 fits the loss of accuracy when N is too large.

Eq.(22) using the output data is straightforward for the right hand side. For the time derivative operators, one has to rebuild the time derivatives using alternative schemes. We have used here finite difference up to order eight. Similarly to the Runge Kutta integration verified in Appendix A, the finite difference derivatives are checked independently by comparing the numerical derivative of $\sin(t)$ to the analytic value $\cos(t)$, Figure (9). The measured errors E_{fd} are observed to compare well with the expected orders of the finite difference schemes until precision reaches the machine noise. The number of operations is then too large, no precision can be gained due to the numerical scheme, but the impact of the numerical noise, increasing with the number of steps, overwhelms the accuracy of the schemes. This governs an increase of the error with slope 1.

For each point i, position x_i at time t_i of the trajectories, one can then compute the error $E^{(r)}_{ok,i}$ as:

$$E^{(r)}_{ok,i} = \left[\frac{d^2x}{dt^2}\right]^{(r)}_{ok,i} - \ \mathrm{RHS}^{(r)}_{ok,i} \tag{23a}$$

$$\mathrm{RHS}^{(r)}_{ok,i} = -2\pi\ B\Big(\ \sin\left(2\pi x_i\right) + \sin\left(2\pi(x_i - t_i)\right)\Big) - \nu\ J_i \tag{23b}$$

where $\left[d^2x/dt^2\right]^{(r)}_{ok,i}$ is the reconstructed (superscript (r)) second derivative of x with respect to t, computed with finite difference schemes at order k, indicated by the subscript ok. To simplify the notations, the superscript (r) will be dropped in the following. The error $E^{(r)}_{ok,i}$ then depends on that of the reconstruction scheme, but for the issues of interest it mostly depends on the error made to generate the trajectory, typically governed by the time step of the Runge-Kutta integration scheme and the order of the latter scheme. To illustrate this procedure we consider *case b* with control parameters $\sigma_{chir} = 2.3$, $\nu = 0.8$,

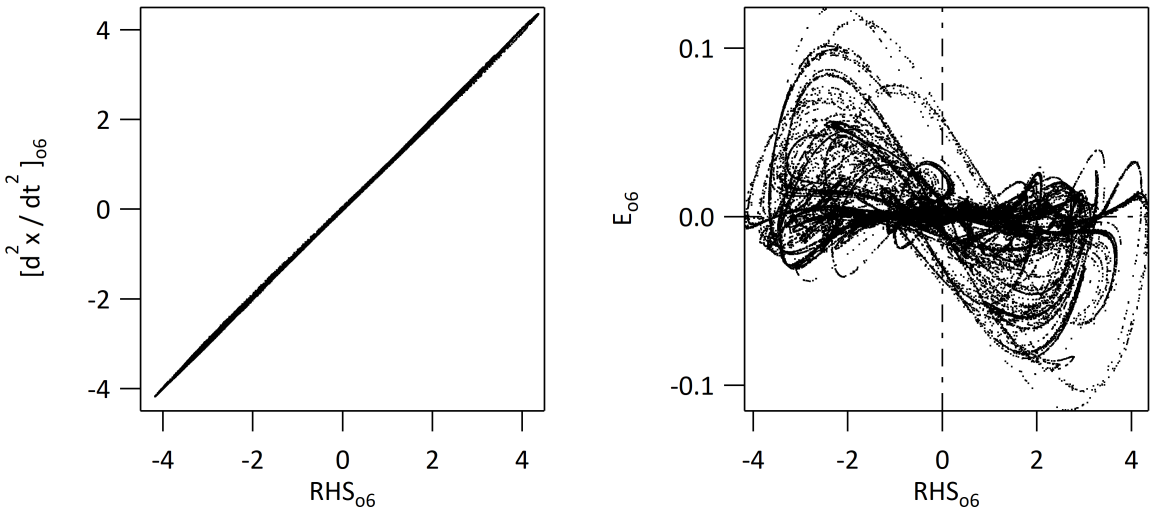

Figure 10: PoPe error for *case a* $\sigma_{chir} = 2.3$, $\nu = 0.8$, RK4 integration with $N = 2^8$ points per unit time, reconstruction with finite difference of order 6. Left hand side: Reconstructed second order derivative of x versus the Right Hand Side (RHS) of Eq.(22). Right hand side: Error E_{o6} versus the RHS.

integration scheme RK4 and number of steps per unit time $N = 2^8$. The second derivative of x, reconstructed with the finite difference scheme of order 6, is plotted versus the right hand side $\mathrm{RHS}_{ok,i}$ (the superscript (r) being omitted) of Eq.(22), Figure 10 left hand side. As expected for a computation with good accuracy, the points lie close to the diagonal. However, one can notice for this case with a low-resolution integration that a thickness is noticeable. Stepping to the error, hence the distance to the diagonal, Figure 10 right hand side, one finds that the error reaches 0.1 and exhibits a structure somewhat reminiscent of that of the strange attractor organised in self similar sheets, together with some form of symmetry regarding the amplitude and the sign. Some properties of the error are better seen when considering its logarithm, Figure 11 left hand side. One can notice that most of the data of $\log_{10}(|E_{o6}|)$ appears to lie in the range -2 ± 1, but excursions can be seen towards small errors while there seems to be a clear upper bound. The structure in the error is still visible, which underlines the fact that the error is not homogeneous. This is all the more visible that the number of points used here is large: it corresponds to the finite time integration of $N_t = 10^4$ unit times multiplied by the number of time steps per unit time $N = 2^8$. Increasing N governs an increase of the statistics, highlighting some details of the results. The histogram of the logarithm of the error contracts the heavy tail effect towards the large errors while expanding the region of small error. It is to be noted that the exponential reduction of the bin size towards the smallest errors drives an exponential reduction of the number of counts. An exponential fall-off of the number of counts towards the small values of $\log_{10}(|E_{o6}|)$ is then indicative of a near constant distribution of $|E_{o6}|$. The histogram of the error, Figure 11 right hand side, illustrates these characteristic features. Towards the large errors, the histogram indicates that the interpretation in terms of a maximum error appears to hold as highlighted by the sharp transition from close to maximum probability to near zero probability for a small increase of the error. Near the maximum of the histogram, a Gaussian like feature could describe the data. Localised peaks close to the maximum, more readily noticeable for a plot of

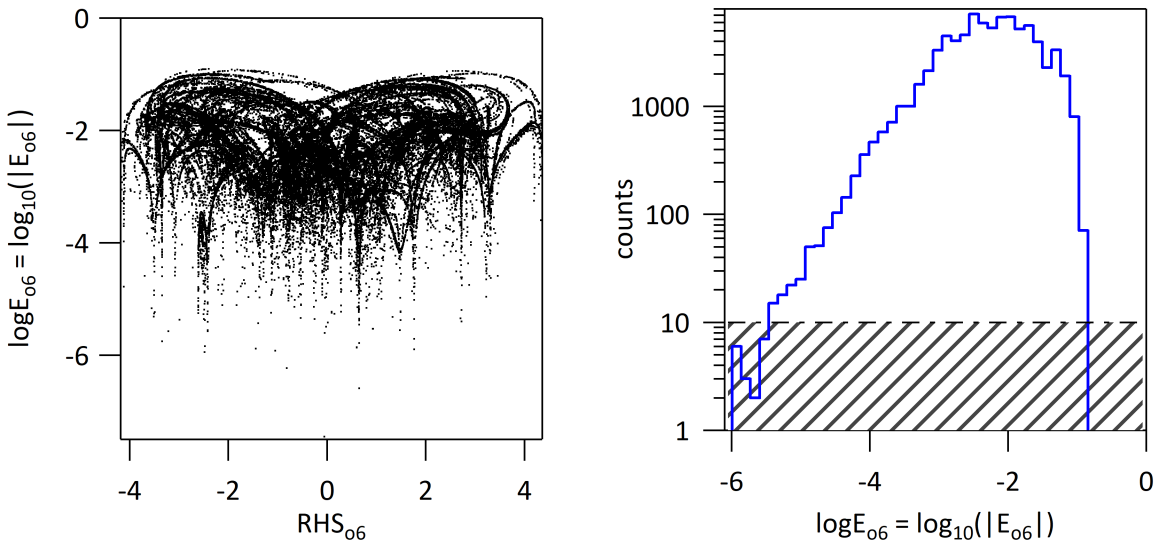

Figure 11: PoPe error for *case a*, $\sigma_{chir} = 2.3$, $\nu = 0.8$, RK4 integration with $N = 2^8$ points per unit time, reconstruction with finite difference of order 6. Left hand side: Left hand side $\log_{10}(|E_{o6}|)$ versus the RHS of Eq.(22). Right hand side, histogram of this error, the shaded area indicating the region with small statistics and consequent large relative fluctuations. All histograms are build using 5 bins par standard deviation.

the histogram in linear scale, could be reminiscent of the observed inhomogeneity of the error. Finally, towards the smallest errors, the histogram exhibits an exponential decay, hence the signature of a near constant distribution when the error tends towards zero. On the figure, the dashed region is that with reduced statistics, namely a number of counts smaller that 10 and therefore a typical statistical error of the order of $1/\sqrt{10}$.

The analysis of the error is made either by setting E as random variable or considering $logE = \log_{10}(|E|)$. The former is more sensitive to the large values and is sign dependent while the latter is sensitive to the small errors, ignoring their sign. However, as recalled above, the interpretation in terms of probability distributions is less straightforward for the latter given the changing bin size, which must be properly taken into account. For standard situations with small amplitude error, $|E| < 1$, the random variable $logE$ is negative. When computing the standard deviation $\delta logE$, one can decide for either signs. Usually it is defined as the square root of the variance, hence positive, but when comparing its value to the mean value $\langle logE \rangle$, negative in a standard case with small errors, the convenient choice is the negative sign. When considering the range of values $\langle logE \rangle \pm \delta logE$ the sign of $\delta logE$ is not an issue.

Of interest in the error analysis are particular dependences of the error, for instance variations in the phase space. Such an analysis can be performed by splitting the data according to a range of values of $J = dx/dt$. For each subset of the data one can determine the average $\langle logE \rangle$ and standard deviation $\delta logE$. The normalised error function can then be defined as $(logE - \langle logE \rangle)/\delta logE$ with the same meaning for each chosen range of values of J. This procedure provides a way to investigate the inhomogeneity of the error, Figure 12 left hand side. The data used to build this histogram is that of *case*

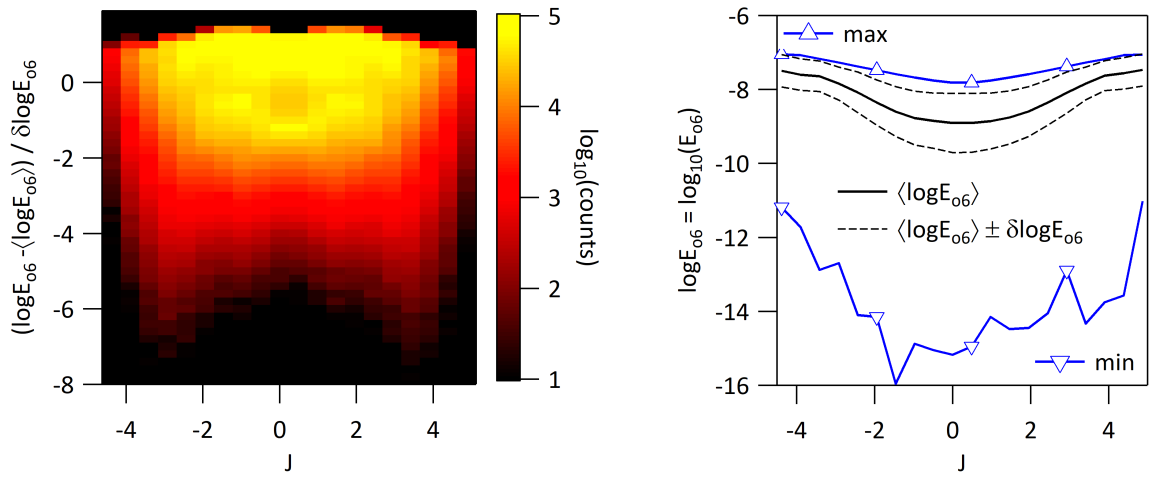

Figure 12: PoPe error for *case a*, $\sigma_{chir} = 7.0$ and $\nu = 0.2$, RK4 integration with $N = 2^8$ points per unit time, reconstruction with finite difference of order 6. Left hand side: Histograms for different values o J and normalised statistics. Right hand side, variation of the mean $\langle logE_{o6}\rangle$, max and min of $logE_{o6}$ as well as the characteristic width of the distribution function determined by the standard deviation $\delta logE_{o6}$, namely $\langle logE_{o6}\rangle \pm \delta logE_{o6}$.

a with control parameters $\sigma_{chir} = 7$, $\nu = 0.2$, with the RK4 time integration scheme and stepping with 2^{10} points per unit time and finally with an order 6 finite difference reconstruction scheme. One finds that the histogram exhibits a dependence on J combining a change in occurrences, these decreasing for larger $|J|$ as well as a structure with double peaking for smaller $|J|$. The latter can be indicative of further structure in the error, such as a dependence on x and on time, here understood as the phase shift in the time dependent potential. A full separation of such a 3D investigation of the error would help determining the origin of the error and means to improve the numerical scheme. However, for the simple problem at hand, brute force precision increase is possible and the need for such a detailed numerical analysis is not required.
The same analysis of the dependence on J is done for the mean $\langle logE\rangle$, standard deviation $\delta logE$ and maximum and minimum value, Figure 12 right hand side. The mean and maximum of $logE$ exhibit a comparable dependence on J with the larger values for the larger $|J|$. The standard deviation is also found to vary but the change is small. The largest variation is observed for the minimum value, which is the most sensitive to poor sampling. However, a trend to smaller error at small $|J|$ can also be seen regarding the latter. In the following, the dependences on phase space location of the points used in computing the error will not be taken into account. One has to keep in mind however that some aspects features of the result, for instance the cut-off at large error, can be related to an underlying inhomogeneity. As a final remark, one can remark that $\langle logE\rangle + \delta logE$, is comparable to the maximum value that can be achieved.

In the reconstruction process, we have underlined the need to use a scheme with better or at least equivalent precision to that of the code. This is tested by comparing different orders of the finite difference schemes used to reconstruct the second time derivative of x from the time trace of x provided by the code output. One can then compare the

histograms of the error obtained for each reconstruction procedure. As shown on Figure 13 left hand side, the histograms obtained for the order 6 and order 8 reconstruction are identical. The error is therefore checked to be generated by the code and not the reconstruction scheme with finite difference of order 6 and 8. Conversely, for the chosen RK4 simulation of *case a*, the order 4 and order 2 reconstruction schemes lead to different histograms, these being shifted towards the large errors. In these two cases the error of the code is less important than that of the reconstruction schemes and verification cannot be achieved. The results of this figure have been obtained with 2^{10} points per unit time and are averages over the 20 points describing the J-dependence illustrated on Figure 12 left hand side. Compared to the histogram Figure 11 obtained at low resolution, one can remark the sharp cut-off at highest error, a structure in the vicinity of the maximum and the exponential fall-off towards the smallest errors as expected for a constant distribution with exponential reduction of bin size.

The analysis of the error $logE$ can also be used to recover the order of the integration scheme of the code, Figure 13 right hand side. For *case a* $\sigma_{chir} = 7$, $\nu = 0.2$ the RK4 and RK2 schemes are compared. One finds that the error $logE$ scales with the expected scaling, N^{-4} for RK4 and N^{-2} for RK2. Furthermore, as for the error analysis with the Method of Return Solution Figure 6, one finds that the actual value of the error depends on the case that is investigated. This is shown by *case b* $\sigma_{chir} = 2.3$, $\nu = 0.8$ and RK4 scheme which exhibits the RK4 scaling N^{-4} but with a smaller error, typically by two and three orders of magnitude. As observed in the test of the Runge Kutta schemes Figure 47 one can also notice an increase of the error at largest values of N when the error drops to the level of machine precision. The aim of the PoPe analysis is to provide a figure of merit in terms of accuracy of a given production simulation. The average logarithm of the error $\langle logE \rangle$ plotted on Figure 13 right hand side can be regarded as such a figure of merit. The smaller $\langle logE \rangle$ the more accurate the simulation. However, a crucial point is then to provide a criterion to assess that a simulation is acceptable, which is an issue since computer resources give access to finite accuracy simulations. One can readily consider that $\langle logE \rangle \gtrsim 0$, hence a mean error exceeding 100%, is a criterion to reject simulations. One then finds that for the control parameter $\sigma_{chir} = 7$, $\nu = 0.2$ the RK4 simulation with $N = 2^3$ and the RK2 simulations with $N = 2^3$ and $N = 2^4$ can be considered too inaccurate and rejected on the basis of this criterion. Compared to the set of simulations that yield a wrong Lyapunov exponent, Figure 8, and an inappropriate strange attractor structure in phase space, Figure 7, the present chosen criterion only reject 3 out of 6 simulations identified as being with poor resolution and generating misleading results. Of course, one could tune the threshold value for rejection, but with the likely result that this critical value is case dependent, hence yielding a criterion without universality. There is therefore a need for a more effective criterion. One possibility is to take into account the features of the histogram of the error $logE$, Figure 14. Given the standard deviation $\delta logE$, which is negative since $logE$ is related to the logarithm of the error, a more appropriate criterion would be to reject simulations such that $\langle logE \rangle + \delta logE \gtrsim 0$. The RK2 simulation with $N = 2^5 = 32$ is then added to the previous list. For the RK2-simulation with $N = 2^6 = 64$ one can notice that the maximum value of $logE$ is larger than zero, while $\langle logE \rangle + \delta logE \gtrsim -0.3$, hence an error E larger than 50%. Given the sharp cut-off towards the large values of $logE$, it appears reasonable to extend the rejection criterion to this simulation. Still, one simulation with a wrong Lyapunov exponent, the RK2 run

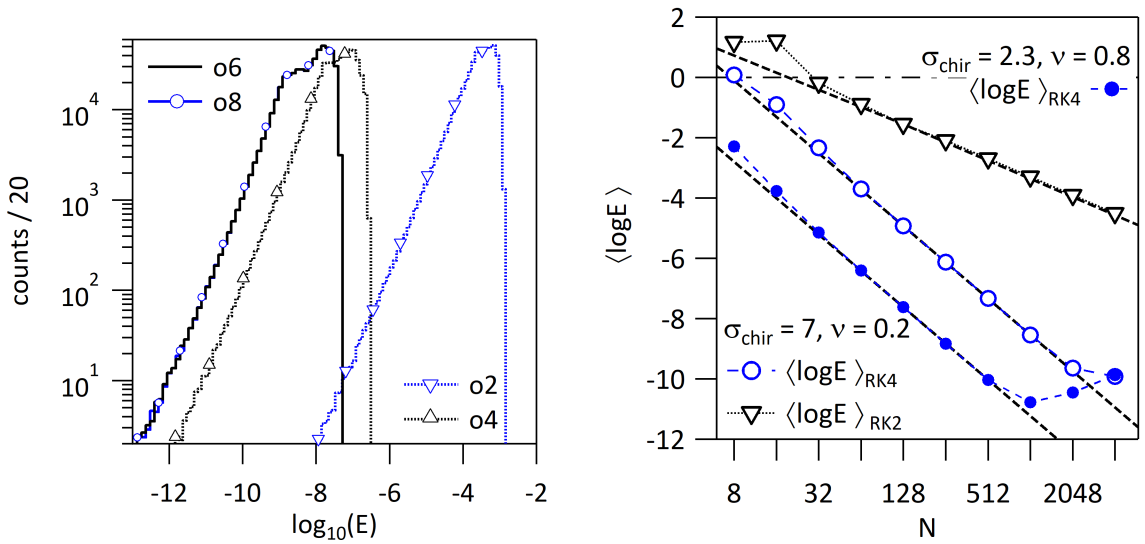

Figure 13: Left hand side: PoPe error for *case a* $\sigma_{chir} = 7$ and $\nu = 0.2$, RK4 integration with $N = 2^10$ points per unit time, comparison of the histograms of $logE$ obtained with finite difference scheme of order 2 head down triangles dashed blue curve, and order 4 head up triangles dashed black curve, order 6 black plain line and order 8 blue plain line open blue circles. Right hand side: variation of the mean error $\langle logE \rangle$ with the number of steps of the integration scheme for the Runge Kutta schemes of order 2, black curve head down open triangles, and order 4, blue curves with circles and for the control parameters of *case b*, $\sigma_{chir} = 2.3$ and $\nu = 0.8$, closed symbols, and *case a*, $\sigma_{chir} = 7$ and $\nu = 0.2$ open symbols.

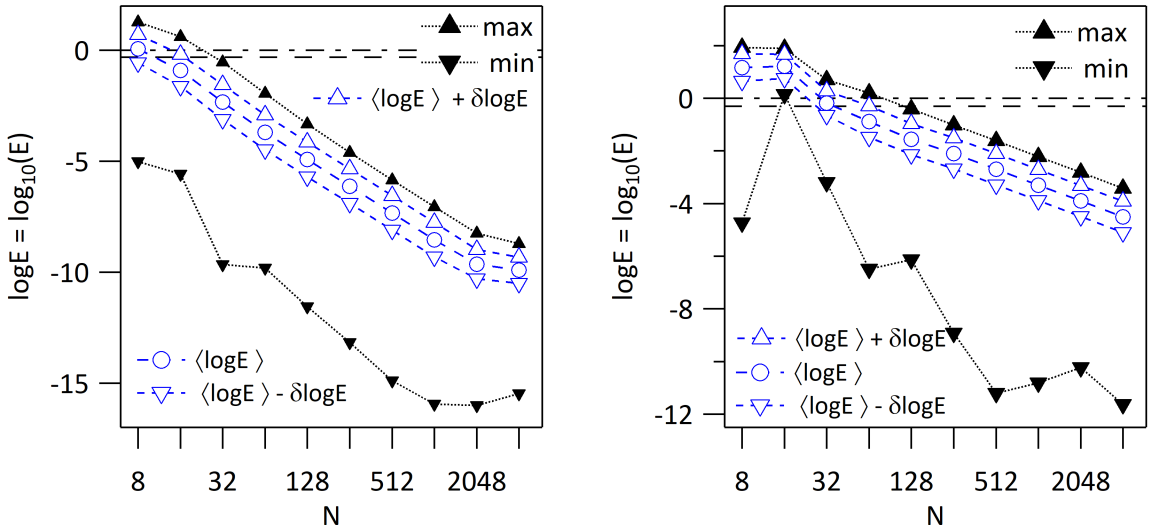

Figure 14: Key features of the histogram of the error $logE$, the mean $\langle logE \rangle$ blue open circles, the mean value plus the standard deviation $\langle logE \rangle + \delta logE$, blue open head-up triangles, the mean value minus the standard deviation $\langle logE \rangle - \delta logE$, blue open head-down triangles, and finally minimum and maximum of the distribution of $logE$, respectively head-down full black triangles, and head-up full black triangles. The dashed horizontal line locates $logE \approx -0.3$ hence an error E of 50%. Left hand side: *case a*, $\sigma_{chir} = 7$ and $\nu = 0.2$ RK4 integration scheme. Right hand side: *case a*, $\sigma_{chir} = 7$ and $\nu = 0.2$, RK2 integration scheme.

with $N = 2^7$, is not excluded by this extended criterion.
We find therefore that with PoPe analysis of the error, it is possible to define a criterion based on the magnitude of the error to exclude simulations with accuracy smaller than a prescribed limit. However, we have found that there is not clear-cut way to assess the exactness of the physics for simulations with reduced precision. The simulations with order 2 Runge Kutta integration for *case a* $\sigma_{chir} = 7$, $\nu = 0.2$ exemplify this issue. While the simulation with $N = 2^6$ exhibits a correct Lyapunov exponent and phase space portrait, the simulation with $N = 2^7$ has a wrong Lyapunov exponent and different phase space portrait. Characterising the accuracy with $\langle logE \rangle + \delta logE$ as figure of merit, one finds $\langle logE \rangle + \delta logE \approx -0.30$ for the former and $\langle logE \rangle + \delta logE \approx -0.96$ for the latter.

The analysis of the PoPe error performed in this Section yields useful insight into the accuracy of the simulations that are performed. One recovers the verification results obtained with the MRS in Section 3.2. One also finds that the error is not homogeneous in the phase space, being larger for large momentum J than for $J \approx 0$. A similar behaviour is observed for kinetic simulations of turbulence with larger error at large velocity [9, 10]. The statistics of the error E are found to be close to Gaussian near zero error with an apparent sharp cut towards the larger values of $|E|$. At this stage, the magnitude of the PoPe error does not appear to provide a robust and universal criterion that would allow identifying systematically the simulations that have too poor resolution.

