Density-Induced Variations of Local Dimension Estimates for Absolutely Continuous Random Variables

Paul Platzer, Bertrand Chapron

Density-Induced Variations of Local Dimension Estimates for Absolutely Continuous Random Variables

Paul Platzer[1] and Bertrand Chapron[1]

[1]Laboratoire d'Océanographie Physique et Spatiale (LOPS), Ifremer, 1625 route de Sainte-Anne, Plouzané, 29280, Bretagne, France.

Contributing authors: paul.platzer@ifremer.fr;

Abstract

For any multi-fractal dynamical system, a precise estimate of the local dimension is essential to infer variations in its number of degrees of freedom. Following extreme value theory, a local dimension may be estimated from the distributions of pairwise distances within the dataset. For absolutely continuous random variables and in the absence of zeros and singularities, the theoretical value of this local dimension is constant and equals the phase-space dimension. However, due to uneven sampling across the dataset, practical estimations of the local dimension may diverge from this theoretical value, depending on both the phase-space dimension and the position at which the dimension is estimated. To explore such variations of the estimated local dimension of absolutely continuous random variables, approximate analytical expressions are derived and further assessed in numerical experiments. These variations are expressed as a function of 1. the random variables' probability density function, 2. the threshold used to compute the local dimension, and 3. the phase-space dimension. Largest deviations are anticipated when the probability density function has a low absolute value, and a high absolute value of its Laplacian. Numerical simulations of random variables of dimension 1 to 30 allow to assess the validity of the approximate analytical expressions. These effects may become important for systems of moderately-high dimension and in case of limited-size datasets. We suggest to take into account this source of local variation of dimension estimates in future studies of empirical data. Implications for weather regimes are discussed.

Keywords: local dimension, information dimension, attractor dimension, random variable, distribution, curse of dimensionality, weather regimes

1 Introduction

Local dimension estimations are tools to study multifractal measures with local density exhibiting multiple scaling exponent. A first approach to study such measures is global, looking at these scaling exponents over the measure's whole attractor, through what is called the spectrum of generalized dimensions [1–3]. The other approach is local, examining variations of estimated dimensions at different points of the attractor. Such an approach is widely used to study dynamical properties of atmospheric circulation [4–12], building on mathematical developments linking dynamical systems theory and extreme value theory [13]. These local dimensions allow to assess the probability distributions of distances for "analogs" [14], often used in atmospheric science for several applications [e.g. 15–18]. Local and global approaches can be reconciled, as [19] showed that the spectrum of generalized dimensions can be deduced from the ensemble of local dimensions estimates.

These dimension-estimation tools are designed for multifractal measures, and should in principle give trivial results when applied to random variables with smooth probability density functions. However, in practice, dimension estimates can be biased. [20] showed that in high dimension, the curse of dimensionality induces dimension

estimates inferior to what is expected from the multi-fractal formalism of dynamical systems (i.e., that the local dimension should equal the phase-space dimension). It has also been noted that the dimension estimates are anomalously high in areas of low density [6], such as the borders of the wings of the three-variable convective Lorenz system [21].

In this work, we explore these seemingly intrinsic variations with position in phase-space of the estimates of local dimension for random variables possessing an absolutely continuous probability density function. We use Taylor expansions with the hyper-sphere-radius used to compute local dimensions, to derive analytical approximate expressions for the estimates of local dimension, leading to a typical formula that can be used to compute the latter from empirical data. These expressions are then compared to true empirical estimates of local dimension from numerically generated data corresponding to 1. a one-dimensional double-well stochastic system, 2. a two-dimensional Gaussian Mixture Model, and 3. a standard multivariate Gaussian of arbitrary dimension.

Section 2 recalls the basic definitions and provides analytical derivations for the approximate deviation of local dimension estimates from the phase-space dimension. Section 3 provides particular cases of the analytical expressions, and describes numerical experiments used to validate these expressions. Finally, section 4 gives concluding remarks and discusses implications for studies of weather regimes based on dynamical indicators.

2 Theoretical background

2.1 Definitions

Let us consider a dynamical system with invariant measure μ. For any point x in the support of μ, the local, r-resolution dimension at point x follows:

$$d(x,r) := \frac{\log \mu(B_{x,r})}{\log r}\,, \tag{1}$$

where $B_{x,r}$ is the ball of radius r centered on x, and $r > 0$. The limit for small r of this local dimension, when it exists, is denoted $d(x)$. If $d(x)$ exists for all x and if μ is ergodic, then $d(x)$ is constant μ-almost everywhere and the system is said to be exact-dimensional [22]. In this case the typical value of $d(x)$ is noted D_1 and is called the first-order Renyi dimension or information dimension. It is also equal to the μ-average of the local dimensions :

$$D_1 := \lim_{r\to 0} \frac{\int \log\left(\mu(B_{x,r})\right) d\mu(x)}{\log r}\,. \tag{2}$$

[19] showed that the local dimensions $d(x,r)$ follow a large deviation principle around their μ-average value D_1 as $r \to 0$. This gives information on the probability density of $d(x,r)$ when $|d(x,r) - D_1|$ exceeds a given threshold.