4.2 Projection of the error, PoPe verification

4.2.1 Simplified PoPe analysis: 2 operator reduction

Given the computed error, the proposed way to evaluate the accuracy with PoPe is to determine the class of equations that yield a comparable behaviour and that cannot be discriminated. Let us rewrite the system Eq.(23) in terms of two operators $O_{1,2}$ and O_3.

$$E_{ok,i} = \left[\frac{d^2x}{dt^2}\right]^{(r)}_{ok,i} - \mathrm{RHS}^{(r)}_{ok,i} \tag{24a}$$

$$\mathrm{RHS}^{(r)}_{ok,i} = O^{(r)}_{1,2,ok,i} + \ O^{(r)}_{3,ok,i} + \mathrm{R}^{(r)}_{ok,i} \tag{24b}$$

$$O^{(r)}_{1,2,ok,i} = 2\pi\ B\Big(\sin\big(2\pi x_i\big) + \sin\big(2\pi(x_i - t_i)\big)\Big) \tag{24c}$$

$$O^{(r)}_{3,ok,i} = \nu\ J_i \tag{24d}$$

As highlighted by the notation the former operator $O_{1,2}$ is in fact the sum of the operators identified as $O_1 = 2\pi\ B\sin\big(2\pi x_i\big)$ and $O_2 = \pi\ B\sin\big(2\pi(x_i - t_i)\big)$. The reduction to two operators and the possibility of defining the relevant operators to be addressed by the PoPe verification scheme is part of the freedom and versatility of the method. Beyond simplifying the presentation of the results, the choice made in splitting the operators can be seen as governed by the properties of these operators. Indeed, both O_1 and O_2 are computed analytically given x_i and t_i, while O_3 is reconstructed with a finite difference scheme. The label i is the index of the saved data of a given simulation, ranging typically from 1 to N_{max}. In the present subsection the reference to the order of the reconstruction scheme, order k labelled ok and the superscript (r) are omitted to simplify the notations. We now want to determine the coefficients $\delta c_{O_{1,2}}$ and δc_{O_3} as well as the residue R defined by:

$$E_i = \delta c_{O_{1,2}} O_{1,2,i} + \delta c_{O_3} O_{3,i} + \mathrm{R}_i \tag{25a}$$

$$E = \delta c_{O_{1,2}} O_{1,2} + \delta c_{O_3} O_3 + \mathrm{R} \tag{25b}$$

In Eq.(25a) the two coefficients $\delta c_{O_{1,2}}$ and δc_{O_3} are defined as independent of the realisation i of the error E_i. One can then define the vector $E = \{E_i\}$ as an N_{max}-dimension vector with components E_i, similarly for $O_{1,2}$, O_3 and R. Equation (25b) is then the vector form of Eq.(25a) for each vector component. This equation can be understood as the projection of E on the two vectors $O_{1,2}$ and O_3 plus the vector R which stands for the part of E with zero projection on $O_{1,2}$ and O_3. Let us use the notation $\langle E|O\rangle$ for the projection of E on O, one can then split Eq.(25b) into:

$$\delta c_{O_{1,2}} \langle O_{1,2}|O_{1,2}\rangle + \delta c_{O_3} \langle O_3|O_{1,2}\rangle = \langle E|O_{1,2}\rangle \tag{26a}$$

$$\delta c_{O_{1,2}} \langle O_{1,2}|O_3\rangle + \delta c_{O_3} \langle O_3|O_3\rangle = \langle E|O_3\rangle \tag{26b}$$

Provided the projection is actually defined, then the system (26) is a set of two coupled linear equations with unknowns $\delta c_{O_{1,2}}$ and δc_{O_3} that can readily be solved provided the determinant is different from zero, namely that the two vectors $O_{1,2}$ and O_3 are not co-linear.

$$\langle O_{1,2}|O_{1,2}\rangle \langle O_3|O_3\rangle - \langle O_3|O_{1,2}\rangle^2 \neq 0 \tag{27a}$$

Given $\delta c_{O_{1,2}}$ and δc_{O_3} solution of Eq.(26), the residue is then computed as the part of the error that is not aligned along $O_{1,2}$ or O_3, R $= E - \delta c_{O_{1,2}} O_{1,2} - \delta c_{O_3} O_3$. When the system is solved with no error, hence $E = 0$, one finds $\delta c_{O_{1,2}} = \delta c_{O_3} = 0$ and R $= 0$. The two coefficients $\delta c_{O_{1,2}}$, δc_{O_3} and the residue R therefore characterise the numerical error.

One can first remark that one only needs two linear equations of the form Eq.(25a) to determine a set of coefficients $\delta c_{O_{1,2}}$ and δc_{O_3}. Let us consider one of the possible pair (i, j) of points belonging to the strange attractor, one can determine the coefficients $\delta c_{O_{1,2}}(i, j)$ and $\delta c_{O_3}(i, j)$ associated to the pair (i, j). Considering several pairs (i, j) then determines an ensemble of values for the pair $(\delta c_{O_{1,2}}, \delta c_{O_3})$ which can be analysed statistically. This procedure holds insofar that the points i and j are not co-linear, hence:

$$O_{1,2,i} O_{3,j} - O_{3,i} O_{1,2,j} \neq 0 \tag{27b}$$

In practise, the issue of co-linearity can occur when the determinant is small, hence when $(O_{1,2,i}, O_{3,i})$ is close to being co-linear to $(O_{1,2,j}, O_{3,j})$ but the error (E_i, Ej) is not aligned on these vectors. Since this property is governed by the error, some randomness in this difficulty can be expected. The generation of spurious values for $(\delta c_{O_{1,2}}, \delta c_{O_3})$ is therefore expected as a consequence of co-linearity but all the cases characterised by a small determinant will not lead to large values of $(\delta c_{O_{1,2}}, \delta c_{O_3})$ that are obviously not correct.

A means to overcome this issue is to define the projection scheme based on the method of least square minimisation. For the present example, one defines the relative position d_i as:

$$d_i = E_i - \delta c_{O_{1,2}} O_{1,2,i} - \delta c_{O_3} O_{3,i} \tag{28a}$$

and one then determines the coefficients $\delta c_{O_{1,2}}$, δc_{O_3} as those minimising the distance:

$$\tfrac{1}{2} d^2 = \tfrac{1}{2} \sum_i d_i^2 = \tfrac{1}{2} \sum_i \Big[E_i - \delta c_{O_{1,2}} O_{1,2,i} - \delta c_{O_3} O_{3,i} \Big]^2 \tag{28b}$$

Setting the derivatives of d^2 with respect to $\delta c_{O_{1,2}}$ and δc_{O_3} to be equal to zero, one obtains:

$$\delta c_{O_{1,2}} \Big[\sum_i O_{1,2,i}^2 \Big] + \delta c_{O_3} \Big[\sum_i O_{1,2,i} O_{3,i} \Big] = \Big[\sum_i O_{1,2,i} E_i \Big] \tag{29a}$$

$$\delta c_{O_{1,2}} \Big[\sum_i O_{1,2,i} O_{3,i} \Big] + \delta c_{O_3} \Big[\sum_i O_{3,i}^2 \Big] = \Big[\sum_i O_{3,i} E_i \Big] \tag{29b}$$

If a single point is chosen the two equations Eq.(29a) and Eq.(29a) are identical. The summation must therefore be made with at least two points and can be extended up to all available points. The latter limit corresponds to the calculation with the scalar product introduced above in an N_{max} dimension space. For the other situations we define an m-dimension space scalar product of two vectors:

$$\langle F | G \rangle_m = \sum_{i=j_1}^{j_m} F_i G_i \tag{30}$$

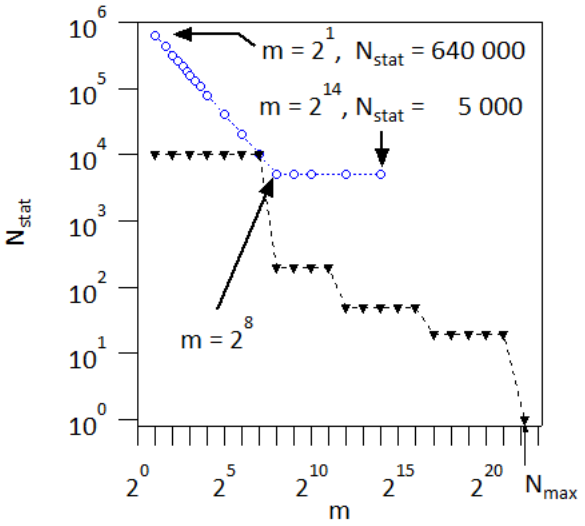

Figure 15: N_{stat}, the number of m-tuples used for the statistics. Two procedures are used, a set with N_{stat} large enough to determine histograms, blue line with open circles, and a series with reduced statistics to only determine the mean and standard deviation, black head-down triangles. The latter series is completed by the calculation for $m = N_{max}$ yielding a single value, thus equal to the mean with zero standard deviation.

The subscript m that is added in these notation stands for the number of points from j_1 to j_m that are used in the sum. The scalar product and the solution can also depend on the choice that is made for the m-tuples. The latter freedom of choice will be used in the following to make statistics on the results at given number of points m but different choices of m-tuples. With only two points $m = 2$ one can show that the problem of co-linearity is identical to that discussed above and one can expect that as m is increased, the weight of the co-linearity generating outliers in the results will be decreasing.

We first investigate the impact of the choice of the number of points m that are used in the summation defining the scalar product; equivalently the dimension of the space where the vectors E, $O_{1,2}$, O_3 are defined. As first indicator, we consider $\left\langle \delta c_{O_{1,2}} \right\rangle_{m,N_{stat}}$, hence the average value of the coefficient $\delta c_{O_{1,2}}$. In the chosen simulation, the time series is of length N_{max} such that $N_{max} = 5\,119\,993$. There is therefore a very large freedom in choosing random m-tuples in this set whenever $m \ll N_{max}$. Let us define N_{stat} the number of chosen m-tuples that also determines the size of the data set used for the statistics on the error. Two different procedures have been used to fix N_{stat} and consequently investigate the statistics of the coefficients $\delta c_{O_{1,2}}$ and δc_{O_3} Figure 15. In a first set of verification tests labelled *ext*, the number N_{stat} of randomly chosen m-tuples is first fixed in such a way that $m \times N_{stat} = N_{max}/4$ for m ranging from 2^1 to 2^8. When m becomes large, $2^8 \leq m \leq 2^{14}$, N_{stat} is maintained constant, $N_{stat} = N_{max}/(4 \times 2^8)$ to have a sufficiently large data set for the statistics. In this set, the number of points involved in the calculation is first maintained constant as m is increased, but then increases proportionally to m. This first procedure is depicted by the points of N_{stat} versus m with blue open circles on Figure 15 and labelled *ext*. A second series of m-tuples is made with reduced statistics, and labelled *red*. In this procedure only the mean and standard deviation are computed and a reduced set of N_{stat} points is sufficient for such a purpose. Depending on the value of

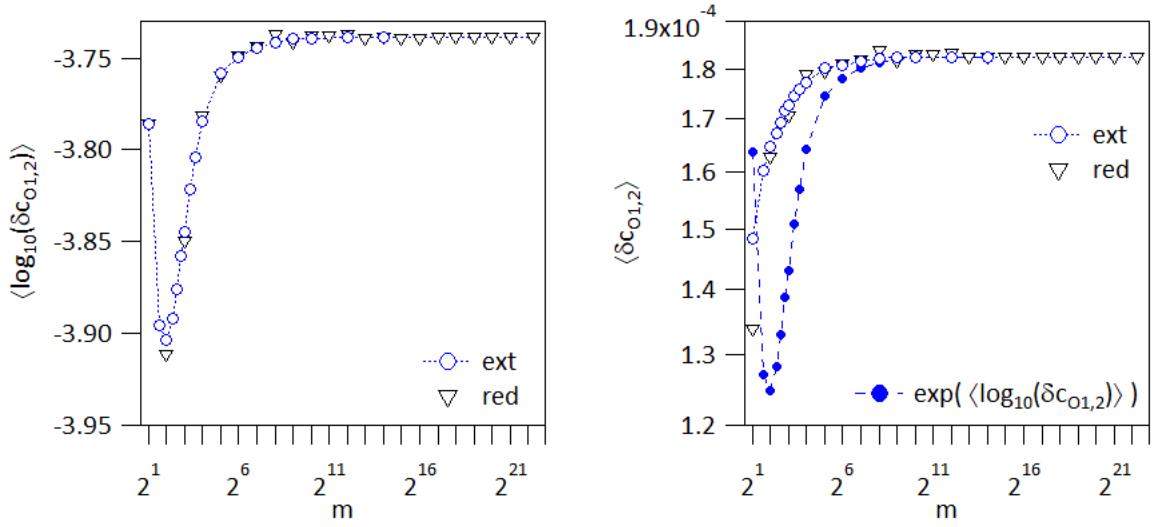

Figure 16: For the operator $O_{1,2}$, effect of the number of m-tuples on the statistical results, blue open circle labelled *ext* for extended data bases with N_{stat} large, black head-down triangles labelled *red* for reduced data bases, hence N_{stat} small. Left hand side: mean value of $\log_{10}(|\delta c_{O_{1,2}}|)$. Right hand side: mean value of $\delta c_{O_{1,2}}$ and comparison to $\exp(\langle \log_{10}(|\delta c_{O_{1,2}}|)\rangle)$, closed blue circles.

m, different values of N_{stat} have been chosen. For this second series of data selection, the points (m, N_{stat}) are indicated by black head-down open triangles on Figure 15. This series is completed by the calculation for $m = N_{max}$ yielding a single value, thus equal to the mean with zero standard deviation. One finds that either procedures to determine the points used in the calculation lead to similar values of the mean and standard deviation.

4.2.2 PoPe verification of the drive operator $O_{1,2}$

The statistics are performed both for the random variables $\delta c_{O_{1,2}}(m)$ and $\log_{10}(|\delta c_{O_{1,2}}(m)|)$. The latter data is less sensitive to outliers with very large values and more sensitive to the very small values of the coefficients. These statistics are applied to the RK2 run with resolution $N = 2^9$ steps per unit time for *case a*: $\sigma_{chir} = 7$, $\nu = 0.2$. The mean value of $\delta c_{O_{1,2}}$ is first addressed, Figure 16, left hand side with statistics on $\log_{10}(|\delta c_{O_{1,2}}|)$ and, right hand side statistics on $\delta c_{O_{1,2}}$. For both cases one readily finds a convergence of the error as m is increased, the value for the limit $m = N_{max}$ being identical for the two statistics since only one value is available. One can also notice that the investigation with the reduced statistics, labelled by black head-down triangles, yields appropriate results for large values of m, the reduced number of points for these statistics being compensated by the large number of data points used for the least square calculation. One can also notice that the variation of the error from $m = 4$ to $m \approx 2^8$ is of the order of 30%. This variation is observed for $\langle \log_{10}(|\delta c_{O_{1,2}}|)\rangle$, Figure 16 left hand side, and for $\langle \delta c_{O_{1,2}}\rangle$, Figure 16 right hand side. Although similar, the statistic on $\delta c_{O_{1,2}}$ and $\log_{10}(|\delta c_{O_{1,2}}|)$ yield different mean values for $m \leq 2^8$, Figure 16 right hand side. The variation of the mean values with m is observed to become small for $m > 2^8$, and the values for the reduced and extended data bases are found to agree. Furthermore, the obtained value in this range of m does not seem to depend on the way the analysis is performed, both the statistics on $\delta c_{O_{1,2}}$ and

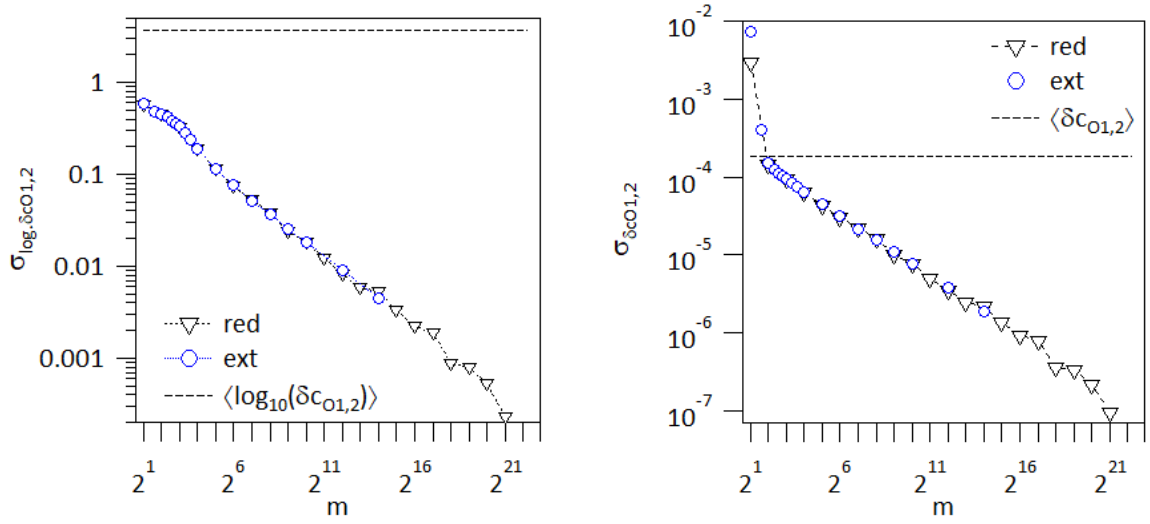

Figure 17: For the operator $O_{1,2}$, effect of the number of m-tuples on the standard deviation, blue open circle, labelled *ext* N_{stat} large, hence large data bases, and for reduced data bases, N_{stat} small, black head-down triangles labelled *red*. Left hand side: standard deviation $\sigma_{log.\delta cO1,2}$ of $\log_{10}(|\delta c_{O_{1,2}}|)$ and comparison to $|\left\langle \log_{10}(|\delta c_{O_{1,2}}|)\right\rangle|$, dashed black line towards the top of the Figure. Right hand side: standard deviation $\sigma_{\delta cO1,2}$ of $\delta c_{O_{1,2}}$ and comparison to $|\left\langle \delta c_{O_{1,2}}\right\rangle|$ dashed black line.

$\log_{10}(|\delta c_{O_{1,2}}|)$ leading to the same value for the mean of $\delta c_{O_{1,2}}$. Finally it is important to underline that the sign of the error on the coefficient of $O_{1,2}$ is given by the statistics on $\delta c_{O_{1,2}}$ and is found to be positive.

The dependence of the standard deviation on m provides a better insight into the changes governed by increasing m, Figure 17. For $\log_{10}(|\delta c_{O_{1,2}}|)$, Figure 17 left hand side, one finds that $|\left\langle \log_{10}(|\delta c_{O_{1,2}}|)\right\rangle| > \sigma_{log.\delta cO1,2}$ for all values of m and that $\sigma_{log.\delta cO1,2}$ decays exponentially with m. Here $\sigma_{log.\delta cO1,2}$ is chosen positive for convenience. Regarding the standard deviation of $\delta c_{O_{1,2}}$, Figure 17 right hand side, one can also observe an exponential decay of $\sigma_{\delta cO1,2}$ for $m \geq 2^2$, which coincides with the point where $|\left\langle \delta c_{O_{1,2}}\right\rangle| \geq \sigma_{\delta cO1,2}$.

Let us now complete this investigation by considering the statistical fluctuations governed by the number of samples that are used, in particular for the scheme with reduced data sets *red*, Figure 18. On Figure 18 left hand side the mean value of $\log_{10}(|\delta c_{O_{1,2}}|)$ is plotted versus m of the m-tuples used in the least square method. The dashed region corresponds to the high probability region of the distribution of $\log_{10}(|\delta c_{O_{1,2}}|)$ between the mean plus the standard deviation and the mean minus the standard deviation. One can note the exponential narrowing of this region towards the mean value. The mean value, blue line with closed circles, is computed with the extended data base while the region of high probability is determined with the reduced data scheme which allows one to have data for $m \geq 2^{14}$. The same data is plotted on Figure 18 right hand side with a zoom for $m \geq 2^8$ and around the asymptotic value of the mean $|\left\langle \log_{10}(|\delta c_{O_{1,2}}|)\right\rangle| \approx -3.739$, dashed blue line. The shaded region within $\left\langle \log_{10}(|\delta c_{O_{1,2}}|)\right\rangle + \sigma_{log.\delta cO1,2}$, closed head-up triangles and $\left\langle \log_{10}(|\delta c_{O_{1,2}}|)\right\rangle - \sigma_{log.\delta cO1,2}$, closed head-down triangles is determined with the reduced data base which provides data for $m \geq 2^{14}$. Five different randomly chosen

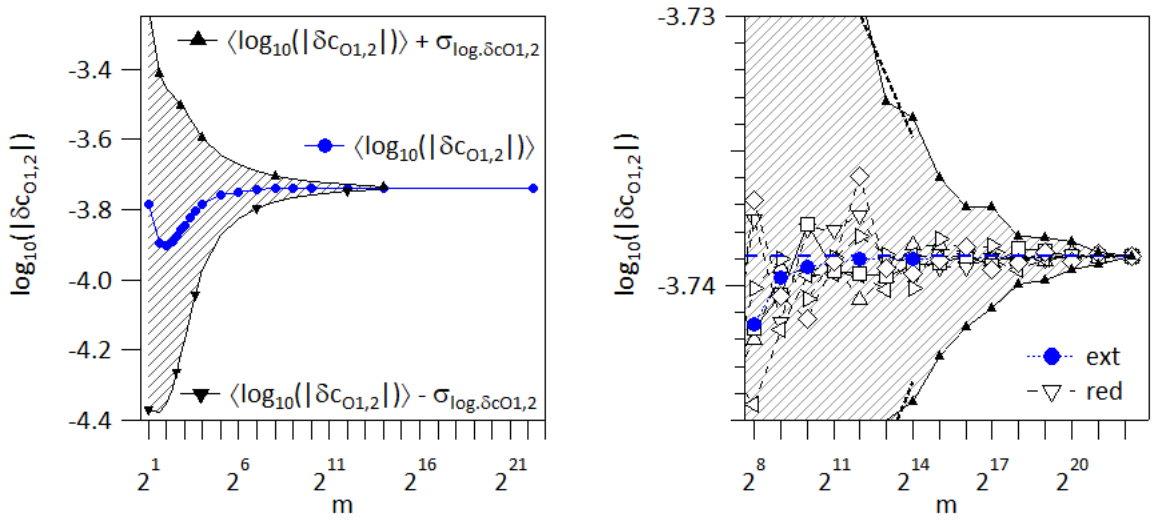

Figure 18: For the operator $O_{1,2}$, effect of the number of m-tuples on the region of highest likelihood of $\log_{10}(|\delta c_{O_{1,2}}(m)|)$, hence between $\left\langle \log_{10}(|\delta c_{O_{1,2}}|)\right\rangle + \sigma_{log.\delta cO1,2}$, closed head-up triangles, and $\left\langle \log_{10}(|\delta c_{O_{1,2}}|)\right\rangle - \sigma_{log.\delta cO1,2}$, closed head-down triangle. The mean value $\left\langle \log_{10}(|\delta c_{O_{1,2}}|)\right\rangle$ is plotted with blue closed circle, $\sigma_{log.\delta cO1,2}$ is the standard deviation chosen positive here. Left hand side: data for the full range of m, $2^1 \leq m \leq N_{max}$. Right hand side, zoom for $m \geq 2^8$ with data from five different random choices of the m-tuples, open symbols.

m-tuples are also plotted using different open markers.

For the operator $O_{1,2}$ Eq.(24c), one finds that when increasing the number of m-tuples used to determine the error $\delta c_{O_{1,2}}$ of the coefficient $c_{O_{1,2}}$, then $\delta c_{O_{1,2}}$ converges towards a well-defined value. The standard deviation of the statistics decreases exponentially with m. The asymptotic value is $\delta c_{O_{1,2}} = 1.824\ 10^{-4}$. It means that the output data for the chosen simulation has the best agreement with the evolution equation when the weight of the operator $O_{1,2}$ is $1 + 1.824\ 10^{-4}$. The relative error $\delta c_{O_{1,2}} = 1.824\ 10^{-4}$ with respect to the input parameter B, leads to a relative error of the Chirikov parameter of $\frac{1}{2} 1.824\ 10^{-4}$.

4.2.3 PoPe verification of the damping operator O_3

For the operator O_3, we first analyse the standard deviation of the relative error δc_{O_3} on the weight of operator O_3, Figure 19. Regarding $\log_{10}(|\delta c_{O_3}|)$, the standard deviation $\sigma_{\log.\delta cO3}$ exhibits a quite different behaviour from that reported for $\sigma_{\log.\delta cO1,2}$, Figure 17 left hand side. Indeed the standard deviation is observed to be nearly constant for $\sigma_{\log.\delta cO3}$ and $m \leq 2^{15}$ while it exhibits an exponential decrease over the whole range of values of m when considering the variable $\log_{10}(|\delta c_{O_{1,2}}|)$. For $m > 2^{15}$, one can observe an exponential decrease of $\sigma_{\log.\delta cO3}$, Figure 17 left hand side. For the standard variable δc_{O_3}, one finds that $\sigma_{\delta cO3}$ decreases over the whole range of values of m and exhibits a constant rate exponential decrease for $2^2 \leq m \leq 2^{17}$. However, the standard deviation then exceeds the mean for $m < 2^{13}$. When analysing the mean value of $\log_{10}(|\delta c_{O_3}|)$, Figure 20 left hand side, open blue circles for the extended statistics, and black open

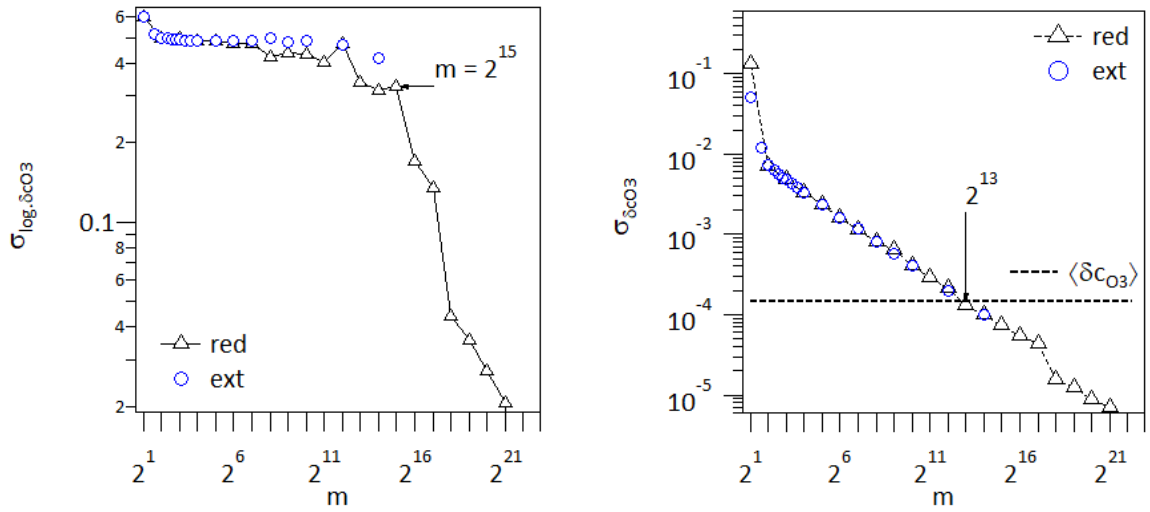

Figure 19: Effect of the number of m-tuples on the standard deviation σ with the extended statistics, *ext* blue open circles, and reduced statistics, *red* black triangles for the operator O_3. Left hand side: Statistics on $\log_{10}(|\delta c_{O_3}|)$, standard deviation $\sigma_{log.\delta cO3}$ versus m. The transition from near constant standard deviation to roughly an exponential decrease occurs for $m \approx 2^{15}$. Right hand side: Statistics on δc_{O_3}, standard deviation $\sigma_{\delta cO3}$ versus m. The asymptotic value of the mean $\langle \delta c_{O_3} \rangle$ is plotted with a dashed black line. The standard deviation becomes smaller than the mean value, $\sigma_{\delta cO3} / \langle \delta c_{O_3} \rangle \leq 1$, for $m \geq 2^{13}$.

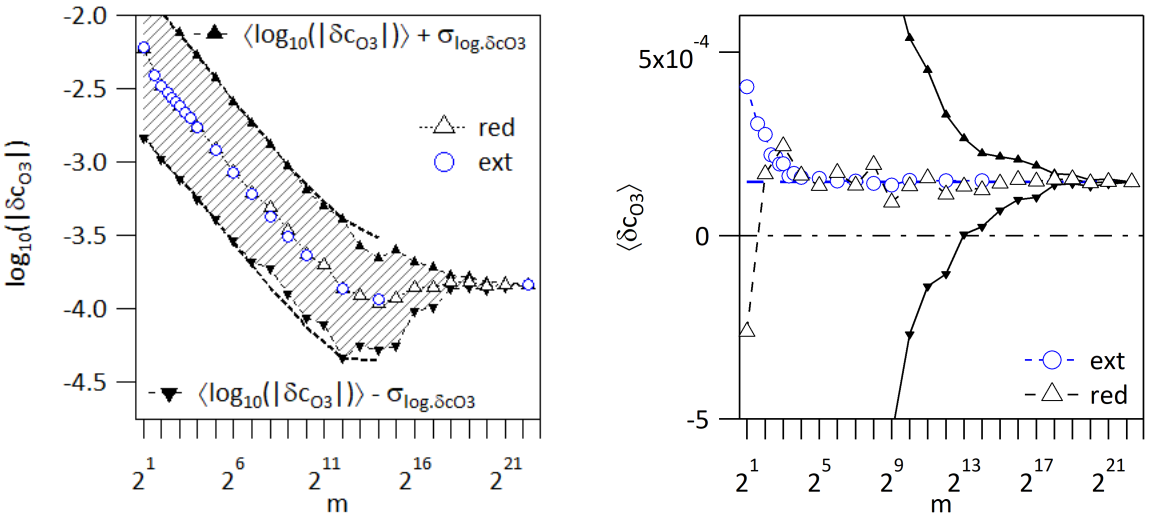

Figure 20: For the operator O_3, effect of the number of m-tuples on the mean value of the coefficient using the extended statistics, *ext* blue open circles, and reduced statistics *red* black triangles. Left hand side: Mean value $\langle \log_{10}(|\delta c_{O_3}|) \rangle$ and region with highest probability between $\langle \log_{10}(|\delta c_{O_3}|) \rangle + \sigma_{log.\delta cO3}$ closed head-up triangles, and $\langle \log_{10}(|\delta c_{O_3}|) \rangle - \sigma_{log.\delta cO3}$, closed head-down triangle. Right hand side: Mean value $\langle \delta c_{O_3} \rangle$ and region with highest probability between $\langle \delta c_{O_3} \rangle + \sigma_{\delta cO3}$ closed head-up triangles, and $\langle \delta c_{O_3} \rangle - \sigma_{\delta cO3}$, closed head-down triangle.

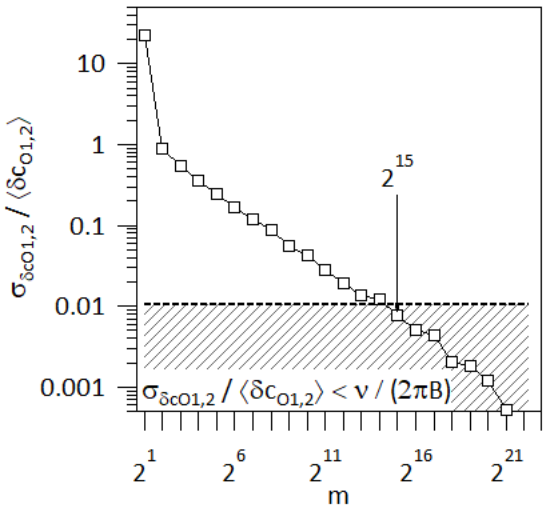

Figure 21: Criterion on the required precision in determining $\delta c_{O_{1,2}}$ to avoid that its fluctuations $\sigma_{\delta cO1,2}$ govern the error of δc_{O_3}. This effect is more important when the typical magnitude operator $O_{1,2}$ is larger than that of operator O_3, $||O_3||/||O_{1,2}|| \approx 0.0104$ in the present simulation.

triangles for the reduced statistics, one can observe first an exponential decrease as m is increased, together with a variation of more than one order of magnitude of δc_{O_3} for $2^1 \leq m \leq 2^{12}$. The mean value $\langle \log_{10}(|\delta c_{O_3}|) \rangle$ is then roughly constant and equal to its asymptotic value: $\langle \log_{10}(|\delta c_{O_3}|) \rangle = -3.837$. The latter range of values corresponds to that with decreasing standard deviation. Prior to this transition, the standard deviation is more or less constant, see Figure 20 left hand side. The lines $\langle \log_{10}(|\delta c_{O_3}|) \rangle + \sigma_{log.\delta cO3}$, black line head-up closed triangles, and $\langle \log_{10}(|\delta c_{O_3}|) \rangle - \sigma_{log.\delta cO3}$, black line head-down closed triangles, are then parallel to the variation of the mean which indicates similarity in the distribution function of the error, Figure 20 left hand side. If one now considers the statistics of δc_{O_3}, Figure 20 right hand side, one finds a first regime for $m \leq 2^4$ with large variation of $\langle \delta c_{O_3} \rangle$. The values then seem to settle close to the asymptotic value $\langle \delta c_{O_3} \rangle \approx 1.455\ 10^{-4}$. However, when analysing the region with highest probability, hence between $\langle \delta c_{O_3} \rangle + \sigma_{\delta cO3}$ black line closed head-up triangles, and $\langle \delta c_{O_3} \rangle - \sigma_{\delta cO3}$ black line closed head-down triangles, one finds that $\langle \delta c_{O_3} \rangle - \sigma_{\delta cO3}$ only becomes positive for $m \geq 2^{13}$.

One finds that recovering converged values for the effective weight of operator O_3 is more demanding than for operator $O_{1,2}$. Suitable precision for the operator O_3 is only reached when very precise values are obtained for operator $O_{1,2}$.

4.2.4 Error contamination of the low amplitude operator

The error δc_{O_3} on the weight of the damping operator O_3 is typically $\langle \delta c_{O_3} \rangle \approx 1.455\ 10^{-4}$. It is found to be quite comparable to that on $\delta c_{O_{1,2}}$, $\langle \delta c_{O_{1,2}} \rangle \approx 1.824\ 10^{-4}$. These errors have the same sign and comparable magnitude, which is consistent with the fact that the error is stemming from the numerical time stepping scheme. There is a marked difference between the coefficient $\delta c_{O_{1,2}}$ that stands close to the asymptotic value for all values of m, and the coefficient δc_{O_3} which is found to require large values of m to exhibit reasonable

convergence. This difference in behaviour can be linked to the order of magnitude of the two operators $||O_{1,2}||$ and $||O_3||$ and their effective weight in the evolution equation. One finds that $||O_{1,2}|| \approx (2\pi)^2 B$ while $||O_3|| \approx \nu J \approx (2\pi)\nu$, therefore $||O_3||/||O_{1,2}|| \approx \nu/(2\pi B)$. We now consider a change in the error of magnitude $\sigma_{\delta cO1,2}$, hence characteristic of the error on the weight $\delta c_{O_{1,2}}$ of operator $O_{1,2}$, such that this fluctuation of the error becomes projected on operator O_3 rather than operator $O_{1,2}$. The contamination of δc_{O_3} would then be of order $\sigma_{\delta cO1,2}||O_{1,2}||/||O_3||$. For such a contamination to be reasonable, one requires that $\sigma_{\delta cO1,2}$ to be small enough that:

$$\langle \delta c_{O_3} \rangle \gg \sigma_{\delta cO1,2} \frac{||O_{1,2}||}{||O_3||} \tag{31a}$$

Taking into account that $\langle \delta c_{O_3} \rangle \approx \left\langle \delta c_{O_{1,2}} \right\rangle$ one can then recast this constraint so that it only depends on the properties of $\delta c_{O_{1,2}}$.

$$\frac{||O_3||}{||O_{1,2}||} \approx \frac{\nu}{2\pi B} \gg \frac{\sigma_{\delta cO1,2}}{\left\langle \delta c_{O_{1,2}} \right\rangle} \tag{31b}$$

On Figure 21 the ratio $\sigma_{\delta cO1,2}/\left\langle \delta c_{O_{1,2}} \right\rangle$ is plotted versus m and shown to decrease exponentially as m is increased. Given $\nu/(2\pi B) \approx 0.0104$ one can then determine in threshold in m such that the criterion Eq.(31b) is marginally fulfilled, see shaded domain on Figure 21. Very high precision means square procedure with $m \geq 2^{15}$ is therefore appears to be required to avoid that the error in determining the coefficient of the operator with largest magnitude overwhelms the uncertainty in determining the coefficient of the operator with smallest amplitude.

Simulations with disparate magnitude of operators, therefore disparate magnitude of physical effects, are not only demanding in terms of numerical resolution, they also require enhanced precision with PoPe to properly evaluate the error and avoid contamination of the error estimated for the low amplitude operator by the large amplitude operator. As for the numerical implementation, a small error on the PoPe projection for the large amplitude operator drives a big error on low amplitude operator.

In this Section, we analyse the statistics of the error generated by PoPe. We examine a case where two operators O_1 and O_2 are combined into $O_{1,2} = O_1 + O_2$ so that only two operators are used in the verification, $O_{1,2}$ and O_3. The standard deviation of the error distribution is observed to decrease as the number of dimension of the vectors used for the projection is increased. With the choice of operators $O_{1,2}$ and O_3 we find a case with operators that have different magnitude so that the error on the large amplitude operator, here $O_{1,2}$ can contaminate and even dominate the error on the small amplitude operator, here O_3.

4.3 Distribution function of the error, PoPe verification

When using a least square method with fewer points than the maximum, a statistical analysis of the projection of the error on the existing operators can be performed, yielding a distribution function characterised in particular by the mean and standard deviation discussed in the previous Section 4.2. The error determined with the output of the

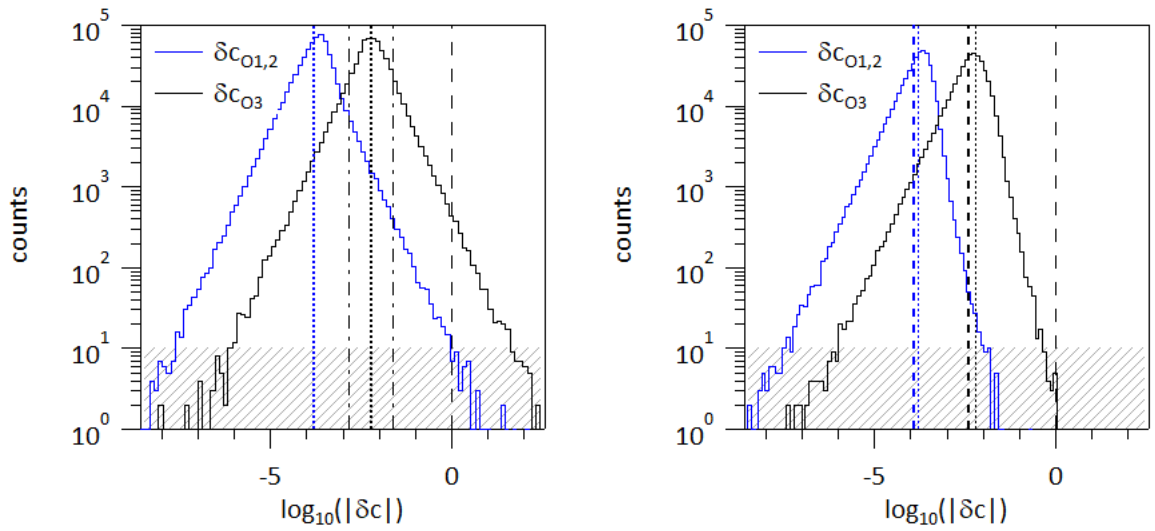

Figure 22: Histograms of the error coefficients $log_{10}(|\delta c_{O_{1,2}}|)$ in blue, and $log_{10}(|\delta c_{O_3}|)$, in black, the maximum value is indicated by the dotted vertical lines, the 100% error line, $log_{10}(|\delta c|) = 0$, by a dashed black line. Left hand side, statistics with $m = 2$. For $log_{10}(|\delta c_{O_3}|)$, the vertical dash-dot lines indicate the standard deviation with respect to the mean. The histograms are close to symmetric and an exponential decay extend towards both small and large errors. Right hand side, same analysis for $m = 3$, the histograms are not symmetric and the exponential decay mainly holds towards the small errors.

simulations of the strange attractor exhibits Gaussian like distribution with maximum probability for a given error. When considering the logarithm of the error, some form of cut-off is found towards the large errors together with an exponential fall-off towards the smaller errors, Figure 13 left hand side. A characteristic error is thus obtained together with rare events that exhibit a very small error and a maximum error slightly larger than the mean characteristic error. In the vicinity the maximum probability of the error distribution and towards the upper limit of the error, with possibly the cut-off feature, a finer structure is apparent.

We first consider the distribution for the random variables $\log_{10}(|\delta c_{O_{1,2}}|)$ and $\log_{10}(|\delta c_{O_3}|)$ with low dimension least square calculation, typically using 2 and 3 different points in phase space, Figure 22. For $m = 2$, the probability of having co-linear vectors is small but not negligible and values with large errors, typically $\log_{10}(|\delta c_{O_{1,2}}|) \geq 0$ and $\log_{10}(|\delta c_{O_3}|) \geq 0$ are found, Figure 22 left hand side. For the coefficient δc_{O_3}, black histogram, the distribution is rather symmetric with respect to its mean $\langle \log_{10}(|\delta c_{O_3}|) \rangle$, indicated by the black doted vertical line, with exponential fall-off in both directions. The mean plus or minus the standard deviation is indicated by the vertical dash-dot black lines. The histogram of $\delta c_{O_{1,2}}$, blue line, appears in first analysis to be shifted towards the smaller errors, typically by a factor 0.07, which is not too different from the magnitude ratio between O_3 and $O_{1,2}$. Towards the maximum, one can observe that the occurrence of large errors, $\log_{10}(|\delta c_{O_{1,2}}|) > \left\langle \log_{10}(|\delta c_{O_{1,2}}|) \right\rangle$ is smaller than that of small errors $\log_{10}(|\delta c_{O_{1,2}}|) < \left\langle \log_{10}(|\delta c_{O_{1,2}}|) \right\rangle$. This leads to a slight asymmetry between the left (broad) and right (narrow) hand sides with respect to the maximum.
With a 3 point least square procedure, Figure 22 right hand side, the probability of having

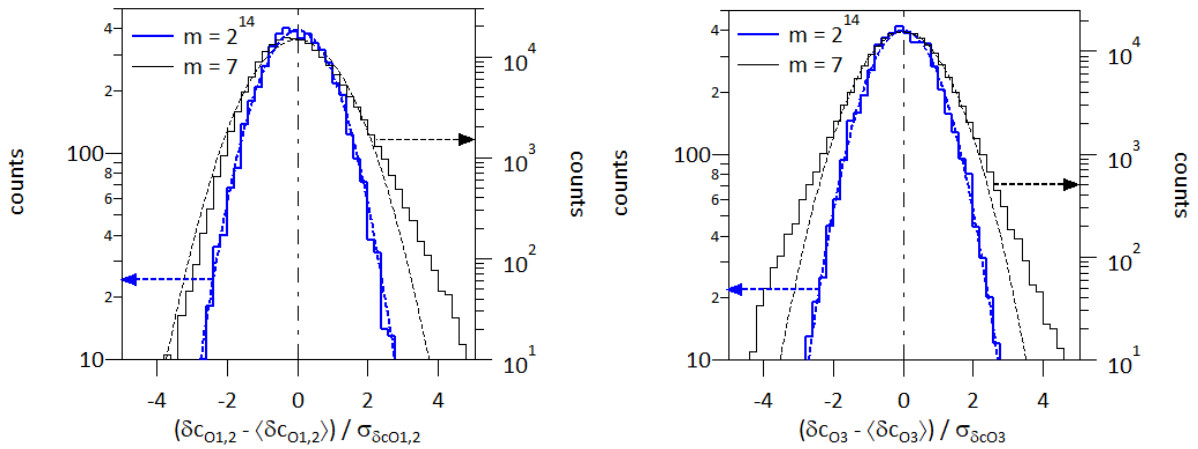

Figure 23: Histograms for $m = 2^{14}$ thick blue line, Gaussian fit dashed blue line, and for $m = 7$, black line, Gaussian fit dotted black line, versus the normalised variation $(c - \langle c \rangle)/\sigma_c$. Left hand side for $c = \delta c_{O_{1,2}}$. Right hand side for $c = \delta c_{O_3}$.

3 co-linear points out of 3 is significantly reduced compared to having 2 co-linear points out of 2. The generation of large errors in the calculation is strongly reduced. The histograms for $m = 3$ Figure 22 right hand side, are characterised by a loss of symmetry, the regions on the right hand side of the mean $\langle \log_{10}(|\delta c|) \rangle$, hence towards the large errors, being depleted. Conversely, little change of the histograms is observed towards the small errors.

The statistics of δc rather than $\log_{10}(|\delta c|)$, besides allowing one to determine the sign of δc is more sensitive to the large errors. We compare the change in the distribution functions when m is increased from $m = 7$ to $m = 2^{14}$ using the normalised variation, namely the distance to the mean divided by the standard deviation $(\delta c - \langle \delta c \rangle)/\sigma_{\delta c}$, Figure 23. This allows comparing the distribution function even when changes in the standard deviation or in the mean value are important. For both $\delta c_{O_{1,2}}$, Figure 23 left hand side, and δc_{O_3}, Figure 23 right hand side, the histogram for $m = 2^{14}$, thick blue line, is well approximated by a Gaussian distribution function, blue dashed lines. For these very large samples the standard deviation is small and the randomness in the choice of the m-tuples only yields weak variation with comparable probability. The limit, when $m = N_{max}$ is a delta function yielding a single value, the asymptotic value. For the smaller values of m, the exponential variation that governs the distribution of the error $logE = \log_{10}(|E|)$, described in Section 4.3, is indicated by a black dashed line. A Gaussian fit is shown with a black dotted line on Figure 23 right hand side. For $\delta c_{O_{1,2}}$ the distribution is skewed towards the values that are larger than the mean value, while for δc_{O_3} heavy tails are observed for both positive and negative deviations from the mean.

The statistics on $\log_{10}(|\delta c_{O_{1,2}}|)$ for $m = 2^{14}$, Figure 24 left hand side blue line histogram, yield the same Gaussian feature, blue dashed line, as that for $\delta c_{O_{1,2}}$. Compared to the $m = 2^{14}$ analysis, Figure 24 left hand side, the $m = 7$ analysis, black line histogram, exhibits a heavy tail towards the small errors, Figure 24 left and right hand side. This heavy tail, blue line histogram on Figure 24 right hand side, can be understood as the sum of the exponential dependence of $\log_{10}(|\delta c_{O_{1,2}}|) \leq \langle \log_{10}(|\delta c_{O_{1,2}}|) \rangle$, dashed black line,

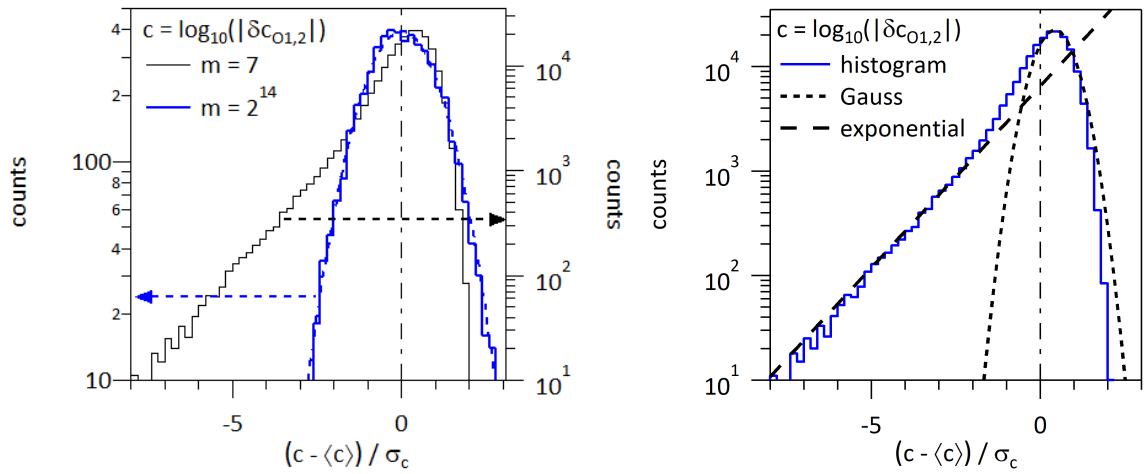

Figure 24: Left hand side: Histograms for $m = 2^{14}$ thick blue line, Gaussian fit dashed blue line, and for $m = 7$, black line, versus the normalised variation $(c - \langle c \rangle)/\sigma_c$ for $c = \log_{10}(|\delta c_{O_{1,2}}|)$. Right hand side: Histogram for $m = 7$ versus for $(c - \langle c \rangle)/\sigma_c$ for $c = \log_{10}(|\delta c_{O_{1,2}}|)$ as on left hand side, with exponential fit towards the small errors and Gaussian fit for the errors comparable to the mean value.

and of a Gaussian distribution near the mean $\langle \log_{10}(|\delta c_{O_{1,2}}|) \rangle$, dotted black line. The Gaussian fit, which is adapted to the shape near the maximum of the histogram, appears to be shifted towards the errors larger than the mean. Furthermore, the histogram in this region appears to decrease faster than the Gaussian, which is reminiscent of the cut-off behaviour discussed in the limit of the large errors. When increasing m, one can observe that the amplitude of the exponential contribution decreases, and is found negligible for $m = 2^{14}$, while the Gaussian contribution is roughly unchanged and shifted to the left and close to symmetric.