2.2 Numerical estimation

We assume that we are provided with a long time-series of $\{x_i\}_{1\leq i\leq N}$ from the dynamical system defining μ, where N is a large integer.

Computing $d(x,r)$ at fixed x through Eq. (1) with a Birkhoff sum to estimate $\mu(B_{x,r})$ gives a slow convergence to D_1 for small values of r. Instead, methods relying on several values of $\mu(B_{x,r})$ for small r give more satisfying results. Let $K \in \mathbb{N}$, such that K/N is small enough to ensure small bias but large enough to ensure small variance. Note $r_1 < \ldots < r_K$ the ordered distances to the K nearest neighbours of x in the dataset $\{x_i\}_{1\leq i\leq N}$. Then the following expression is an estimator of $d(x, r_K)$ (see [18]):

$$\hat{d}(x,r_K) := \left\{ \sum_{k=2}^{K} \frac{k}{K} \log\left(\frac{r_k}{r_{k-1}}\right) \right\}^{-1}\,. \tag{3}$$

Note also that, if we assume that there are constant values d, $\mu_0 > 0$ such that, for $r < r_K$, $\mu(B_{x,r}) = \mu_0 r^d$, then the integral in the right-hand side of Eq. (4) equals exactly d^{-1}. In the right-hand side of Eq. (3), the sum is an approximation of:

$$\sum_{k=2}^{K} \frac{k}{K} \log\left(\frac{r_k}{r_{k-1}}\right) \approx \int_0^{r_K} \frac{\mu(B_{x,r})}{\mu(B_{x,r_K})} \frac{\mathrm{d}r}{r} \tag{4}$$

by using the approximation $\log\left(\frac{r_k}{r_{k-1}}\right) \approx \frac{r_k - r_{k-1}}{r_k}$, which is valid only when $\frac{r_k - r_{k-1}}{r_k}$ is small. Note that from [18], we have the scaling $r_k \sim k^{1/D_1}$, so that the previous approximations holds in particular for medium-to-large values of the dimension, and for medium-to-large values of k. For instance, if $D_1 = 1$ and $k = 1$, we have the scaling $\frac{r_k - r_{k-1}}{r_k} \sim 1$ and so this approximation barely holds. On the contrary, if $D_1 = 3$ and $k = 4$, then $\frac{r_k - r_{k-1}}{r_k} \sim 0.07$. Thus for practical applications, this approximation should hold. This allows to directly reassess $\hat{d}(x, r_K)$ as a function of r:

$$\hat{d}(x, r) \approx \left\{ \int_0^{r} \frac{\mu(B_{x,r'})}{\mu(B_{x,r})} \frac{\mathrm{d}r'}{r'} \right\}^{-1}. \tag{5}$$

In the following, we will focus on the behaviour of $\hat{d}(x, r)$ using this expression.

2.3 Expansion for absolutely continuous random variables

2.3.1 Fixed radius

In this section, the attractor measure μ describes the probability of an absolutely continuous random variable, i.e. the following formula is true for any n-dimensional phase-space volume V:

$$\mu(x \in V) = \int_V p(x) \mathrm{d}^n x \tag{6}$$

where $p(x)$ is the probability density function of the random variable, and a smooth function of x. We also assume that $p(x)$ has no zeros, and no singularity (*i.e.* for all x, $0 < p(x) < +\infty$). This condition is necessary, as one could otherwise build

absolutely continuous random variables that have a continuous spectrum of generalized dimensions, as in [23]. With our hypothesis, the quantity $\mu(B_{x,r})$ admits a Taylor expansion for small r:

$$\mu(B_{x,r}) = \int_{B_{0,r}} p(x+u)\mathrm{d}^n u \tag{7}$$

$$= \int_{B_{0,r}} \left\{ p(x) + \nabla p(x) \cdot u + \frac{1}{2} u \cdot [H(p)(x)\, u] + \mathcal{O}(u^3) \right\} \mathrm{d}^n u \tag{8}$$

where $\nabla p(x)$ denote the n-dimensional gradient of p at x, and $H(p)(x)$ denotes the $n \times n$ - dimensional Hessian matrix of p at x, the matrix of second-order derivatives, and centered dot "$\cdot$" denotes scalar product.