For the statistics on $\log_{10}(|\delta c_{O_3}|)$, Figure 25, one finds a different behaviour. For most of the values of m, the distribution are essentially exponential like towards the small errors. Only at the largest values of m, here $m = 2^{14}$ can one split the distribution into a sum of exponential and Gaussian distribution functions.

The projection of the error on the operators of the system to be solved exhibit a departure from 1, the target value for perfect accuracy. Two random variables are used to analyse this effect, first δc the departure from 1 of coefficient c, a direct measure of the error, second $\log_{10}(|\delta c|)$. The distribution of the error of δc is close to Gaussian, nearly symmetric with respect to the mean value. At low number of points in the least square method, outliers with large error generate heavy tails. As the number of points is increased, the distribution gets closer to a Gaussian. The width of the Gaussian narrows and the heavy tails shrink as the occurrence of outliers is reduced. Ultimately, when all available points are used a δ distribution is obtained.
Apart from the sign of the error and the characteristic value of the error, this distribution is also useful to analyse the outliers that are generated when the operators are transiently co-linear and the error exhibits a finite amplitude. Cases where the co-linear events are frequent indicate that the chosen operators exhibit a too strong correlation, which is an

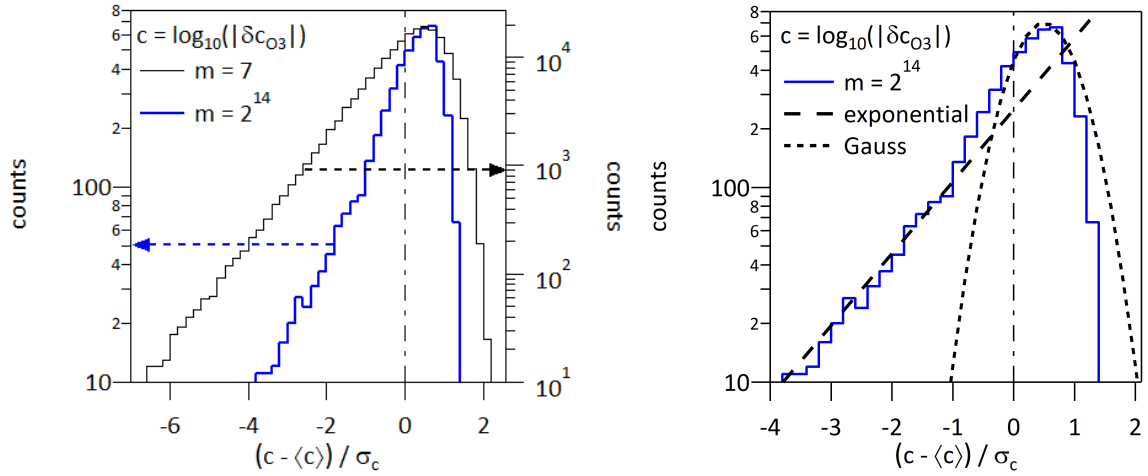

Figure 25: Left hand side: Histograms for $m = 2^{14}$ thick blue line, Gaussian fit dashed blue line, and for $m = 7$, black line, versus the normalised variation $(c - \langle c\rangle)/\sigma_c$ for $c = \log_{10}(|\delta c_{O_3}|)$. Right hand side: Histogram for $m = 7$ versus $(c-\langle c\rangle)/\sigma_c$ for $c = \log_{10}(|\delta c_{O_3}|)$ as on left hand side with exponential fit towards the small errors and Gaussian fit for the errors comparable to the mean value.

important information, and that a more appropriate choice of the operators should be considered.
The distribution of the random variable $\log_{10}(|\delta c|)$ provides a different information. In the problems with relatively few points for the least square projection, one can observe a distribution combining an exponential behaviour towards the error with small magnitude and a Gaussian feature towards the large magnitude. As one increases the number of points in the least square projection the Gaussian feature tends to become dominant, thus retrieving the behaviour observed for the distribution of δc. One thus finds that the error is characterised by a typical value with a randomly distributed departure from the mean value leading to a Gaussian distribution feature.

In this Section, we analyse the statistics that can be addressed with the PoPe verification. When using a few points to define each element of the projection, one can generate very large data base that tend to exhibit heavy tails in the distribution with the occurrence of large errors generated by spurious apparent co-linearity of the operators. As the number of points defining each element of the projection is increased, the statistics of the error δc for each control parameter tend to Gaussian with nearly constant mean error and narrower and narrower width. Ultimately, when all available points are used to define each element, a single value of the error for each retained control parameter is obtained. A key aspect when using the statistics of $\log_1 0(\delta c)$ and a relatively large number of points for each element of the projection, is that the error exhibits a Gaussian distribution around the most probable error and together with a cut-off towards the large errors.

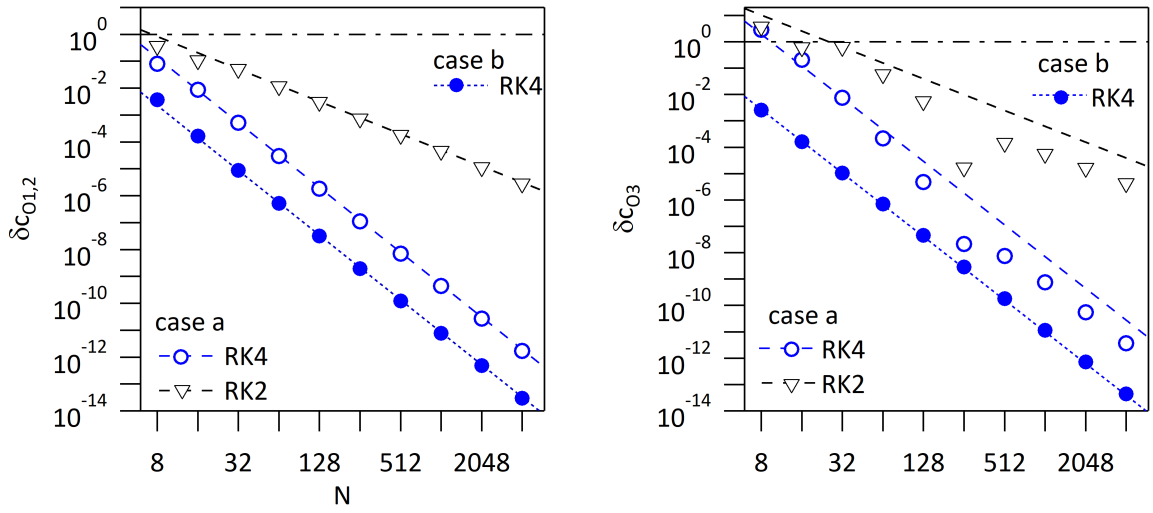

Figure 26: Dependence of the PoPe projection of the error for RK4 and RK2 integration schemes, RK4 with blue circles and RK2 with black head-down triangles, on the number of steps per unit time N. Left hand side for $\delta c_{O_{1,2}}$. Right hand side for δc_{O_3}. The dashed lines indicate the scaling laws of the error proportional to N^{-4} for RK4 and to N^{-2} for RK2.

4.4 Scaling law of the error on the weight of the operators

We consider here the actual PoPe procedure to verify the code runs, namely we compute the coefficients $\delta c_{O_{1,2}}$ and δc_{O_3} as well as the residue, using the maximum number of points N_{max} for the PoPe projection. When varying the time stepping of the integration scheme, as well as the order of the Runge Kutta scheme itself, one can check that the error measured by $\delta c_{O_{1,2}}$ and δc_{O_3} exhibits the expected scaling law. The results are summarised on Figure 26 and present the same trends as that previously reported for the error. The coefficient $\delta c_{O_{1,2}}$ Figure 26 left hand side is found to follow the appropriate scaling law indicated by the dashed lines for $N \geq 2^5 = 32$, respectively N^{-4} and N^{-2} for the RK4 and RK2 time stepping schemes. One finds that $\delta c_{O_{1,2}}$ is smaller than 1 for all the values of N that have been investigated. In terms of the chosen control parameters, an error $\delta c_{O_{1,2}}$ leads in fact to a relative error on the Chirikov parameter of $\frac{1}{2}\delta c_{O_{1,2}}$. If one considers that a more appropriate criterion is a 10% relative error on the Chirikov parameter, then one finds that for *case a*, $\sigma_{chir} = 7$, $\nu = 0.2$, the RK2 run with $N = 2^5$ is marginal while the RK4 run with $N = 2^3$ and the RK2 runs with $N = 2^3$ and $N = 2^4$ exceed the 10% error on the Chirikov parameter. One also recovers here that the error is case dependent since the RK4 error on $\delta c_{O_{1,2}}$ for *case b*, $\sigma_{chir} = 2.3$, $\nu = 0.8$ is smaller by more than one order of magnitude than that for *case a* at identical time stepping scheme. As can be expected from the prior analysis, the projection on the operator O_3 with absolute error δc_{O_3}, which is therefore an effective error made on the viscosity ν, exhibits larger values and consequently requires higher performance numerical schemes to achieve a comparable accuracy, Figure 26, right hand side. Using the same criterion of a maximum relative error of 10%, the RK4 runs with $N < 2^4$ and the RK2 runs with $N < 2^5$ exceed the 10% threshold and the RK2 case with $N < 2^6$ is marginal. For *case b*, $\sigma_{chir} = 2.3$, $\nu = 0.8$, one finds the expected trend for the scaling law of the error δc_{O_3} with time stepping, hence that δc_{O_3} scales like N^{-4}. Significant departure from the expected

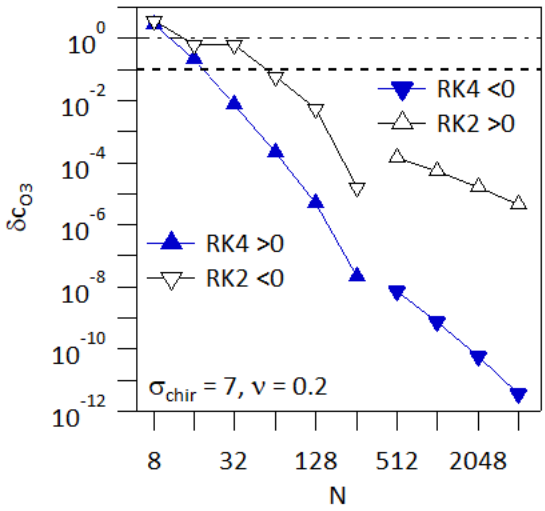

Figure 27: Sign and scaling law of the PoPe projection of the error δc_{O_3} for given integration scheme, RK4 closed blue triangles, and RK2 open black triangles, versus the number of steps per unit time N for *case a*, $\sigma_{chir} = 7$, $\nu = 0.2$. Head-up triangles positive absolute error, head-down triangles negative absolute error.

scaling law is found however for *case a*, control parameters $\sigma_{chir} = 7$, $\nu = 0.2$, Figure 27. When analysing the sign of δc_{O_3}, Figure 27, one finds that the distortion with respect to the expected scaling laws stems from a change of sign of the absolute error for both RK4 and RK2 schemes in the resolution interval $2^8 < N < 2^9$, Figure 27. The sign of the error with the RK4 and RK2 schemes is found opposite and remains opposite when the sign changes occur Figure 27. When examining the sign of $\delta c_{O_{1,2}}$ one finds that it does not depend on N for $N \geq 2^4$. For *case a*, $\sigma_{chir} = 7$, $\nu = 0.2$, the error $\delta c_{O_{1,2}}$ is negative with the RK4 scheme and positive with the RK2 scheme. Conversely, for *case b* $\sigma_{chir} = 2.3$, $\nu = 0.8$, not shown on the Figure, both errors $\delta c_{O_{1,2}}$ and δc_{O_3} are negative for the whole range of values of N that have been investigated. Within the PoPe framework, and for the particular case we investigate, the two coefficients $\delta c_{O_{1,2}}$ and δc_{O_3} are determined by the error made on the time derivatives. A change of sign of the error is then indicative that the asymptotic regime for the error behaviour has not been reached. Should the latter be chosen as accuracy constraint, then *case a* would then require a high precision time stepping with $N \geq 2^9$ for both the RK4 and RK2 schemes.

The PoPe procedure allows one to consider various projections. Up to this point, we have considered the absolute error $\delta c_{O_{1,2}}$ for the projection on the operator $O_1 + O_2$. It can be understood as driving an effective absolute error on the Chirikov control parameter, while the projection on O_3, yielding δc_{O_3}, would be the effective absolute error on the control parameter ν. However, one can also consider the absolute error $\delta c_{O1,3}$ for the projection on the operator $O_1 + O_3$ and absolute error δc_{O2} for the projection on the operator O_2. In that case the two reference operators $O_1 + O_3$ and O_2 have the same magnitude; $\delta c_{O1,3}$ and δc_{O2} should exhibit a comparable behaviour. The same analysis is made as for the previous projection, Figure 28. One finds that the two coefficients $\delta c_{O1,3}$ Figure 28 left and side, and δc_{O2} Figure 28 right and side, exhibit the same behaviour and appropriate scaling with N. A slight departure from the expected scaling law is observed

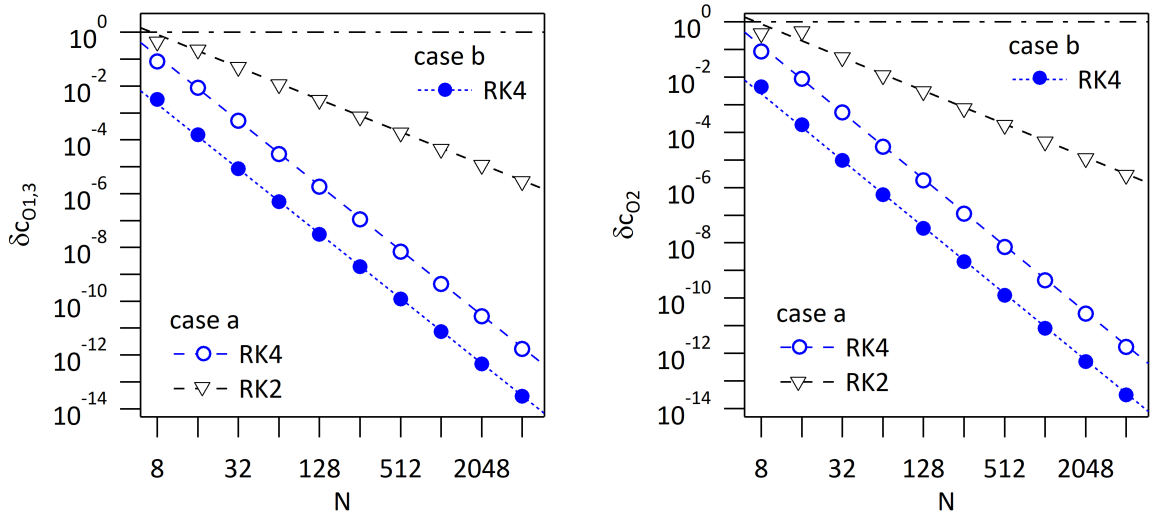

Figure 28: Dependence of the scaling law of the PoPe projection of the error for given integration scheme, RK4 with blue circles and RK2 with black head-down triangles, in terms of the number of steps per unit time N. Left hand side for $\delta c_{O_{1,3}}$. Right hand side for δc_{O_2}. The dashed line are the scaling of the error N^{-4} for RK4 and N^{-2} for RK2.

for the RK2 integration scheme at lowest values of N, $N \leq 2^5$. The latter feature is more readily seen on Figure 29 where $\delta c_{O1,3}$, δc_{O2} and $\delta c_{O_{1,2}}$ determined by the projection of the error made with the RK2 scheme are plotted together. One finds that for $N \geq 2^5$ the values of these three coefficients are comparable and exhibit the expected order 2 scaling N^{-2}. Differences are only observed for $N < 2^5$, which also corresponds to absolute errors exceeding 10%. Finally, for all values of N, $N > 2^3$, the signs of the coefficients $\delta c_{O1,3}$ and δc_{O2} are identical.

At this stage, one finds that the results of the PoPe analysis allow one to discard four simulations out of the six that have been identified as being misleading. These four simulations are in fact those that already exhibit large errors using the other verification procedures. Conversely, the two remaining simulations, both with RK2 time stepping with $N = 2^6$ or $N = 2^7$, pose a problem since, regardless of the verification method, no sharp criterion has been found that would discard them. It is therefore important to step back and revisit why in first place they have been listed as faulty. In fact, there is no measure to indicate that the simulation with RK2 and $N = 2^6$ is not correct. Indeed, both the phase space portrait of the strange attractor and the largest Lyapunov exponent agree with the highest resolution simulation. The issue is the next simulation in the series with higher resolution $N = 2^7$. Indeed, this simulation exhibits a fixed point after a chaotic transient and consequently yields a Lyapunov exponent that clearly departs from the expected range of values, see Figure 8 left hand side.
Since the PoPe projection that yields the coefficient $\delta c_{O_{1,2}}$ δc_{O_3} determines in fact the ensemble of control parameters that yield equivalent results, one must analyse within this uncertainty on the control parameters if all the simulations yield comparable results and behaviour. The sensitivity of the target solution to small variations of the control parameters is an issue. In the particular example of the strange attractor, there is a known possibility of a transition from chaotic attractor to fixed point with small variations of

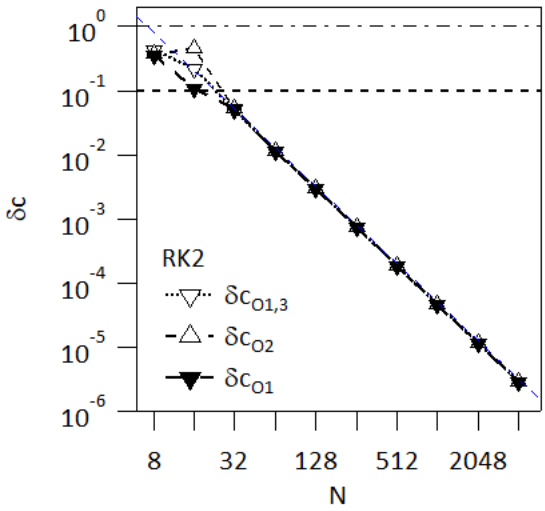

Figure 29: PoPe projection of the error for RK2, coefficients: $\delta c_{O1,3}$ head-down open triangles, δc_{O2} head-up open triangles, and δc_{O1} head-down closed triangles.

the control parameters. The PoPe verification method provides a means to address this issue. Indeed, one computes the projection of the error on the operators that govern the evolution of the problem at hand. This yields therefore the effective control parameters of the particular simulation. Furthermore, the residue, which is the part of the error transverse to the operators that govern the evolution, is a perturbation that can be regarded as some particular noise that is added to the dynamics by the numerical scheme. For the strange attractor, both the shift of the actual control parameters and the properties of the noise-like perturbation, identified as the residue, can play a role on the occurrence of fixed point solutions as well as the duration of the chaotic transients prior to the convergence to the fixed points.

In this Section, we use PoPe to investigate the accuracy and the scaling law of the error δc in terms of the number N of time steps in a period. Various combinations of the existing operators are tested that yield comparable results to that obtained with the MRS in Section 3.2. The PoPe output is of two kinds, part of the error that can be projected on the existing operators of the equations, and the residue, which has zero projection on the operators of the system. Both the control parameter error and the residue amplitude decrease when increasing the precision of the scheme. The scaling laws of the error in step size are found to agree with order of the chosen numerical scheme.

4.5 Sensitivity to small changes of the control parameters

In order to investigate the possible sensitivity of the trajectories to small changes of the control parameter, we first map the parameter space with $1 - 0.01 \leq \nu/\nu_{ref} \leq 1 + 0.01$ and $1 - 0.01 \leq \sigma_{chir}/\sigma_{chir,ref} \leq 1 + 0.01$ for *case a*, therefore $\sigma_{chir,ref} = 7$ and $\nu_{ref} = 0.02$. We use 11 values in each direction and for each pair of values of the control parameters we run the same simulations in terms of initial condition and duration. Each point of the phase portrait (ν, σ_{chir}) is characterised by the largest Lyapunov exponent Λ_+, Figures 30 and 31. The vertical and horizontal dashed lines highlight the reference val-

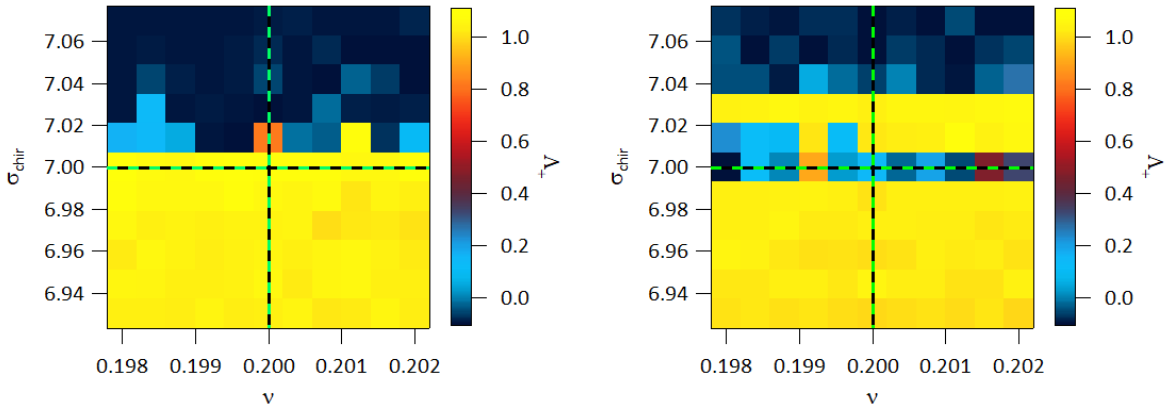

Figure 30: Control parameter space in the vicinity of ($\nu_{ref} = 0.2$, $\sigma_{chir,ref} = 7$) with $\pm 1\%$ variation and 11 points in each direction. Each simulation is characterised by the largest Lyapunov exponent Λ_+. Left hand side: Phase portrait for $N = 2^6$ steps per unit time and RK2 time stepping. Right hand side: Phase portrait for $N = 2^7$ steps per unit time and RK2 time stepping.

ues of the parameters. The values $\Lambda_+ \approx 1$ appear in yellow for the chosen colour scale while those for the fixed points appear in dark blue for $\Lambda_+ \leq 0$. The occurrence of long transients before converging towards the fixed points yields intermediate value typically with $\Lambda_+ \leq 0.5$. The phase portrait is generated using the RK2 time stepping scheme for $N = 2^6$ Figure 30 left hand side, $N = 2^7$ Figure 30 right hand side, and $N = 2^8$ Figure 31 left hand side, and using the RK4 scheme with $N = 2^7$ Figure 31 right hand side. For the case with $N = 2^6$ steps per unit time, Figure 30 left hand side, the phase portrait exhibits two *phases*, a chaotic phase $\Lambda_+ \approx 1$ for $\sigma_{chir} \leq 7.007$, and fixed point $\Lambda_+ \lesssim 0.5$ for $\sigma_{chir} > 7.007$. This phase transition is observed for all computed values of ν but for $\nu \approx 0.02012$ where the chaotic region $\Lambda_+ \approx 1$ extends up to 7.021. To be rigorous in this description of the phase portrait, one must understand by chaotic, the trajectories that exhibit chaotic transients that are longer than the chosen duration of the simulation. Indeed, one cannot exclude that at later times the trajectory might converge towards a fixed point. Conversely, in the region with fixed point, the calculation of the Lyapunov exponent includes the chaotic transients. This measure can converge towards negative values indicative of fixed points as well as small positive Lyapunov exponent. These can correspond either to a low dimensionality attractor or to a long transient before a fixed point with asymptotic value $\Lambda_+ < 0$.

The phase portrait for $N = 2^6$ is characterised by a phase transition, from chaotic to fixed point, in the vicinity of the reference values of the control parameters. The latter is found to belong to the chaotic region of the phase portrait. The relative distance along σ_{chir} of the reference simulation to the fixed-point / chaotic-attractor of order 10^{-3}. However, this is an upper bound estimate constrained by the chosen meshing along σ_{chir}. This maximum value would correspond to an error on $\delta_{O_{1,2}} \approx 5.\ 10^{-2}$. The latter is comparable to the PoPe estimated error made on $\delta_{O_{1,2}}$ for $N = 2^5$, and therefore larger that made for $N = 2^6$, typically of order 10^{-2}. It appears possible that the resolution with $N = 2^6$ is sufficient to assess that the reference point is at a distance larger than the numerical error from the phase transition chaotic-attractor / fixed point.

We now consider the phase portrait with higher resolution, $N = 2^7$ steps per unit time,

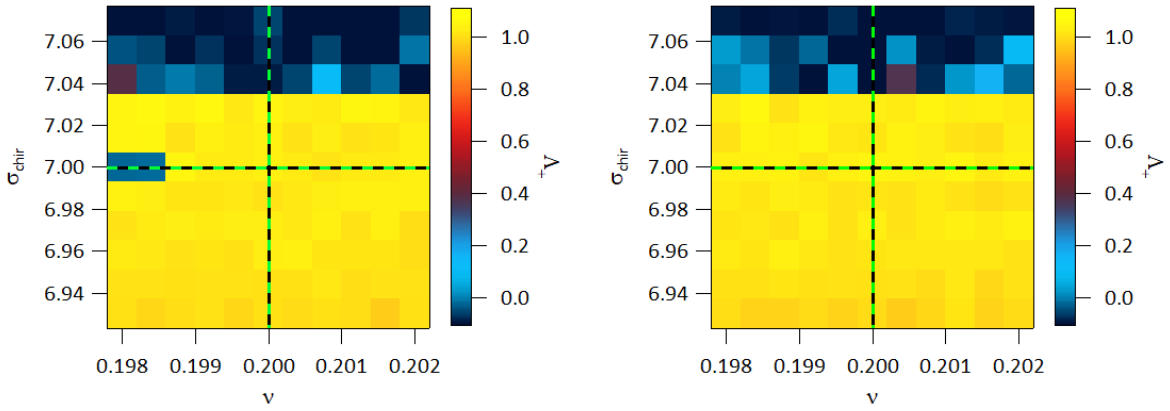

Figure 31: Control parameter space in the vicinity of ($\nu_{ref} = 0.2$, $\sigma_{chir,ref} = 7$) with $\pm 1\%$ variation and 11 points in each direction. Each simulation is characterised by the largest Lyapunov exponent Λ_+. Left hand side: Phase portrait for $N = 2^8$ steps per unit time and RK2 time stepping. Right hand side: For comparison, change of Runge Kutta scheme to RK4 time stepping, phase portrait for $N = 2^7$ steps per unit time.

figure 30 right hand side. This phase portrait appears to be more complex. One still recognises the chaotic phase for $\sigma_{chir} \leq 6.993$ and the fixed point region for $\sigma_{chir} \geq 7.035$, but the intermediate region exhibits two stripes, one with fixed points for $\sigma_{chir} \approx 7.$ and above a chaotic stripe for $\sigma_{chir} \approx 7.028$. The width of these stripes appears to vary slightly with ν. It is to be noted that the meshing of the phase portrait is a bit coarse with respect to these variations since the width of these stripes in some parts is equal to one. However, very clearly for this value of the resolution, the reference control parameters lies in the stripes of fixed points. Further increasing the resolution to $N = 2^8$, Figure 31 left hand side, one finds new changes in the phase portrait with typically a chaotic region for $\sigma_{chir} \leq 7.035$ and fixed points for $\sigma_{chir} \geq 7.035$, see figure 31 left hand side. For the smallest values of ν and $\sigma_{chir} \approx 7$, one finds a region of fixed point within the chaotic region of the phase portrait. The description of the phase portrait with RK4 time stepping and resolution $N = 2^7$ points per unit time, Figure 31 right hand side is quite similar to that obtained with RK2 time stepping and $N = 2^8$. Because of the chosen meshing of the phase portrait, it is to be underlined that horizontal stripes of changed properties with a width smaller than $\delta\sigma_{chir} = 0.017$ can escape detection. The description given to the phase portrait has to be understood with this uncertainty. The similarity between the two phase portraits of Figure 31 does not mean that the integration scheme has enough accuracy that the phase portrait are identical. It means that within the precision used to describe the phase portrait, the two sets of simulations exhibit comparable properties up to the resolution of their mesh. Phase portrait structures finer than the mesh step are unresolved. This aspect of the problem is illustrated on Figure 32. On Figure 32 left hand side, the phase portrait is very similar to that of Figure 30 right hand side, namely the case with $N = 2^7$ and RK2 time stepping. We have used different initial conditions for this set of simulations. The phase portrait properties appear to be near identical, although the transients towards the fixed points are different, leading to changes in the values of the Lyapunov exponent in the range of values $0 \leq \Lambda_+ \leq 0.5$. The fixed point stripe in the vicinity of $\sigma_{chir} = 7(1 \pm 0.001)$ exhibits the same change in width with ν as indicated previously, with an apparent width of $\delta\sigma_{chir} \approx 2 \times 7 \times 0.002$ for $\nu \lesssim 0.199$.

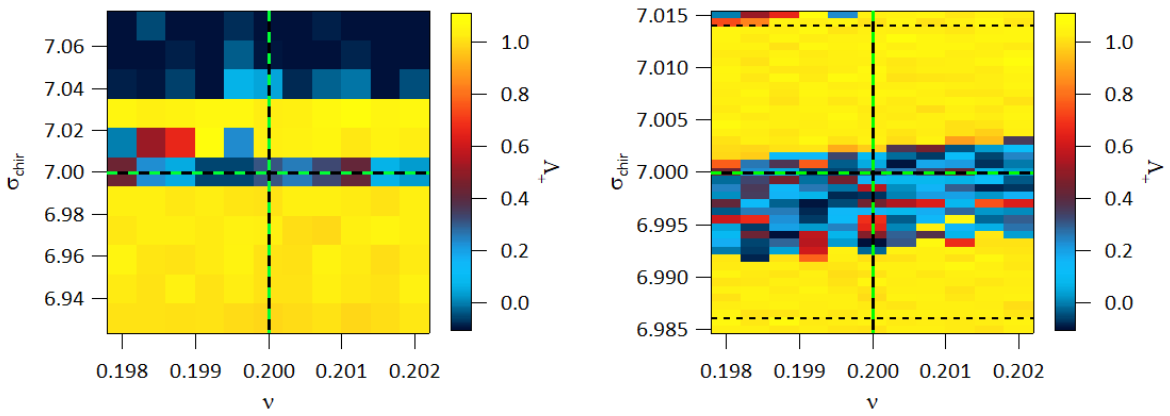

Figure 32: Control parameter space in the vicinity of ($\nu_{ref} = 0.2$, $\sigma_{chir,ref} = 7$) with $\pm 1\%$ variation and 11 points in each direction. Each simulation is characterised by the largest Lyapunov exponent Λ_+. Left hand side: Phase portrait for $N = 2^7$ steps per unit time and RK2 time stepping and different initial conditions. Right hand side: Zoom of the phase portrait for $N = 2^7$ steps per unit time and RK2 time stepping.

However, with finer meshing of the phase portrait, but reduced range of values for the Chirikov parameter σ_{chir}, and unchanged meshing for the viscosity ν, Figure 32 right hand side, one finds that this stripe is now split into two stripes. A fixed point stripe for $\sigma_{chir} \lesssim 7.$ and a chaotic stripe for $\sigma_{chir} \lesssim 7.010$, prior to a new region of fixed points. The latter is apparent at largest values of σ_{chir} and smallest values of ν.

This analysis of the phase portrait therefore indicates that the chosen control parameter lies in a region where phase transitions occur between fixed-point and chaotic regions. The structure of the phase portrait is complex and exhibits inter-layered chaotic and fixed point stripes, depending on the Chirikov parameter σ_{chir}, with comparatively small dependence on ν. The width and location of these stripes, as well as the numerical uncertainty of the effective control parameters of the simulations thus contribute to making impractical the evaluation of the correctness of the simulation on the basis on the largest Lyapunov exponent. For a coarse description of the phase portrait, both RK2 simulations with $N = 2^6$ and $N = 2^7$ can be considered to be sufficiently accurate despite the fact that they have different local values of the Lyapunov exponent. This holds because the phase portrait structure are observed to be comparable although the precise location of the change of phase from fixed-point to chaotic-attractor is resolution dependent for a given mesh of the phase portrait. Should one require finer agreement on the structure of the phase portrait, one must step to higher accuracy of the numerical scheme, however knowing that the overall sensitivity of the phase portrait structure will exclude any definitive conclusion.

In this particular set of a simulation performed with control parameters that lie in a region that exhibits a strong sensitivity of the results on the precise value of the control parameter, an alternative to evaluate the accuracy is to set the precision that one targets in the description of the phase portrait. For the chosen examples the relative precision with respect to both control parameters is typically $\pm 10^{-3}$. Consistently, one should then require that the error on the control parameters evaluated by PoPe, $\delta c_{O_{1,2}}$ and δc_{O_3}, be smaller than 2. 10^{-3} and 10^{-3} respectively, Figure 33. The difference stems from the

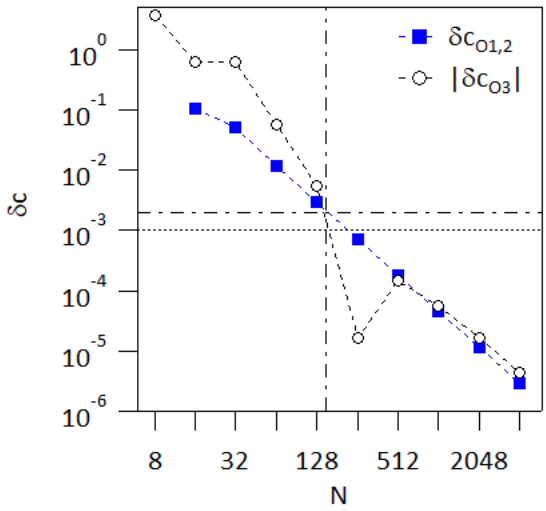

Figure 33: Variation of the relative error determined by PoPe, $\delta c_{O_{1,2}}$ for σ^2_{chir}, closed blue squares, and δc_{O_3} for ν, open black circles. The dashed, respectively dotted line indicates the maximum error for δc_{O_3}, respectively $\delta c_{O_{1,2}}$ so that the error is smaller than the mesh size of the phase portrait of Figures 30 and 31.

square root dependence of the Chirikov parameter on the control parameter B, here akin to $c_{O_{1,2}}$. When plotting the PoPe error δc versus the precision N, one finds that the accuracy increases with the N but only drops below the chosen phase portrait precision with respect to both control parameters when $N \geq 2^8$, Figure 33. The criterion of a required phase portrait precision thus leads one to discarding the two simulations $N = 2^6$ and $N = 2^7$.
The present analysis therefore indicates that the accuracy and verification of the code is case dependent, not only in terms of the chosen parameters but also in terms of the physics. Each simulation must be evaluated according to the physics that is to be addressed. For a rather loose description of the properties, hence retaining the two simulations RK2 $N = 2^6$ and $N = 2^7$, the Method of Return Solution presented in Section 3.2, performed for each simulation, and the PoPe projection of the error yield comparable criteria. However, the former has higher computing cost and requires running a different version of the code. A finer description of the properties requires enhanced numerical precision, discarding these RK2 $N = 2^6$ and $N = 2^7$ simulations.

Investigating the simulation accuracy and the criterion that allows identifying a simulation as correct thus leads us to analyse the sensitivity on the control parameters. Indeed, the PoPe projection determines the ensemble of control parameters that yield equivalent simulation output given the numerical errors. In most cases, this small uncertainty has little effect on the behaviour of the system. However, as observed with the present example, the phase portrait of the system can be quite sensitive to the values of the control parameters. We have observed bifurcations between fixed-points and chaotic-attractor for changes of the control parameter that are comparable to the effective error on the control parameters as determined by PoPe. The PoPe analysis then leads us to refine the precision to adjust the simulation result to the accuracy one chooses as target for the phase portrait description. In this discussion, another issue is of importance, namely the

role of the residue, the part of the error that is orthogonal to the operators found in the equations. The latter can be seen as a particular noise-like perturbation when following chaotic trajectories or transients. A complete description of the effective system corresponding to the simulation output is both an effective error on the control parameters and an effective noise like perturbation added to the system and accounting for the residue.

4.6 iPoPe error analysis

The first step in the PoPe analysis, both PoPe in the previous Sections and iPoPe addressed here, is to determine the error E, namely the difference between a reconstructed operator from the simulation output and the value obtained with the reference equation and the elementary reconstructed operator contributing to that equation, see Section 4.3. The error is projected on the latter operators yielding the corrections δc to the weight of these operators. The residue R is then defined as the part of the error orthogonal to all the elementary operators, see Eq.(14b). As an example of this procedure we consider the simulation of *case a* with control parameters $\sigma_{chir} = 7$, $\nu = 0.2$ and RK2 time stepping with 2^9 steps per unit time. The probability distribution function (PDF) of the error and of the residue is shown on Figure 34 left hand side. These two distribution function exhibit similar shapes, they are close to symmetric, peaked in the vicinity of zero, and exhibit a cut-off, with no data for $|E| \gtrsim 0.247$ and $|R| \gtrsim 0.185$. One can thus observe that these cut-off values are consistent with the expectation that the residue $|R|$, the remnant error transverse to the implemented operators as obtained with PoPe, is smaller than the error $|E|$. The number of counts is also observed to decrease faster for the residue than for the error as the amplitudes of $|E|$ and $|R|$ are increased, Figure 34 left hand side. The distribution function for R is therefore found to be narrower than that of E. Note that the number of counts cannot be directly compared since the bin size of the histograms are chosen to be proportional to the standard deviation of the data, and consequently different for E and R. The decrease of R with respect to E, found statistically, is recovered when considering the distribution function of $\log_{10}(|E|)$ and $\log_{10}(|R|)$, Figure 34 right hand side. As discussed previously one can observe that the histograms exhibit exponential like features with the same cut-off behaviour for both E and R. One also finds that the most probable value is shifted towards the smaller values of $\log_{10}(|R|)$, $\log_{10}(|R|) \approx -2.72$ for the most probable value hence $|R| \approx 1.9\ 10^{-3}$, compared to $\log_{10}(|E|)$, $\log_{10}(|E|) \approx -2.28$ for the most probable value hence $|E| \approx 5.4\ 10^{-3}$. The reduction factor is therefore typically of ≈ 2.8. One thus finds that part of the error is projected on the existing operators of the driving equation, and that the remnant error, the residue R, has been reduced when compared to the original error E.

The iPoPe procedure is applied to the error analysis of the strange attractor now considering the three independent operators $O_1 = -(2\pi)B\sin(2\pi x)$, $O_2 = -(2\pi)B\sin(2\pi(x-t))$ and $O_3 = -\nu J$, see Eq.(18b). The coefficients δc_{O_1}, δc_{O_2} and δc_{O_3} are then determined by 5 different ways using iPoPe, depending on the order followed in this staged approach.

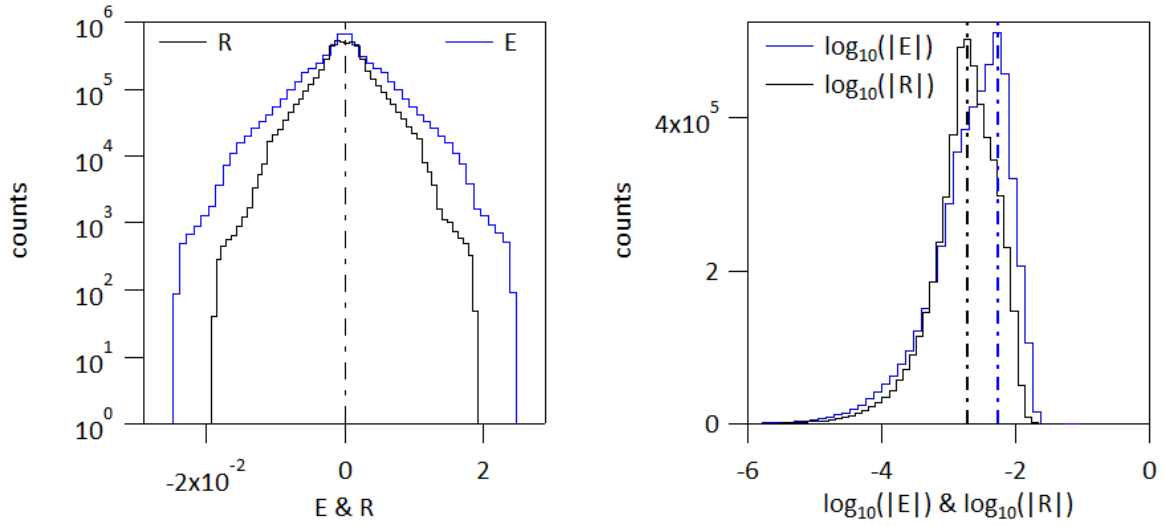

Figure 34: Left hand side: Histograms of the error E, blue line, and residue R, black line. Right hand side: Histograms of $logE = \log_{10}(|E|)$, blue line, and $\mathcal{R} = \log_{10}(|R|)$, black line. The vertical dashed-dot lines for the latter indicate the most probable values. Simulation with RK2 scheme time step $1/N$, $N = 2^9$.

For example, when computing δc_{O_1}, one finds the various results:

$$\delta c_{O_1}^{(1,2,3)} = \frac{\langle E|O_1\rangle}{\langle O_1|O_1\rangle} = \delta c_{O_1}^{(1,3,2)}$$

$$\delta c_{O_1}^{(2,1,3)} = \frac{\langle R_2|O_1\rangle}{\langle O_1|O_1\rangle} \quad ; \quad \delta c_{O_1}^{(3,1,2)} = \frac{\langle R_3|O_1\rangle}{\langle O_1|O_1\rangle}$$

$$\delta c_{O_1}^{(2,3,1)} = \frac{\langle R_{2,3}|O_1\rangle}{\langle O_1|O_1\rangle} \quad ; \quad \delta c_{O_1}^{(3,2,1)} = \frac{\langle R_{3,2}|O_1\rangle}{\langle O_1|O_1\rangle}$$

Here the three superscript labels of the coefficients indicate the order of the iPoPe projection starting from the label on the left. In the first step, for instance starting with the projection on O_1, the value of the error $\delta c_{O_1}^{(1,2,3)} = \delta c_{O_1}^{(1,3,2)}$ because this initial step does not depend on the subsequent projection. These expressions depend on the staged values of the residues, which are defined by:

$$R_2 = E - \delta c_{O_2}^{(2,1,3)} O_2 \quad ; \quad R_{2,3} = E - \delta c_{O_2}^{(2,3,1)} O_2 - \delta c_{O_3}^{(2,3,1)} O_3$$

$$R_3 = E - \delta c_{O_3}^{(3,1,2)} O_3 \quad ; \quad R_{3,2} = E - \delta c_{O_3}^{(3,2,1)} O_3 - \delta c_{O_2}^{(3,2,1)} O_2$$

The five different series of values of the iPoPe coefficients δc_{O_i} obtained for each simulation can then be investigated and compared to the PoPe result. One can note that for a routine use in production runs, a single series out of the five is sufficient to characterise the accuracy, with the benefit of avoiding the matrix inversion required with the standard PoPe. The five different series are only useful for a more complete analysis of the numerical scheme, in particular to understand how the error made on one operator propagates on the other operators, with particular attention to those that play a particular role in the symmetries or bifurcation features.

We consider here simulations of *case a* with the RK2 time stepping scheme and use the scaling law of the error with number of steps per unit time N to compare the different

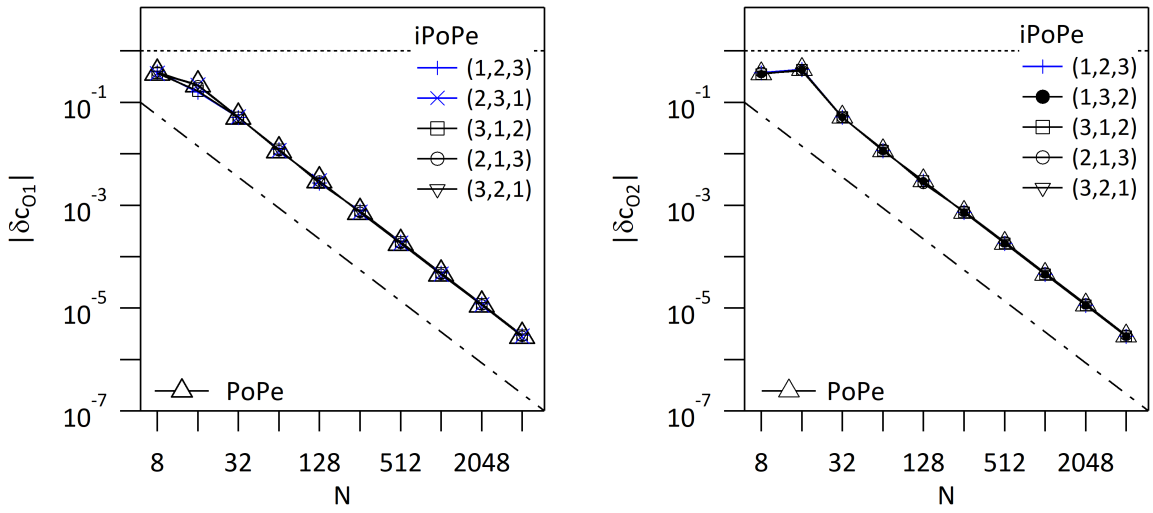

Figure 35: Determination of the coefficients δc_{O_1}, left hand side, and δc_{O_2}, right hand side with both PoPe, black open head-up triangles, and iPoPe. The five different ways of computing the coefficients are labelled according to the order used for the iPoPe staged projection procedure. Each iPoPe result is characterised by the triplet indicating which coefficient is determined first, second and third: $(1,2,3)$ blue plus + marker, $(2,3,1)$ blue cross $\times$ marker, $(1,3,2)$ black closed circle, $(3,1,2)$ black open square, $(2,1,3)$ black open circle, $(3,2,1)$ black open head down triangle. The dash-dot line indicates the order 2 scaling of the error and the dotted line indicates the relative error equal to 1. Data of simulations of *case a* with RK2 time stepping.

iPoPe series of results. For both coefficients δc_{O_1}, Figure 35 left hand side and δc_{O_2}, Figure 35 right hand side, the iPoPe set of values and the PoPe value are comparable. For $N \geq 2^5$ one also finds that $\delta c_{O_1} \approx \delta c_{O_2}$. Conversely, differences are observed for the coefficient δc_{O_3}, Figure 36 left hand side. One can notice that three out of the five possible iPoPe series are similar to the PoPe result; those corresponding to the sequences $(1,2,3)$, $(2,1,3)$ and $(2,3,1)$. The sequences $(3,1,2)$ and $(1,3,2)$ are comparable to the other iPoPe and PoPe results for $N \leq 2^7 = 128$. As shown previously, the drop of the PoPe result for $N = 2^8$ is governed by a change of sign of δc_{O_3} that occurs for $2^8 < N < 2^9$. The sequences $(1,2,3)$, $(2,1,3)$ and $(2,3,1)$ exhibit the sign change for $2^7 < N < 2^9$, while the sequences $(3,1,2)$ and $(1,3,2)$ are characterised by a sign change between $N = 2^9$ and $N = 2^{10}$. For the coefficient δc_{O_3}, with RK2 time stepping, one finds that the iPoPe depends of the order in which the staged projections are performed. Furthermore, the scaling law of the error when changing the time step is less precisely observed compared to the result for the other two coefficients. The analysis of the error on Figure 13 also indicates that the error is large, of order 10^{-1}, for $N < 2^6$. It becomes comparable to that of the other coefficients for $N \geq 2^9$.

The time trace of the various operators LHS, O_1, O_2, O_3 and R normalised by the mean value of RHS plus its standard deviation, are plotted on Figure 36 right hand side. One finds that O_1 O_2 and LHS have comparable magnitude, typically two orders of magnitude larger than that of O_3 and four orders of magnitude larger that the residue R. The small relative magnitude of O_3 compared to that of the other operators is an issue for the

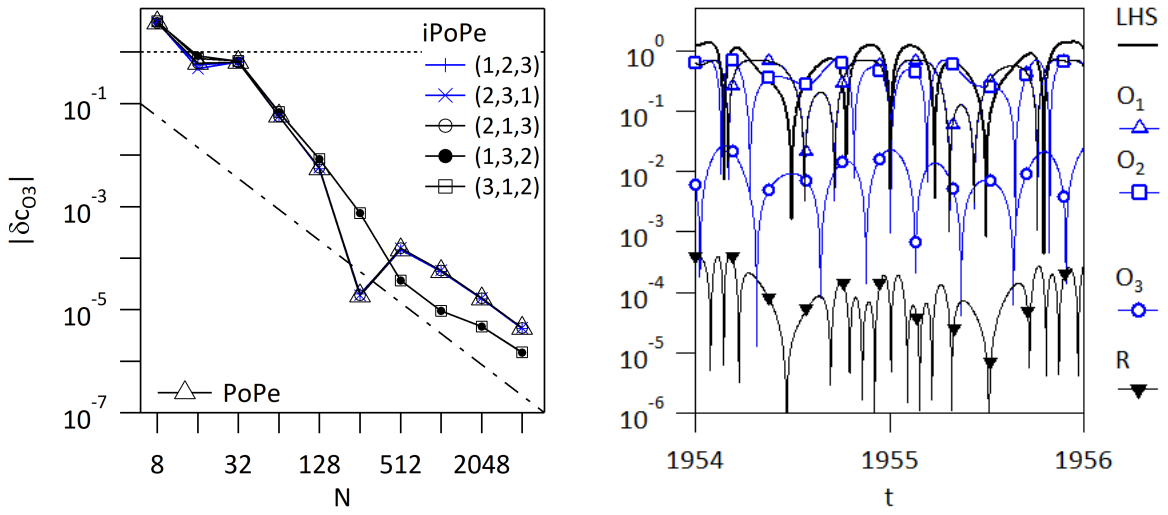

Figure 36: Left hand side: Determination of the coefficients δc_{O_3} with both PoPe, black open head-up triangles, and iPoPe: $(1,2,3)$ blue plus + marker, $(2,3,1)$ blue cross $\times$ marker, $(1,3,2)$ black closed circle, $(3,1,2)$ black open square, $(2,1,3)$ black open circle, $(3,2,1)$ black open head down triangle. The dotted line indicates the relative error equal to 1. Right hand side: Time trace over 2 periods of the Left Hand Side (LHS) of the evolution equation plain black line, and of the three operators that contribute to the right hand side of Eq.(22), O_1 open blue head-up triangles, O_2 open blue squares, O_3 open blue circles and the residue R, closed black head-down triangle, RK2 simulation of *case a*, time step $1/N$, $N = 2^9$. All operators are normalised by the mean plus standard deviation of RHS.

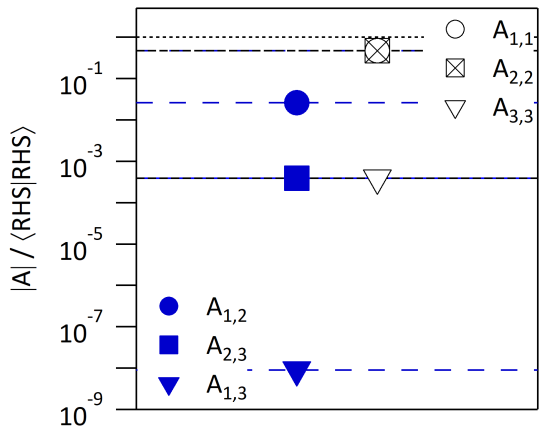

Figure 37: Values of the symmetric matrix $\mathcal{A}/\langle RHS|RHS\rangle$, see Table 1, open symbols diagonal elements, closed symbols off-diagonal elements, RK2 simulation of *case a* with $N = 2^9$. The dotted line indicates the value of the sum of all the matrix elements, equal to one for the normalisation by $\langle RHS|RHS\rangle$.

Table 1: Elements of matrix $\mathcal{A}$ normalised by $\langle RHS|RHS\rangle$ where RHS is the Right Hand Side of Eq.(22) equal to the second time derivative of x, see also Figure 37. RK2 simulation of *case a* with $N = 2^9$.

$$\mathcal{A} = \begin{pmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} = \begin{pmatrix} 4.73\ 10^{-1} & 2.60\ 10^{-2} & 8.97\ 10^{-9} \\ 2.60\ 10^{-2} & 4.75\ 10^{-1} & 3.86\ 10^{-4} \\ 8.97\ 10^{-9} & 3.86\ 10^{-4} & 3.86\ 10^{-4} \end{pmatrix}$$

numerical resolution. In the present example, this operator controls the shrinking of the phase space towards the strange attractor. This a major part of the physics that must be properly addressed. One finds that the dynamics determined by Eq.(22) are twofold: when the time t is equal to zero modulus 1, the operators O_1 and O_2 add to each other and their combined effect governs the dynamics, the phase space contraction due to O_3 has a small effect. Conversely, when $t = 0.5$ modulus 1 the two operator O_1 and O_2 are opposite and cancel out so that the evolution is only governed O_3, phase space contraction is then the leading effect. To properly account for this particular behaviour, several integration steps must take place when O_3 is the leading operator, therefore in a narrow time window of typical half width $3.\ 10^{-3}$. The time step for N equal to 2^7, 2^8, 2^9 is typically $8.\ 10^{-3}$, $4.\ 10^{-3}$, $2.\ 10^{-3}$ respectively. This indicates that $N \geq 8$ is the minimum value to have several integration steps in the time window when dissipation enforced by O_3 is the main mechanism at play. The iPoPe calculation being equivalent to the PoPe method when the matrix $\mathcal{A}$ with elements $A_{k,k'} = \langle O_k|O_{k'}\rangle$ is diagonal, the comparison of the 5 different series of values of the iPoPe coefficients δc_{O_i} with that computed directly with PoPe depends on the relative values of the diagonal and off-diagonal elements of the symmetric matrix $\mathcal{A}$. The calculation of the last coefficient with iPoPe, the third one with the present example, is the same as with PoPe. Therefore, if the two first coefficients in the present series are accurately determined the calculation of the third will also be accurate even if the non-diagonal elements $A_{3,1}$ and $A_{3,2}$ are comparable to the diagonal element $A_{3,3}$. In the chosen example of RK2 simulations of *case a*, the matrix elements can be computed, see table 1. In this table the elements are normalised by $\langle RHS|RHS\rangle$ where $RHS = O_1 + O_2 + O_3$. The sum of all matrix elements is equal to 1 by definition.
As found in table 1 and shown on Figure 37, the leading terms are the diagonal elements for the two first rows $A_{1,1} \approx A_{2,2} \gg A_{1,2} \gg A_{2,3} \gg A_{1,3}$. One could therefore expect the observed agreement between iPoPe and PoPe results for the coefficients δc_1 and δc_2, Figure 35. The calculation of δc_3 leads to different results because the magnitude of operator O_3 is small $A_{3,3} \ll A_{1,1} \approx A_{2,2}$ and because its cross product with O_2 is comparable to its magnitude $A_{3,3} \approx A_{2,3}$, see table 1 and Figure 37. This coupling governed by $A_{2,3}$ determines the contamination of the error coefficient δc_3 by any change in the calculation of δc_2. Conversely, changes in the value of δc_3 have little effect on δc_2 because $A_{2,2} \gg A_{2,3}$. This explains the increased precision achieved for the coefficient c_3 with iPoPe, smallest value of δc_3 see Figure 36 left hand side, when its calculation precedes that of δc_2.

The iPoPe calculation is an alternative to the PoPe verification, which strongly simplified calculation, and therefore very efficient to analyse the accuracy at reduced cost in computing resources. The use of iPoPe consists of a staged projection of the error on the

various operators. The result then depends on the order chosen to proceed. Comparing the various possibilities proves to be an indicator of possible correlation between the operators or that of contamination of the error between operators with large amplitude and those with small amplitude. An even simpler use of iPoPe is to project the total error on each operator, thus generating the worse possible case for the error calculation of each control parameter. The difference that is observed between PoPe and iPoPe results is small so that iPoPe can be regarded as the most efficient means to address verification and accuracy checks of production runs.

4.7 PoPe analysis with missing operator

In this Section, we analyse the effect of assuming the dependence on an operator that is not present in the equations addressed by the simulations. We only use the order 6 finite difference scheme to rebuilt the time derivative from the stored data and omit the superscript specifying the order of the reconstruction scheme. The equation that is solved numerically has been written as:

$$\left[\frac{d^2x}{dt^2}\right] = c_1O_1 + c_2O_2 + c_3O_3 + c_4O_4 + R \tag{34}$$

For the actual equation to be solved one has $c_1 = c_2 = c_3 = 1$ since the three operators govern the evolution, and determine therefore the Right Hand Side (RHS) of Eq.(34). The operator O_4 is the chosen missing operator and consequently one has $c_4 = 0$ for the theoretical equation. Similarly, the residual error R is equal to zero for the theoretical equation. In practise the equation that governs the evolution determined numerically is Eq.(34) but where $c_1 = 1 + \delta c_1$, $c_2 = 1 + \delta c_2$, $c_3 = 1 + \delta c_3$, $R \neq 0$ and possibly $c_4 \neq 0$. The error is then defined according to Eq.(24a).

$$E = \left[\frac{d^2x}{dt^2}\right] - \big(O_1 + O_2 + O_3\big) \tag{35a}$$

and therefore:

$$E = \delta c_1O_1 + \delta c_2O2 + \delta c_3O_3 + c_4O_4 + R \tag{35b}$$

where we further assume that the projection of R on O_1, O_2, O_3 and O_4 is equal to zero. One can note that when defining E, the role given to O_4 is quite different from that of the two operators. This can be regarded as a bias in the analysis. To show that this is not the case lest us define another error function E_{O3} such that:

$$E_{O3} = \left[\frac{d^2x}{dt^2}\right] - \big(O_1 + O_2\big) \tag{36a}$$

so that:

$$E_{O3} = \delta c_1O_1 + \delta c_2O_2 + c_3O_3 + c_4O_4 + R_{O3} \tag{36b}$$

In this last case we allow the residual error R_{O3} to be slightly different from R. We shall see that the PoPe analysis readily handles this difference and provides the appropriate weight for the operator O_3 with both definitions of the error. The choice of O_4 is quite arbitrary. For this example, we choose:

$$O_4 = 2\pi\ B\ \cos\big(2\pi(x - t)\big) \tag{37}$$

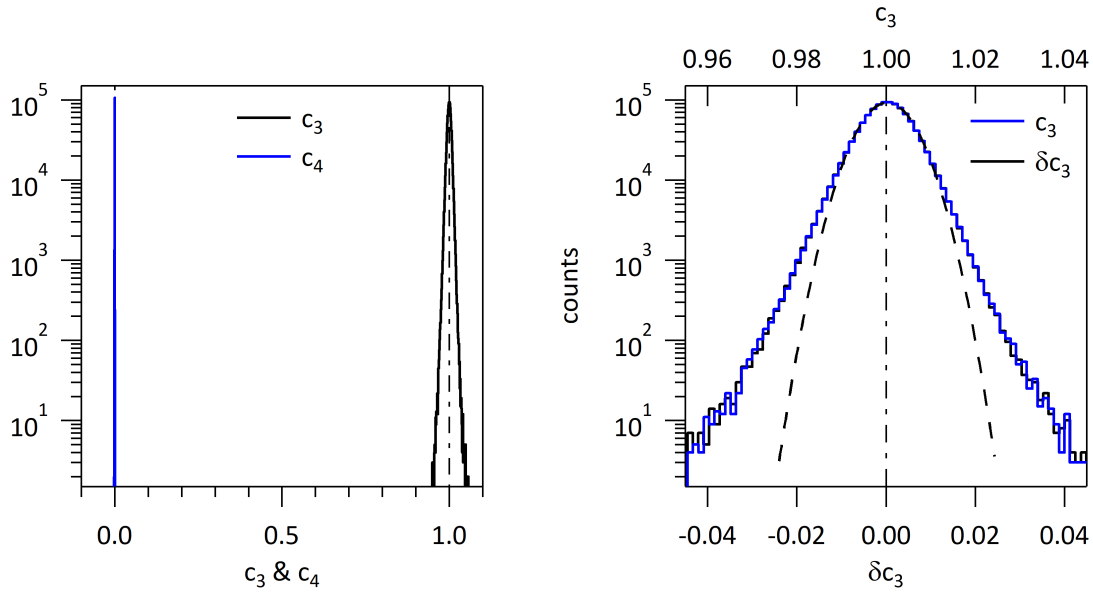

Figure 38: Histogram for *case a* simulations with RK2 time stepping and time step $1/2^9$. Left hand side: Histogram of coefficients c_3, black histogram, and c_4, blue histogram, computed with E_{O3}, Eq.(36a). Right hand side: top scale, blue histogram coefficient c_3 computed with the error E_{O3}, Eq.(36a), lower scale, black histogram, coefficient δc_3 computed with the standard expression of the error Eq.(35a).

Regarding O_4 defined in Eq.(37), we analyse the projection of the error on this operator and how this modifies the values of the other coefficients, in particular δc_3.

For the first step of this analysis we use error E_{O3} Eq.(36a) to determine c_3 and c_4, Figure 38 left hand side. One then finds that the histogram of c_3 is centred on the value 1 as it should given the evolution equation while c_4 is centred on 0 clearly indicating that the operator O_4 is not present on the right hand side of the evolution equation implemented for the simulation. One also readily notices that the width of the histogram of c_3 is much larger than that of c_4. The zero value of c_4 is recovered with better precision than the 1 value of c_3. For this calculation and the others of this Section, eight randomly chosen times of the output are used for the least square calculation of the coefficients, and a sample of 2^{20} ($\approx 10^6$) is used for the statistics. For the same simulation, we also compare the statistics of c_3 using error E_{O3} Eq.(36a) and δc_3 given by E Eq.(35a), Figure 38 right hand side. With this more precise scale, the coefficient c_3 blue histogram top scale, is clearly centred on 1. The Gaussian fit, dashed black line, yields the average $1 + 2.8\ 10^{-4}$ with standard deviation $7.6\ 10^{-4}$. For the coefficient δc_3 lower scale, black histogram, the statistics obtained with different samples, are near identical but shifted to zero, the average value is $2.8\ 10^{-4}$ about half the standard deviation $7.6\ 10^{-4}$. The theoretical relation $c_3 = 1 + \delta c_3$ is perfectly recovered here.
We find therefore that PoPe clearly discriminates the case of operator O_3, that contributes to the RHS of the evolution equation of the simulation, from operator O_4 that is not implemented. Furthermore, the calculation of c_3 and δc_3 match perfectly showing that either form of the error yield the same result.

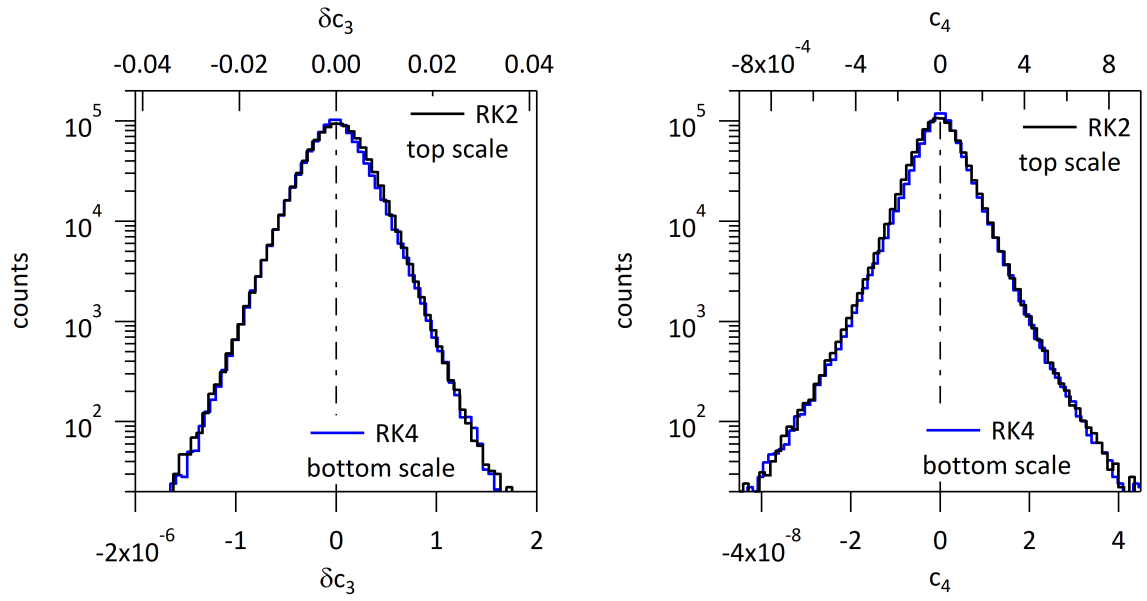

Figure 39: Histogram for *case a* simulations with RK2 compared to RK4 time stepping and time step $1/2^9$. Left hand side: Histogram of coefficients δc_3, black histogram top scale with RK2 time stepping, blue histogram bottom scale with RK4 time stepping. Right hand side: Histogram of coefficients c_4, black histogram top scale with RK2 time stepping, blue histogram bottom scale with RK4 time stepping.

Let us now compare the statistics of δc_3 and c_4 obtained with simulations with RK2 and RK4 time stepping and the same time step $1/2^9$. For δc_3 one finds rather similar statistics with the two integration schemes, typically Gaussian centred on 0 with symmetric close to exponential heavy wings. The most significant difference is a reduction by a factor $\approx 2.\ 10^4$ of the histogram width obtained with RK4 compared to RK2 simulations, Figure 39 left hand side. A similar result is obtained for c_4, same shape and same ratio of the distributions, Figure 39 right hand side. However for these statistics the symmetric, close to exponential heavy tails feature is more pronounced than the Gaussian feature near the distribution maximum. One can also notice in Table 2 that the values for c_4 are typically 50 times smaller than for δc_3 and that the mean values are typically 30 times smaller than the standard deviation except for c_4 with the RK4 scheme where this ratio increases to nearly 50. The coefficient c_4 thus appears to be closer to zero than δc_3 with smaller mean values and reduced standard deviation. As side remark, one finds that the mean value of both δc_3 and c_4 changes sign when changing the integration scheme from RK2 to RK4, which is not too surprising since the expressions of the error that can be computed are different.

The statistics of the coefficient δc_1 and $\log_{10}(|\delta c_1|)$ are shown on Figure 40 left hand side for δc_1, right hand side for $\log_{10}(|\delta c_1|)$. The data for RK2 time stepping, top scale with black PDF is compared to that of RK4 simulations, bottom scale blue PDF. For the latter a distribution departing from a Gaussian is observed for δc_1, Figure 40 left hand side. In this case, the negative values seem to exhibit features that are reminiscent of a Log-normal distribution. This appears to be superimposed to a broader distribution generating in particular a heavy tail towards the positive values. The Gaussian fit is

Table 2: PoPe coefficients δc_3 and c_4 of the strange attractor simulation of *case a* with RK2 and RK4 integration and $N = 2^9$ steps per unit time. First line mean value of the histograms Figure 39, second line "std", the standard deviation for the same data.

	δc_3 (RK2)	c_4 (RK2)	$\delta c_3(RK4)$	c_4 (RK4)
mean	$2.0\ 10^{-4}$	$-4.7\ 10^{-6}$	$-8.1\ 10^{-9}$	$1.2\ 10^{-10}$
std	$6.0\ 10^{-3}$	$1.3\ 10^{-4}$	$2.8\ 10^{-7}$	$5.8\ 10^{-9}$

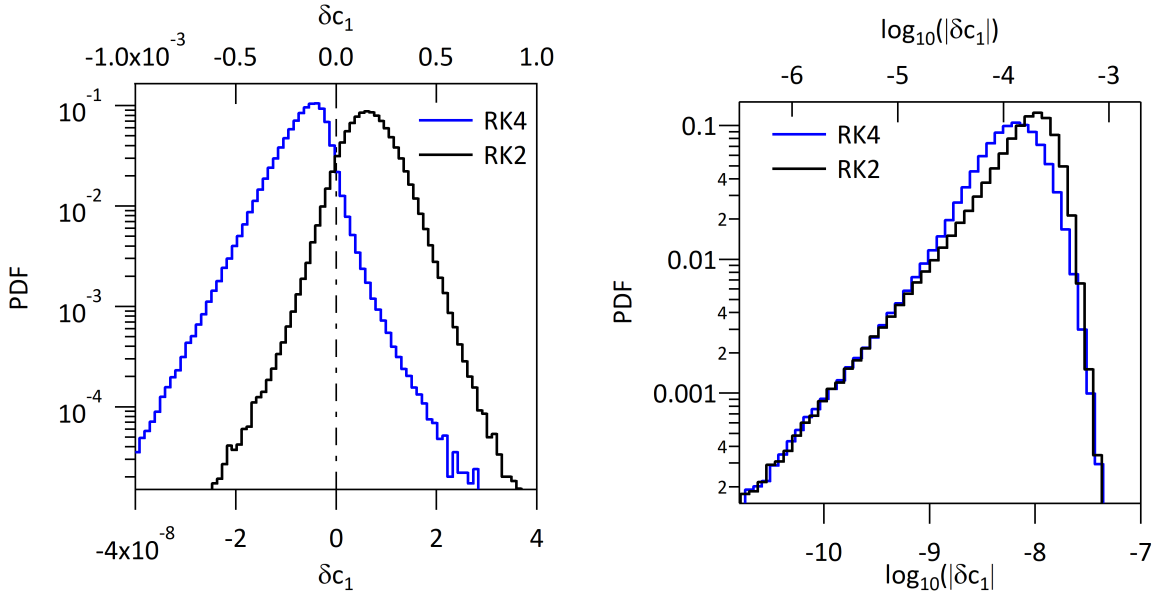

Figure 40: PDF to compare data from RK2 simulation, top scale black PDF, and data from RK4 simulation, bottom scale blue PDF. Left hand side: PDF of δc_1. Right hand side : PDF of $\log_{10}(|\delta c_1|)$. All simulations of *case a* with time step $1/2^9$ and setting $c_4 = 0$ in the PoPe analysis.

rather poor even close to the maximum, dashed line on Figure 40 left hand side. For the simulation with RK2 time stepping a closer to Gaussian symmetry is found, black PDF, Figure 41 left hand side. The statistics of $\log_{10}(|\delta c_1|)$, Figure 41 right hand side, allow one to recover these features. Towards the small errors, one finds the exponential dependence indicating that a near constant value of the PDF towards vanishing values. In the vicinity of the most probable event $\log_{10}(|\delta c_1|)$ exhibits a broader Gaussian behaviour with the RK4 data, bottom scale blue PDF, than with the RK2 data, top scale black PDF. For this data, the top and bottom scales are shifted with respect to one another by 4.3, which corresponds to a decrease of the error by typically $5.\ 10^{-5}$. It is to be noted that these PoPe results are obtained when setting $c_4 = 0$, thus ignoring the possibility of a dependence on the missing operator O_4.

The statistics of the coefficients δc_1 and δc_2, Figure 41 left hand side black histogram for δc_1, blue histogram for δc_2, are found to be remarkably similar. For the chosen RK4 time stepping, the departure from a Gaussian feature is observed for both coefficients. As for the previous results, the missing operator O_4 is ignored in this PoPe analysis. These statistics indicate that as expected the operators O_1 and O_2 play a comparable role in the structure of the error, while a different behaviour is found for O_3.

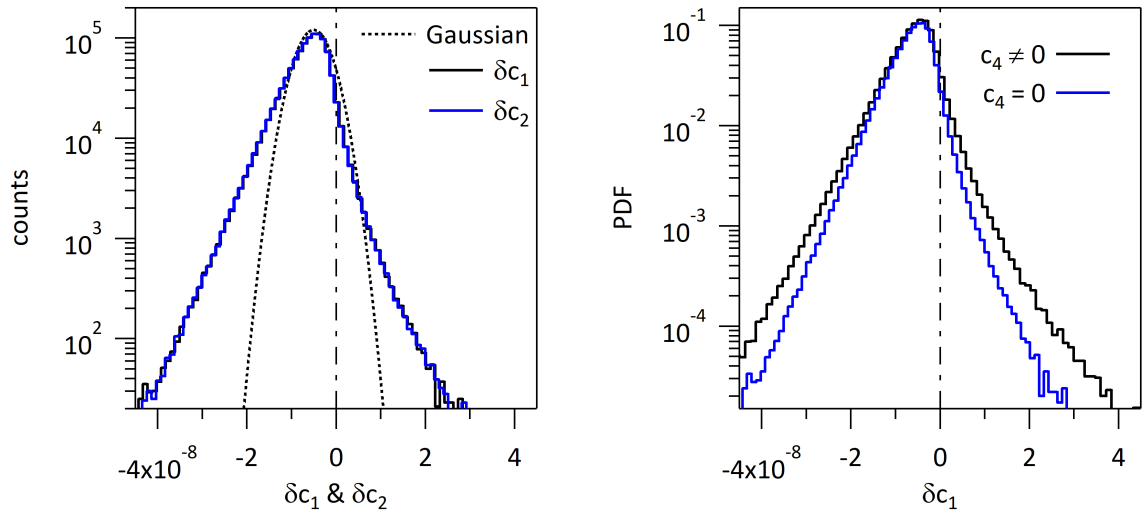

Figure 41: Left hand side: Histograms of δc_1 and δc_2 setting $c_4 = 0$ in the PoPe analysis. Right hand side: Histogram of the coefficient δc_1, blue histogram setting $c_4 = 0$ and when determining c_4 with PoPe, black histogram. All simulations of *case a* with RK4 and time step $1/2^9$.

When allowing the missing operator O_4 in the PoPe analysis, one can observe changes in the heavy tails of the PDFs of all coefficients, Figure 41 right hand side for δc_1 and Figure 42 left hand side for δc_2 and right hand side for δc_3. For all three coefficients one finds that the bulk of the PDFs are close to constant with ($c_4 = 0$ and blue PDF) or without ($c_4 \neq 0$ and black PDF) the missing operator O_4 in the PoPe analysis. Conversely the heavy tail part of the PDFs are systematically broader for $c_4 \neq 0$ compared to $c_4 = 0$.

In this Section, we have analysed the projection of the data on an operator that is not found in the equations solved numerically, operator O_4 in this example. The PoPe analysis very clearly identifies that there is no signature of this operator in the data generated by the simulation, consequently the weight of the operator is found to be close to zero, clearly different from the other operators with weight close to 1. The distribution of the error in the vicinity of these values is observed to depend on both the operators and the order of the time stepping scheme. For the present example, the error is typically Gaussian for the coefficients c_3 and c_4 while a Log-normal feature can be identified for c_1 and c_2 at high precision with RK4 integration. One also finds that trying to identify the operator O_4 tends to broaden the heavy tail part of the distribution of the error for all three coefficients c_1, c_2 and c_3, thus yielding a larger error for these coefficients and not an improved accuracy. All these results confirm that the operator O_4 is not present in the equation solved numerically while the other three operators are present as expected with the appropriate weight. This exemplifies the verification by PoPe of production simulations.

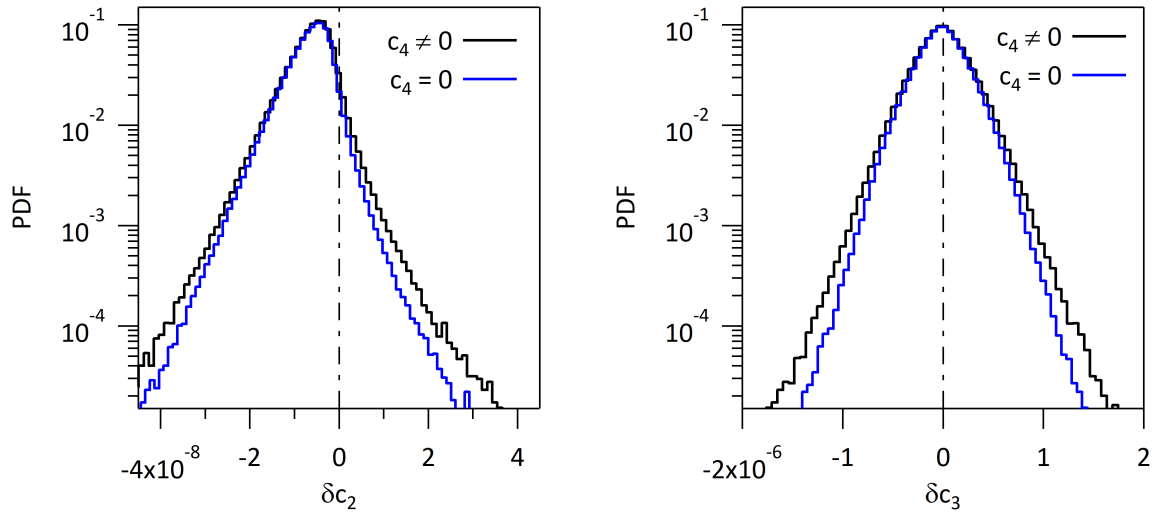

Figure 42: Compared PDF with and without operator O_4 in the PoPe analysis, blue PDF $c_4 = 0$, black PDF $c_4 \neq 0$ determined by PoPe. Left hand side: PDF of δc_2. Right hand side: PDF of δc_3. Simulations of *case a* with RK4 and time step $1/2^9$.

4.8 PoPe simulation index

In this Section, we analyse and revisit the results obtained with the PoPe verification scheme. The first step is determining an error, here the difference between the effective and expected value of d^2x/dt^2, Eq.(22). The effective value is determined using the simulation output and "recomputing" this LHS with higher accuracy than achieved during the simulation. Here an order 6 finite difference scheme yields higher precision than both the RK2 and RK4 time stepping scheme used in the simulations. The expected value of d^2x/dt^2 is recomputed using the same simulation data to determine the right hand side RHS of the evolution equation Eq.(22). The difference between LHS and RHS then defines an error E. The relative value of the error E/RHS is the first indicator of the verification procedure. For the chosen example of the strange attractor, these values are obtained for nearly all points of the computed trajectories (the end and initial points are not computed with the chosen centred finite difference scheme). The projection of the error on the operators that contribute to the right hand side RHS correlates the error E to any particular operator. The coefficient δc_k, the correlation between the error E and operator O_k, is the absolute error made for the contribution of operator O_k to RHS. When all coefficients δc_k of the chosen splitting of the right hand side RHS into a sum of operators O_k are small one can consider that the code is verified, the simulation output is consistent with the equations to be solved.
The simulation accuracy is determined by statistics on the error E, the set of the different coefficients δc, the residue R and the difference $\delta E = E - R$.

$$E = \delta c_1 O_1 + \delta c_2 O_2 + \delta c_3 O_3 + R \tag{38a}$$
$$\delta E = \delta c_1 O_1 + \delta c_2 O_2 + \delta c_3 O_3 \tag{38b}$$

For the simulation of *case a* with RK2 integration scheme and $N = 2^9$ steps per unit time, the coefficients δ_c are given in Table 3 and the statistics on E and R are illustrated on Figure 34. All three coefficients are of order 10^{-4}. These indicate that the relative value

Table 3: The three PoPe coefficients δc_1, δc_2 and δc_3 of the strange attractor simulation of *case a* with RK2 integration and $N = 2^9$ steps per unit time.

δc_1	δc_2	δc_3
$1.827\ 10^{-4}$	$1.822\ 10^{-4}$	$1.453\ 10^{-4}$

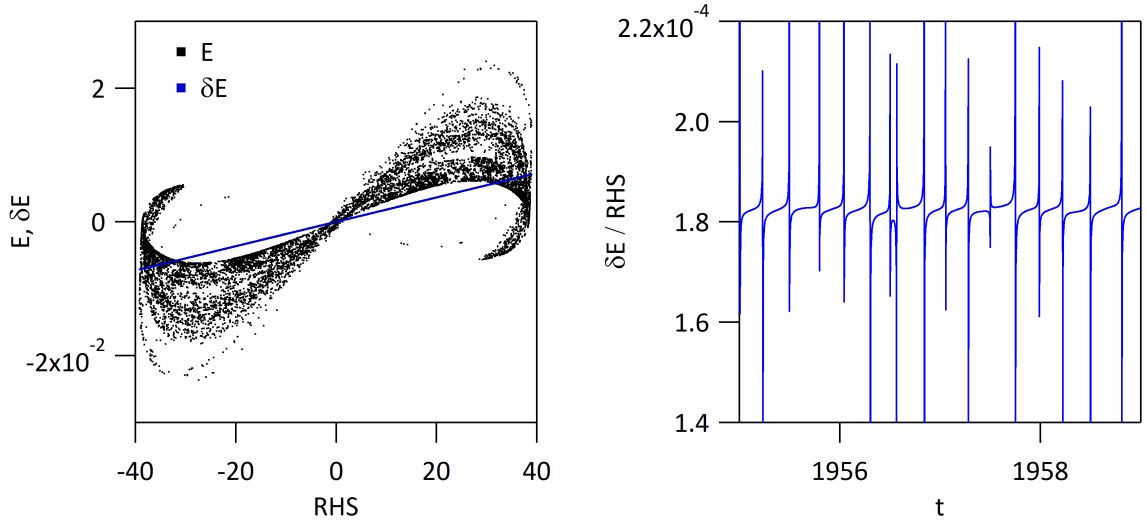

Figure 43: Left hand side: E black dots and δE blues quasi-aligned dots plotted versus RHS, data at $t = 0$ modulus 1. Right hand side: time trace of $\delta E/RHS$ for the same data showing that the ratio is nearly constant but not actually constant. Both figures RK2 simulation of *case a* with $N = 2^9$.

of the control parameters in Eq.(22) can be changed by $\delta c \approx 1.8\ 10^{-4}$ without inducing noticeable changes to the simulation output. This holds provided the phase portrait does not exhibit a bifurcation like transitions between different regimes for such a specific range of values of the control parameter, see discussion Section Sensitivity to control parameter small variation. In such a particular example, the simulation precision must therefore be adapted to the sensitivity to the control parameters one wants to address.
For the present cases, having $\delta c_1 \approx \delta c_2 \approx \delta c_3$ leads to δE Eq.(38b) close to proportional to $RHS = O_1 + O_2 + O_3$. When considering the values of the error E compared to that of the right hand side RHS taken at $t = 0$ modulus 1, Figure 43 left hand side black dots, one finds that the error E is typically proportional to RHS. There is some scatter in the proportionality factor together with a roll-over of the error towards smaller magnitude at the largest magnitude of RHS. On the same graph, the values of δE are plotted, close to aligned blue dots. These appear to be proportional to RHS as expected from the values of δc_1, δc_2 and δc_3 Table 3. In fact the time trace of the ratio $\delta E/RHS$ indicates that $\delta E/RHS$ is close to being constant $\approx 1.8\ 10^{-4}$ but exhibits a pattern which exhibits a clear departure from a constant line.

One can now consider the residue R, see Figure 44. The data for $t = 0$ modulus 1 is clearly the combination of E and δE Figure 43 so that the amplitude of R is slightly reduced and more importantly that the values of R at the largest values of $|RHS|$ are closer to being symmetric with respect to zero that when considering E. These features can be observed on the histograms Figure 34. The cut-off at the largest values does not change much as shown by the comparison of the linear distribution of E and R, Figure

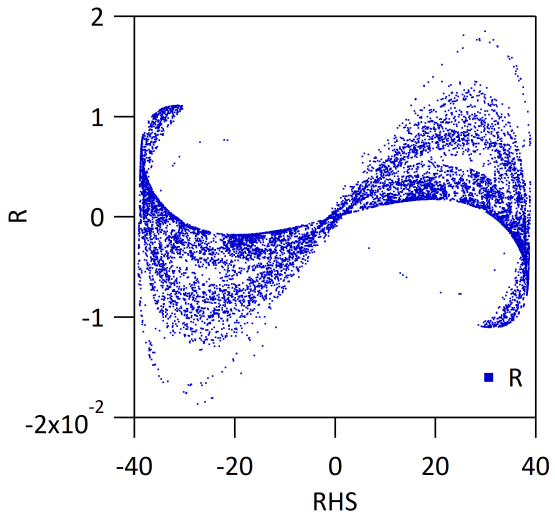

Figure 44: Value of the residue R versus RHS, data at $t = 0$ modulus 1 of the RK2 simulation of *case a* with $N = 2^9$.

34 left hand side, however the distribution of R is narrower than that of E so that one can notice a shift in the distribution of $\log_{10}(R)$ from that of $\log_{10}(E)$ indicating a factor 2.8 reduction of the value at maximum occurrence, Figure 34 right hand side. One finds therefore that about half of the error can be understood in terms of a multiplicative up-shift of order 2. 10^{-4} of the control parameter while the other half is the residue R. This residue still exhibits a structure, Figure 44. With the PoPe verification method, it is built to be orthogonal to O_1, O_2 and O_3, alternatively there is no correlation of R with either operators O_1, O_2 and O_3. In that respect R can be regarded as a low amplitude noise-like perturbation, of order 2. 10^{-4}, added to the evolution equation Eq.(22). If one investigates precise features in a system that exhibits bifurcations between different solutions, this noise-like contribution as well as the small change in the control parameters must be accounted for. In most situations, one addresses more robust properties and one can expect that this small noise-like contribution and the small error on the control parameters will have a weak effect on the simulation results.

The PoPe analysis that has been performed in this paper can be simplified by defining a figure of merit for each simulation. We first define δc_{max} the maximum of all the δc values obtained with the PoPe analysis; the worst error generates the largest value of δc_{max}. We then define the PoPe index as $-\log_{10}(\delta c_{max})$. The smallest values of the PoPe index characterise the worst error, 0 stands for a relative error of 100%, and the upper bound, a PoPe index of order 14 for an accuracy close to machine precision. To illustrate, the PoPe index, we have determined its value for two series of simulations of *case a*, with RK2 and RK4 integration scheme and number of steps per unit time ranging from $N = 2^3$ to $N = 2^{12}$. In order to compare the PoPe index for these two series of simulation we define N_{rhs} as the number of calculations of the RHS performed per reference time scale. For the RK2 scheme one then has $N_{rhs} = 2 \times N$ and for the RK4 scheme $N_{rhs} = 4 \times N$. The PoPe index for each of these 20 simulations is plotted on Figure 45. As expected the PoPe index increases twice faster for the RK4 scheme closed blue circles, than for the RK2 scheme closed black triangles, as the number of operation N_{rhs} is increased when N

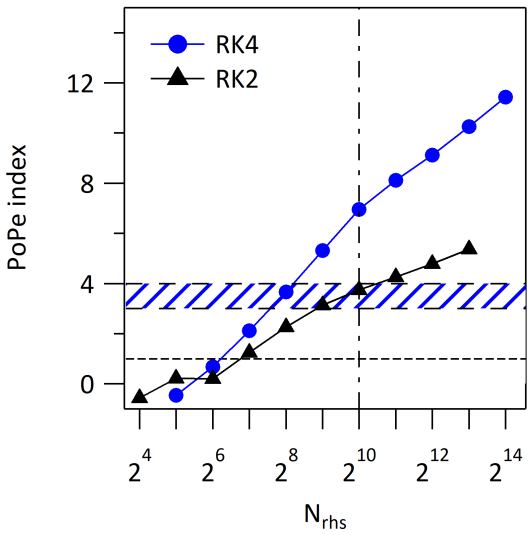

Figure 45: Value of the PoPe index versus the number N_{rhs} of operations to integrate the equations over one time unit. Data obtained with the RK2 integration scheme closed black triangles and with the RK4 integration scheme requiring twice the number of operation per step compared to RK2, closed blue circles. The vertical dash-dot line for 2^{10} operations during one unit time integration indicates the gain in precision achieved by the order 4 scheme at given computing resources. Conversely, the horizontal dashed region with PoPe index comprised between 3 and 4, indicated the gain in performance when using the high order scheme at comparable simulation accuracy.

is scanned. Let us impose the constraint of the number of operation per unit time to be $N_{rhs} = 2^{10}$ dash-dot vertical black line, hence $N = 2^9$ for the RK2 scheme and $N = 2^8$ for the RK4 scheme. One then finds that the PoPe index of the RK2 simulation is ≈ 3.7 while that of the RK4 simulation is significantly higher ≈ 7.0. Conversely, setting as target that the PoPe index should stand between 3 and 4, shaded horizontal region on Figure 45, one finds that the RK4 numerical cost is typically $N_{rhs} \approx 2^8$ for a PoPe index of 3.7, while the numerical cost with RK2 is $N_{rhs} \approx 2^{10}$ for a comparable PoPe index of 3.7. At prescribed accuracy increasing the order of the numerical scheme leads for this example to a factor 4 gain in run time.
Producing such a PoPe index for all simulation provides a figure of merit for each simulation. A PoPe index larger than 1 gives an estimate of the accuracy of the simulation, while a value close to zero or negative is most likely indicative that the accuracy and eventually the verification of the simulation could be an issue.

5 Discussion and conclusion

We have presented in this paper the PoPe and iPoPe verification methods. We have shown that these two novel verification schemes also allow addressing the simulation accuracy. Furthermore, in the course of the verification procedure specific features of the numerical scheme used for the simulation are identified as well as some key properties of the physics addressed by the simulation. PoPe and iPoPe are very similar verification methods based on Big Data analysis of the simulation output. The highlight of these methods is that

the verification process is applied directly to production simulations and not to modified numerical tools designed for the sake of verification. Furthermore, the verification can and should be applied to all production simulations, either as a post-treatment, as for the examples chosen in this paper, or on the fly during the simulation. Statistics are generated by PoPe and iPoPe. These drive the overhead in terms of computing resources when applying PoPe or iPoPe on the fly. Alternatively, necessary data for verification must be saved during the simulation for PoPe post-treatment. For a rather standard case with large statistics, the typical overhead has been estimated to be less than 10% of the simulation cost [9, 8, 10, 7], either to save extra data or to perform on the fly calculations. It can be scaled down by reducing the statistics. We have not addressed possible optimisation of such verification processes that is most likely case dependent.

The backbone of the method is to define numerically various operators, these being combined in the equations solved numerically. For a set of K operators, $m \geq K$ sets of data points can be used to determine the relative weight of these operators in the equations. For $m > K$ a least square method can be used, reducing the statistical scatter of the weight. We show that this least square procedure defines a scalar product and that increasing m reduces the weight of the occurrence of transient co-linearity between the operators. It ensures that the operators tend to become orthogonal. We have found that for $K = 2$, setting $m = 3$ with the least square method is enough to significantly reduce the effect of co-linearity and to generate numerous values of the coefficients, and consequently investigate the statistics of these errors. One also finds that taking m equal to the number of all available points is also possible, presumably yielding the best estimate of the error, but without giving insight into the statistics of the error.
The PoPe verification and accuracy analysis proceeds in three steps. In a first step the numerical error is determined. The data then gives directly insight into possible verification issues. This would occur in particular when the order of magnitude of the error is too large, or when the scaling law of the error does not match the order chosen for the numerical scheme. Conversely, we also show that when correct, the scaling law of the error (for instance with the time step) gives a first insight into the accuracy of the simulations as determined by PoPe. The second step is the projection of the error on the existing operators of the system at hand. There the PoPe and iPoPe methods depart. The PoPe method requires a matrix inversion, which can be cumbersome when the number of operators is large. It can be replaced by the iPoPe method, with a staged resolution of the linear system and possibly a dependence on the order chosen for this staged resolution. In most cases that have been analysed the difference between the PoPe and iPoPe output is small and either ways lead to comparable verification results. This projection step yields the relative error made on each coefficient of the operators that contribute to the system. One thus finds that an infinite set of control parameters, in the vicinity of that determined by the PoPe or iPoPe projection, would yield comparable simulation data. There is here a common feature with the problem on uncertainty propagation. We find with PoPe or iPoPe which uncertainty of the coefficients cannot be identified by the simulation output. Finally, the third step is to determine the residual error, transverse to the operators used in the equations.
The actual verification can be split into two different parts. The crucial one is to assess that one is actually solving numerically the equations that are claimed to be solved. However, when this part is completed arises the second part, namely the question of accuracy

of the numerical resolution. This becomes a matter of trade-off between perfect accuracy, which requires infinite computing resources, and very poor accuracy, which can be an issue for the validity of the simulations. When addressing this issue, we have found two different cases. The high accuracy simulations that are readily considered to be on the safe side and those belonging to the grey zone when the relative error is larger than 1%, of order of 10% and up to 100%. The problem is then to establish criteria that allow discriminating the safe from the unsafe simulations. We have found that some aspects of the simulation output are quite robust and recovered even when the error is large [10, 7]. We show here that specific simulations, close to bifurcation points, can be more demanding, moving the safe zone towards much higher accuracy. This underlines that the verification procedure is case dependent and each simulation will have different verification properties. Then depending on the sensitivity of the problem that is addressed, the simulation accuracy can be considered to be sufficient, thus on the safe side, or can fall short leading to possibly misleading results.

In this work, we present the interesting case of projecting the error on an operator that does not appear in the equation actually solved numerically. This corresponds to two different problems. First, the operator ought to have been part of the equations but for some reason is by-passed by the numerical scheme. Then the PoPe verification indicates that the operator has a weight 0 and not 1 as it should. The numerical scheme is not verified. Second, using a spurious operator with respect to the equations allows one to test whether this operator can describe part of the information generating the residue. This would yield some understanding of the residue and how the problem that one aims at solving can be modified by the numerical error. Rather generally, the residue being orthogonal to the operators that govern the numerical simulation, one can consider the residue to be in some ways a bit like a noise added to the system. This is the case when numerical errors are assumed to introduce a spurious diffusion so that the effective diffusion in the simulation is the sum of the controlled diffusion implemented in the equations and an uncontrolled diffusion governed by the numerical scheme. Two issues can then occur, the controlled diffusion can be dwarfed by the spurious diffusion and consequently ineffective, or bifurcation properties can be modified driving a completely different behaviour of the system. However, we have shown that for our chosen example the residue exhibits a structure that can potentially be captured, and therefore generated by an operator. Identifying approximately the form taken by the residue can be valuable for a better understanding of the problem effectively solved numerically. It also indicates means to reduce the residual error.

Another important issue arises when discussing the case of an operator that appears to have an amplitude comparable to the numerical error but that plays a crucial role, for instance in symmetry breaking. This is the case of the viscous dissipation operator in the simulations used here to exemplify the PoPe and iPoPe verification. For the dynamical system used in this paper, the viscous damping is critical since it governs the phase space contraction towards the strange attractor. However, the amplitude of this term is comparable to the error made on the larger amplitude operators. We then argue that this term will still play a leading role in the dynamics in the time windows where the amplitude of all the other operators is small. Indeed, in our example, the other operators cancel out during periodic time windows. There is a therefore a difference regarding the effect between short, medium and long time contributions to the error.

In all cases addressed in this paper, the mean value of the error is always small compared

to standard deviation. The PoPe and iPoPe analysis then allows splitting the error into a structured part that is determined by evaluating the effective values of the control parameters, and a contribution that is less structured and defines the residue. The small operator can then contribute via a long-term additive effect that can dominate over terms that exhibit short term fluctuations akin to randomness. However, one cannot exclude at this stage that the structure in the residue, which can indeed be observed, does not build-up to generate long-term effects. This question thus leads to considering a more in depth analysis of the error [9], including the slow dynamics that can occur in the simulations compared to the long-term effect of the error.

We also propose a unique index that would characterise the accuracy of the simulation. It is typically given by the opposite of the base 10 logarithm of the error. A PoPe index equal to zero indicates a 100 % error in the simulation output, and the PoPe index increases as the accuracy is improved to level off at machine precision typically between 12 and 14. Negative values are possible and most likely are a concern for the simulation. The PoPe criterion thus gives a figure of merit that allows discussing where the simulation stands with respect to accuracy that one believes to be required.

The PoPe and iPoPe verification methods have been used for several codes addressing plasma turbulence. The verification of the code TOKAM2D [9, 8], a pseudo-spectral 2D code of plasma turbulence with fluid equations in the boundary layer of magnetically confined plasmas was first performed [12, 17, 23]. The physics of the interchange instability at play is very similar to the Rayleigh-Bénard instability [18]. The TOKAM2D code verification demonstrated no error in the implemented equations as well as very high accuracy of the numerical scheme so that the diffusive coefficients in the equations can be scanned over a broad range of values, in particular towards the small values. A finite volume version of this code was written to alleviate the problem of periodicity in all directions. For this version, the PoPe verification spotted an error in the computation of the diffusive terms, leading to a correction of the code. It also showed that numerical diffusion was significant and overwhelming when the diffusion coefficients are too small, levelling-off the effective diffusion process in the simulations. This work has not been published. For a TOKAM2D simulation, the PoPe method was also used to project the relaxation phenomena on a reduced predator prey model [9]. PoPe has also been used to address codes addressing kinetic equations of plasma turbulence. The TERESA code [9, 6, 8] is a 4D kinetic code for Trapped Ion Modes [25]. The physics of the TERESA code can be viewed as that of Rayleigh-Bénard convection addressed in a kinetic framework. The TERESA code is a reduced version of the GYSELA 5D code [15, 14] using a semi-Lagrangian scheme for time stepping and cubic spline interpolation and derivatives. For the TERESA code verification [9, 8], the equations have been found to be properly implemented but numerical diffusion was observed, which breaks the symmetry of the Vlasov equation by adding a physics akin to collisional diffusion. Another kinetic code, the VOICE code addressing 1D-1V kinetic physics [4] has been verified with PoPe and iPoPe in an Eulerian version using Fourier fast transforms in the velocity direction by enforcing periodicity in that direction [10]. Very high accuracy of the plasma wave interaction is achieved but gradual numerical pollution is observed to propagate slowly from the high velocity region of the phase space towards the bulk velocities. This has led to using order 4 finite differences velocity derivatives rather than Fourier velocity space in subsequent versions of the code.

For these Eulerian versions of the VOICE code, the projection with iPoPe of the error on a diffusion operator in velocity space was found to be negligible. However, despite this problem for the medium duration runs, the physics of the short term Landau damping was found quite robust to numerical errors [10]. Finally, the verification with iPoPe of the edge plasma turbulence code TOKAM3X [26] with fluid equations exemplified the robustness of the physics even for marginal resolution [7], a further incentive for improving the resolution as achieved since, in particular in SOLEDGE3X [5], the renewed version of the code.

The PoPe and iPoPe verification methods thus provide a comprehensive verification tool that allows addressing the verification and accuracy of production runs and consequently of simulations of interest. This big data based analysis provides an in depth analysis of the simulation and numerical scheme. For the latter it will identify which operator governs the numerical error and the effective order of the resolution. This understanding can help solving some numerical issues. The analysis will also indicate which operators are small contributing to the calculation with terms that are small and that can be comparable to the error. The analysis will also be quite sensitive to operators that are close to co-linear, either requiring different definitions of the operators to be handled in the verification process or suggesting alternative ways of addressing the problem numerically. The PoPe method is quite versatile and can be used in many different ways to assess the verification of the simulation and its accuracy. Finally, this method can be used to investigate model reduction, as presented in [8] or methods to filter the simulation output to reduce contamination of the solution by the residual error.

Acknowledgements

One of the authors (PhGh) is most indebted to Paul Manneville who introduced him to the world of numerical simulation and encouraged him to investigate the transition to chaos of the compass in an alternating magnetic field. This work has been carried out thanks to the support of the A*MIDEX project (ANR-11-IDEX-0001 02, TOP project) funded by the 'Investissements d'Avenir' French Government program, managed by the French National Research Agency (ANR). This work has been also supported by the French National Research Agency grant SISTEM (ANR-19-CE46-0005-03) and by the French National Research Agency grant AIM4EP (ANR-21-CE30-0018-01). This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No. 633053 and Grant Agreement No 101052200 — EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

A Standard Method of Manufactured Solution

In this Appendix we present the Method of Manufactured Solution in Section A.1 and its application, first to the case of the strange attractor in Section A.1, then for the verification of the Runge Kutta integration schemes in Section A.2. The calculation necessary to use

the MMS for the strange attractor is presented in Section A.3.

A.1 Method of Manufactured Solution for the strange attractor

Let us consider the characteristic problem:

$$\frac{dX}{dt} = F(X) \tag{39a}$$

The standard Method of Manufactured Solution consists of selection a particular function X_0, time independent for simplicity, and modifying the initial equation Eq.(39a) so that X_0 is a steady state solution, typically:

$$\frac{dX}{dt} = F(X) - F(X_0) \tag{39b}$$

For a complex system, the operator F implemented in the code is used to determine $F(X_0)$ so that Eq.(39b) yields exactly $dX_0/dt = 0$ both theoretically and numerically. One can then perform the numerical test that X_0 is indeed a steady state solution, usually by ensuring that the slightly perturbed solution $X_0 + \delta X$ converges back to X_0. Such a procedure is elegant but has two drawbacks: first it assumes that the system of interest is such that the fixed point $X = X_0$ of Eq.(39b) is stable, second one must modify the code to solve both Eq.(39a) of interest and Eq.(39b) for the test. Third, one furthermore assumes that the chosen solution X_0 is representative of the solutions of interest. Given the evolution equation Eq.(39b) one can readily see that the eigenvalues are unchanged when stepping from the strange attractor evolution equation to the MMS evolution system. When the real part of the largest eigenvalue of the fixed point is positive, the fixed point is unstable. Furthermore, due to the explicit time dependence of the potential even a fixed point at initial time will exhibit a positive real part of the largest eigenvalue after an evolution time shorter than 0.5. One can thus expect that in most cases, disturbing the initial condition away from the fixed point X_0 in Eq.(39b) will not drive a relaxation trajectory back to the fixed point. Using the form derived in A.3, one can investigate numerically these features. The modified evolution equation for the Method of Manufactured Solution does yield trajectories of particular interest. Cases that have been tested start from a chosen fixed point $-0.5 \leq x \leq 0.5$ and $J_0 = 0$ since it is shown in A.3 that all possible values of J_0 can be investigated using a change of variable and $J_0 = 0$. An initial distance from the fixed point is chosen $d_0 = 10^{-8}$. The evolution appears to lead to large values of J, either negative or positive and consequently rapid rotation of the phase x, Figure 46 left hand side. As expected and discussed above, in all cases that have been investigated, the trajectories depart from the fixed point as exemplified by the growth of the distance d from the fixed point, Figure 46 right hand side. This standard use of the Method of Manufactured Solution is therefore not fit for the chosen problem that exhibits chaotic trajectories. Since the latter situation is generic, and of particular relevance for complex systems, those which require in particular numerical simulation support, one is led to conclude that this method is of restricted relevance for verification purposes.

A.2 Manufactured Solution testing of the PoPe operators

We verify here the integration schemes used in the simulations of the strange attractor, namely the Runge Kutta schemes, RK4 and RK2 with a standard method, akin to the

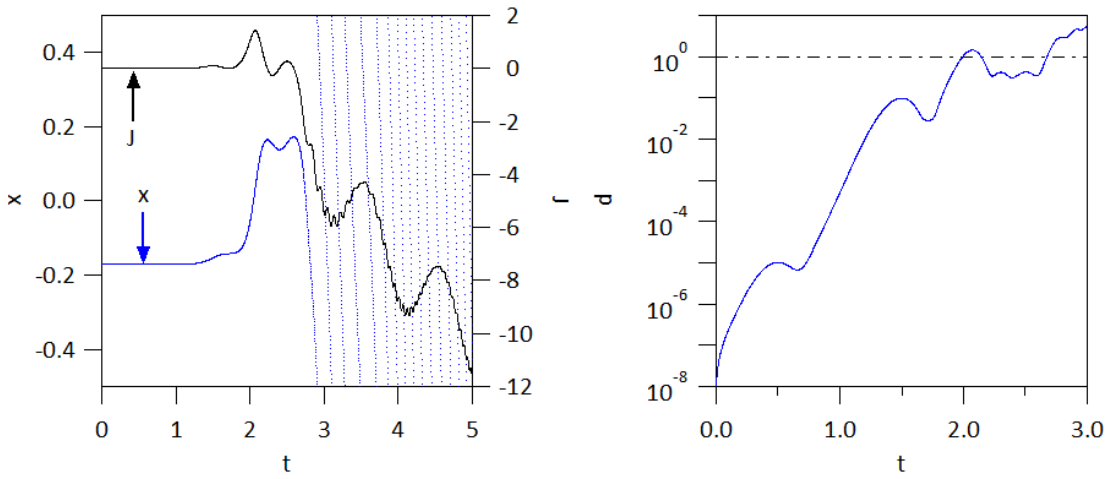

Figure 46: For *case a*, $\sigma_{chir} = 7$ and $\nu = 0.2$, investigation of the fixed point stability for the standard Method of Manufactured Solution, trajectory with chosen fixed point $X = x_0 = -0.17$, $J = 0$ and initial distance from this fixed point $d_0 = 10^{-8}$. Left hand side: trace of x (blue dotted curve) and J (black curve). Right hand side: variation of the distance d from the fixed point, initial value $d_0 = 10^{-8}$ with rapid growth to macroscopic values, $d \approx 1$ on a time scale of $\delta t \approx 2$.

.

Method of Manufactured Solutions. We consider a problem with known solution analytic so that one can measure the error. We thus consider the equation

$$\frac{dJ}{dt} = -\sin\left(2 * \pi t\right) \tag{40}$$

with known solution $J(t) = \cos(t)$ for initial conditions $J = -1$ at $t = -\pi$. One can then compute the error $E_{RKi}(N) = max(|J_{RKi}(N,t) - J_M(t)|)_t$ where J_{RKi} is the value of J computed with the Runge Kutta scheme of order i, and J_M the known analytical solution. We retain here the largest error taken over one period of the solution. Changing the number of steps per period according to $N = 2^n$ with $3 \leq n \leq 25$, hence the step $1/N$, allows one to check the implementation of the Runge Kutta schemes, Figure (47). One can thus observe that the error behaves with the appropriate order until the number of steps is so large that the numerical noise, typically proportional to the number of steps N becomes larger than the error governed by the numerical scheme. One can thus state that this check is a verification of the Runge Kutta schemes used to determine numerically the trajectories that generate the strange attractors. However, as discussed in Section A.1, this verification gives no information regarding the accuracy: *(i)* because the chosen solution has different characteristic properties compared to the problem to be addressed, *(ii)* because a criterion must be defined to be able to discuss the accuracy.

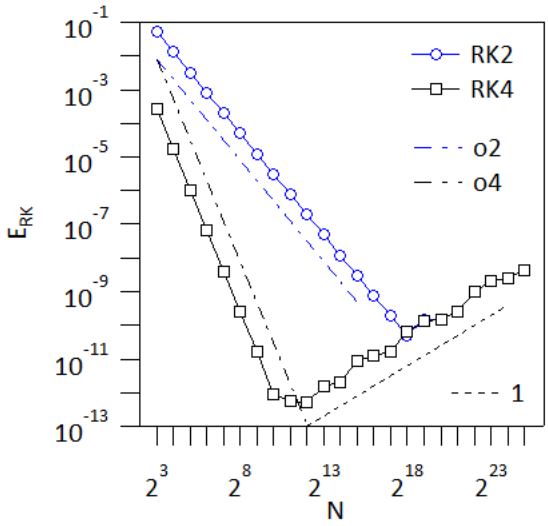

Figure 47: Error E_{RK} obtained with the Manufactured Method Solution for the Runga Kutta schemes, order 2 blue open circles and order 4 black open squares. Dash dot lines, respectively blue for order 2 and black for order 4, indicate the corresponding slopes for order 2 and order 4 error. The dashed black line is indicative of the slope N^1 which fits the loss of accuracy when N is too large.

A.3 Strange attractor MMS evolution equations

The Method of Manufactured Solution leads to a modification of the right hand side of the evolution equations to generate known fixed points (x_0, J_0):

$$\frac{dx}{dt} = J - J_0 \tag{41a}$$

$$\frac{dJ}{dt} = -2\pi\ B\Big(\sin\big(2\pi x\big) + \sin\big(2\pi(x-t)\big)\Big) - \nu\ J + 2\pi\ B\Big(\sin\big(2\pi x_0\big) + \sin\big(2\pi(x_0-t)\big)\Big) + \nu\ J_0 \tag{41b}$$

Given $\sin(A) - \sin(B) = 2\cos(a)\sin(b)$ where $a = (A+B)/2$ and $b = (A-B)/2$, one then obtains:

$$\frac{dx}{dt} = 2Z \tag{42a}$$

$$\frac{dJ}{dt} = -4\pi\ B\sin\Big(2\pi\big(X-x_0\big)\Big)\Big[\sin\Big(2\pi X\Big) + \sin\Big(2\pi\big(X-t\big)\Big)\Big] - \nu\ 2Z \tag{42b}$$

where $Z = (J - J_0)/2$ and $X = (x + x_0)/2 + 1/4$ and therefore:

$$\frac{dX}{dt} = Z \tag{43a}$$

$$\frac{dZ}{dt} = -2\pi\ B\sin\Big(2\pi\big(X-x_0-\tfrac{1}{4}\big)\Big)\Big[\sin\Big(2\pi X\Big) + \sin\Big(2\pi\big(X-t\big)\Big)\Big] - \nu\ Z \tag{43b}$$

One can readily see that $(X = x_0 + \frac{1}{4},\ Z = 0)$, and therefore $(x = x_0,\ J = J_0)$, is the chosen fixed point. The system used for the MMS is therefore close but not identical to that generating the strange attractor since one has to multiply the potential amplitude B by the space dependent function $\sin(2\pi(X - x_0 - \frac{1}{4}))$.

B Scaling of MRS error

B.1 Forward and backward transforms

Let us consider a time integration with initial condition x_0, stored with a time step h generating a trajectory $(x_0, x_1, x_2, \ldots, x_{n-1}, x_n)$ at $t = 0$, $t = h$, $t = 2\ h \ldots t = (n-1)\ h$, $t = n\ h$. We define an approximate trajectory $(y_0, y_1, y_2, \ldots, y_{n-1}, y_n)$ generated by a time integration scheme of order 2α and note F the time derivative of x generating the trajectory and G the function of x to recover the exact trajectory.

$$x_k = x_{k-1} + h\ F(x_{k-1}) + h^{2\alpha+1} G(x_{k-1}) \tag{44a}$$
$$y_k = y_{k-1} + h\ F(y_{k-1}) \tag{44b}$$

Reversing time to step back towards the initial condition we generate the backward trajectories $(\widetilde{x}_n, \widetilde{x}_{n-1}, \ldots, \widetilde{x}_2, \ldots, \widetilde{x}_1, \widetilde{x}_0)$ and $(\widetilde{y}_n, \widetilde{y}_{n-1}, \ldots, \widetilde{y}_2, \ldots, \widetilde{y}_1, \widetilde{y}_0)$ with transform:

$$\widetilde{x}_{k-1} = \widetilde{x}_k - h\ F(\widetilde{x}_k) - h^{2\alpha+1}\ G(\widetilde{x}_k) \tag{45a}$$
$$\widetilde{y}_{k-1} = \widetilde{y}_k - h\ \ G(\widetilde{y}_k) \tag{45b}$$

We can now proceed to defining the n step return transforms made of n step forwards followed by n steps backwards.

B.2 Distance between initial and return point

We are interested in the distance between the upward and downward computed trajectories typically $d_k = \widetilde{y}_y - y_k$. We want to relate d_k to d_{k+1} to determine a series. We split the contribution to d_k into two terms, introducing the distance to the exact trajectory, and reversible, trajectory $x_k = \widetilde{x}_k$.

$$d_k = \widetilde{y}_k - y_k = \widetilde{y}_k - \widetilde{x}_k + x_k - y_k \tag{46a}$$
$$\widetilde{y}_k - \widetilde{x}_k = \widetilde{y}_{k+1} - \widetilde{x}_{k+1} - h\big(F(\widetilde{y}_{k+1}) - F(\widetilde{x}_{k+1})\big) + h^{2\alpha+1} G(\widetilde{x}_{k+1}) \tag{46b}$$
$$y_{k+1} - x_{k+1} = y_k - x_k + h\big(F(y_k) - F(x_k)\big) - h^{2\alpha+1} G(x_k) \tag{46c}$$

One then expands the difference $F(\widetilde{y}_{k+1}) - F(\widetilde{x}_{k+1})$ so that:

$$F(\widetilde{y}_{k+1}) - F(\widetilde{x}_{k+1}) = \big(\widetilde{y}_{k+1} - \widetilde{x}_{k+1}\big) F'(\widetilde{x}_{k+1}) \tag{47a}$$

Similarly, one can expand $F(y_k) - F(x_k)$:

$$F(y_k) - F(x_k) = \big(y_k - x_k\big) F'(x_k) \tag{47b}$$

One can then rewrite Eq.(46b) and Eq.(46c).

$$\widetilde{y}_k - \widetilde{x}_k = \big(\widetilde{y}_{k+1} - \widetilde{x}_{k+1}\big)\big(1 - hF'(\widetilde{x}_{k+1})\big) + h^{2\alpha+1} G(\widetilde{x}_{k+1}) \tag{48a}$$
$$y_{k+1} - x_{k+1} = \big(y_k - x_k\big)\big(1 + hF'(x_k)\big) - h^{2\alpha+1} G(x_k) \tag{48b}$$

One then obtains the two contributions to the distance d_k.

$$\widetilde{y}_k - \widetilde{x}_k = \big(\widetilde{y}_{k+1} - \widetilde{x}_{k+1}\big)\big(1 - hF'(\widetilde{x}_{k+1})\big) + h^{2\alpha+1} G(\widetilde{x}_{k+1}) \tag{49a}$$
$$\big(x_k - y_k\big)\big(1 + hF'(x_k)\big) = x_{k+1} - y_{k+1} - h^{2\alpha+1} G(x_k) \tag{49b}$$

At this stage, the assumption $h|F'(x_k)| \ll 1$ considerably simplifies the calculation, so that:

$$\widetilde{y}_k - \widetilde{x}_k = \widetilde{y}_{k+1} - \widetilde{x}_{k+1} + h^{2\alpha+1} G(\widetilde{x}_{k+1}) \tag{50a}$$
$$x_k - y_k = x_{k+1} - y_{k+1} - h^{2\alpha+1} G(x_k) \tag{50b}$$

We then obtain the recurrence relationship between the distances d_k and d_{k+1}.

$$d_k = d_{k+1} + b_{k,k+1} \tag{51a}$$
$$b_{k,k+1} = h^{2\alpha+1}\big(G(\widetilde{x}_{k+1}) - G(x_k)\big) \tag{51b}$$

Without the previous assumption, the recurrence would also be geometrical, making the final result a bit more complicated.

$$d_0 = d_n + \sum_{k=0}^{n-1} b_{k,k+1} \tag{52}$$

For a return after n steps one enforces $d_n = 0$ removing the contribution of the purely geometrical recurrence. In the general case this leaves various contributions from the coefficients $b_{k,k+1}$, which are all proportional to $h^{2\alpha+1}$, hence of the order determined by the integration scheme. One can then note that:

$$b_{k,k+1} = h^{2\alpha+1}\big(G(x_{k+1}) - G(x_k)\big) \tag{53a}$$
$$\sum_{k=0}^{n-1} b_{k,k+1} = h^{2\alpha+1}\big(G(x_n) - G(x_0)\big) \tag{53b}$$

Two cases are then found if n is not too large, one can expand $G(x_n)$ so that:

$$d_0 = \sum_{k=0}^{n-1} b_{k,k+1} \approx h^{2\alpha+1}\big(x_n - x_0\big)G'(x_0) \approx h^{2\alpha+2}\Big(\sum_{k=0}^{n-1} F(x_k)\Big)G'(x_0) \tag{54}$$

In this case the distance d_0 scales like $h^{2\alpha+2}$. In the other case, when n is too large, one obtains a scaling $h^{2\alpha+1}$.

References

[1] M. Barnes, F. I. Parra, and A. A. Schekochihin. Critically Balanced Ion Temperature Gradient Turbulence in Fusion Plasmas. *Phys. Rev. Lett.*, 107:115003, Sep 2011.

[2] Giancarlo Benettin, Luigi Galgani, Antonio Giorgilli, and Jean-Marie Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for hamiltonian systems; a method for computing all of them. part 1: Theory. *Meccanica*, 15(1):9–20, 1980.

[3] Gert-Jan Both, Subham Choudhury, Pierre Sens, and Remy Kusters. Deepmod: Deep learning for model discovery in noisy data. 2021.

[4] Emily Bourne, Yann Munschy, Virginie Grandgirard, Michel Mehrenberger, and Philippe Ghendrih. Non-Uniform Splines for Semi-Lagrangian Kinetic Simulations of the Plasma Sheath. working paper or preprint, August 2022.

[5] H. Bufferand, J. Bucalossi, G. Ciraolo, G. Falchetto, A. Gallo, Ph. Ghendrih, N. Rivals, P. Tamain, H. Yang, G. Giorgiani, F. Schwander, M. Scotto d'Abusco, E. Serre, Y. Marandet, and M. Raghunathan. Progress in edge plasma turbulence modelling -hierarchy of models from 2D transport application to 3D fluid simulations in realistic tokamak geometry. *Nuclear Fusion*, 61(11):116052, oct 2021.

[6] T Cartier-Michaud. to appear in panoramas & synthèses, *numerical model for fusion.* 2014.

[7] T. Cartier-Michaud, D. Galassi, Ph. Ghendrih, P. Tamain, F. Schwander, and E. Serre. A posteriori error estimate in fluid simulations of turbulent edge plasmas for magnetic fusion in tokamak using the data mining iPoPe method. *Physics of Plasmas*, 27(5):052507, 2020.

[8] T. Cartier-Michaud, P. Ghendrih, Y. Sarazin, J. Abiteboul, H. Bufferand, G. Dif-Pradalier, X. Garbet, V. Grandgirard, G. Latu, C. Norscini, C. Passeron, and P. Tamain. Projection on Proper elements for code control: Verification, numerical convergence, and reduced models. Application to plasma turbulence simulations. *Physics of Plasmas*, 23(2), 2016.

[9] Thomas Cartier-Michaud. *Vérification de Codes et Réduction de Modèles : Application au Transport dans les Plasmas Turbulents.* Theses, Aix-Marseille Universite, June 2015.

[10] Thomas Cartier-Michaud, Philippe Ghendrih, Guilhem Dif-Pradalier, Xavier Garbet, Virginie Grandgirard, Guillaume Latu, Yanick Sarazin, Frederic Schwander, and Eric Serre. Verification of turbulent simulations using PoPe: quantifying model precision and numerical error with data mining of simulation output. *Journal of Physics: Conference Series*, 1125(1):012005, 2018.

[11] Boris V Chirikov. A universal instability of many-dimensional oscillator systems. *Physics Reports*, 52(5):263 – 379, 1979.

[12] X. Garbet, L. Laurent, J.-P. Roubin, and A. Samain. A model for the turbulence in the scrape-off layer of tokamaks. *Nuclear Fusion*, 31(5):967, 1991.

[13] Philippe Ghendrih. *Turbulence faible dans un système mécanique peut dissipatif : étude du processus de transition et caractérisation des états chaotiques.* PhD thesis, Université Pierre et Marie Curie, Paris VI, 1983.

[14] V. Grandgirard, J. Abiteboul, J. Bigot, T. Cartier-Michaud, N. Crouseilles, G. Dif-Pradalier, Ch. Ehrlacher, D. Esteve, X. Garbet, Ph. Ghendrih, G. Latu, M. Mehrenberger, C. Norscini, Ch. Passeron, F. Rozar, Y. Sarazin, E. Sonnendrucker, A. Strugarek, and D. Zarzoso. A 5d gyrokinetic full-f global semi-lagrangian code for flux-driven ion turbulence simulations. *Computer Physics Communications*, 207:35 – 68, 2016.

[15] V Grandgirard, Y Sarazin, P Angelino, A Bottino, N Crouseilles, G Darmet, G Dif-Pradalier, X Garbet, Ph Ghendrih, S Jolliet, G Latu, E Sonnendrücker, and L Villard. Global full- f gyrokinetic simulations of plasma turbulence. *Plasma Physics and Controlled Fusion*, 49(12B):B173, 2007.

[16] B. Malraison, P. Atten, P. Berge, and M. Dubois. Dimension of strange attractors: an experimental determination for the chaotic regime of two convective systems. *Journal de Physique lettres*, 44:897–902, 1983.

[17] A. V. Nedospasov. The enhancement of edge turbulence in tokamaks by a limiter current. *Physics of Fluids B: Plasma Physics*, 5(9):3191–3194, 1993.

[18] Christiane Normand, Yves Pomeau, and Manuel G. Velarde. Convective instability: A physicist's approach. *Review of Modern Physics*, 49:581–624, Jul 1977.

[19] William L. Oberkampf and Christopher J. Roy. *Verification and Validation in Scientific Computing*. Cambridge University Press, USA, 1st edition, 2010.

[20] William L. Oberkampf and Timothy G. Trucano. Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences*, 38(3):209–272, 2002.

[21] Fabio Riva, Paolo Ricci, Federico D. Halpern, Sébastien Jolliet, Joaquim Loizu, and Annamaria Mosetto. Verification methodology for plasma simulations and application to a scrape-off layer turbulence code. *Physics of Plasmas*, 21(6):062301, 2014.

[22] Patrick J. Roache. Code Verification by the Method of Manufactured Solutions . *Journal of Fluids Engineering*, 124(1):4–10, 11 2001.

[23] Y. Sarazin and Ph. Ghendrih. Intermittent particle transport in two-dimensional edge turbulence. *Physics of Plasmas*, 5(12):4214–4228, 1998.

[24] M. Scotto d'Abusco, G. Giorgiani, J.F. Artaud, H. Bufferand, G. Ciraolo, P. Ghendrih, E. Serre, and P. Tamain. Core-edge 2Dfluid modeling of full tokamak discharge with varying magnetic equilibrium: from WEST start-up to ramp-down. *Nuclear Fusion*, 62(8):086002, may 2022.

[25] M. Tagger, G. Laval, and R. Pellat. Trapped ion mode driven by ion magnetic drift resonance in a fat torus. *Nuclear Fusion*, 17(1):109, 1977.

[26] P. Tamain, H. Bufferand, G. Ciraolo, C. Colin, D. Galassi, Ph. Ghendrih, F. Schwander, and E. Serre. The tokam3x code for edge turbulence fluid simulations of tokamak plasmas in versatile magnetic geometries. *Journal of Computational Physics*, 321:606 – 623, 2016.