The first term in the integral is constant and gives $p(x)\alpha_n r^n$ where $\alpha_n > 0$ is the volume of a radius-1, n-dimensional ball. Through symmetry in the ball $B_{0,r}$ the integral $\int_{B_{0,r}} u\mathrm{d}^n u$ of the odd function $u \mapsto u$ vanishes and so does the second term in the integral. Finally the third term can be re-written as a sum of odd and even functions.

$$\int u \cdot [H(p)(x)\, u]\, \mathrm{d}^n u = \sum_{i \neq j} \partial_i \partial_j p \int u_i u_j \mathrm{d}^n u + \sum_i \partial_i^2 p \int u_i^2 \mathrm{d}^n u\,. \tag{9}$$

In this expression, terms that depend on cross-derivatives along different directions vanish, and the sum of non-vanishing terms amounts to $\frac{1}{2n}\Delta p(x)\beta_n r^{n+2}$ where $\Delta p(x) = \sum_i \partial_i^2 p(x)$ is the Laplacian of p at x (i.e. the trace of the Hessian matrix) and β_n is the integral $\int_{B_{0,1}} u^2 \mathrm{d}^n u$. Through vanishing integral of odd functions the fourth term of order $\mathcal{O}(u^3)$ (non-written here) also vanishes, so that one can write:

$$\mu(B_{x,r}) = p(x)\alpha_n r^n + \frac{\beta_n}{2n}\Delta p(x) r^{n+2} + \mathcal{O}(r^{n+4})\,. \tag{10}$$

Coming back to $\hat{d}(x,r)$, one can also estimate the following integral as:

$$\int_0^r \frac{\mu(B_{x,r'})}{r'}\mathrm{d}r' = \left(\frac{1}{n}\right) p(x)\alpha_n r^n + \left(\frac{1}{n+2}\right)\frac{\beta_n}{2n}\Delta p(x) r^{n+2} + \mathcal{O}(r^{n+4})\,, \tag{11}$$

which gives, after manipulation, and using the fact that $(n+2)\beta_n = n\alpha_n$:

$$\hat{d}(x,r) = n\left\{1 + \frac{\Delta p(x)}{p(x)}\left(\frac{r}{n+2}\right)^2\right\} + \mathcal{O}(r^4)\,. \tag{12}$$

This final expression shows that, for absolutely continuous attractor measures μ, the first order deviations of $\hat{d}(x,r)$ from the exact, integer phase-space dimension n is of order r^2. The Laplacian of $p(x)$ is positive (resp. negative) in case of local minima (resp. maxima) of probability. This means that in highly sampled areas, the dimension decreases, while around poorly sampled areas the dimension increases. However, this effect is also balanced by a factor $p(x)^{-1}$, and therefore the position of extrema of $\hat{d}$ differ from those of p in general.

In one dimension, Eq. (12) reads:

$$\hat{d}(x,r) = 1 + \frac{\partial_x^2 p(x)}{9p(x)} r^2 + \mathcal{O}(r^4)\,. \tag{13}$$

where $\partial_x^2 p(x)$ is the second-order derivative of p at x. In two dimensions (x,y), we have:

$$\hat{d}((x,y),r) = 2 + \frac{(\partial_x^2 + \partial_y^2)p(x,y)}{8\,p(x,y)} r^2 + \mathcal{O}(r^4)\,. \tag{14}$$

One can check that when taking the μ-average of $\hat{d}(x,r)$, we have:

$$\int_\Omega \frac{\Delta p(x)}{p(x)}\,\mathrm{d}\mu = \int_\Omega \Delta p(x)\mathrm{d}^n x \tag{15}$$

$$= \int_{\delta\Omega} \nabla p(x)\cdot \mathrm{d}^{n-1}x\,, \tag{16}$$

where the last integral is the flux of the gradient of p at the border of the whole domain Ω, which is zero. This gives:

$$\int \hat{d}(x,r)\mathrm{d}\mu(x) = n + \mathcal{O}(r^4) \,, \tag{17}$$

which is a low-order particular case of the general statement that the μ-average of local dimensions is the order-1 Reiny dimension (here n).

2.3.2 Fixed quantile

In practice, when trying to compute local dimensions, one rarely fixes the radius r, but rather the quantile q which is the proportion of data used to compute the local dimensions. Indeed, fixing the radius can become complicated when data are poorly sampled, as this would imply to rely on very few points for computing d.

The quantile q can be related to the radius and probability density function by noting that $q = \mu\left(B_{x,r}\right)$ by definition, and recalling Eq. (10), which gives at first order:

$$r = \left(\frac{q}{p(x)\alpha_n}\right)^{\frac{1}{n}} \left(1 + \mathcal{O}\left(\frac{r^2}{n}\right)\right) \,. \tag{18}$$

Inserting this in Eq. (12) gives:

$$\hat{d}(x,q) \approx n \left\{1 + \frac{\Delta p(x)}{p(x)^{1+2/n}} \frac{(\Gamma(\frac{n}{2}+1)q)^{2/n}}{\sqrt{\pi}(n+2)^2}\right\} \,, \tag{19}$$

where Γ is the Gamma function that enters into the expression of the volume of a radius-1, n-ball: $\alpha_n = \pi^{n/2}/\Gamma(n/2+1)$. In the case of large n, one can recover Eq. (17) as the second-term of the right-hand side of Eq. (19) is still approximately proportional to $\Delta p(x)/p(x)$.

Eq. (19) has a less straightforward dependency with n than Eq. (12), highlighting the dependency of r with n when q is fixed. However, another dependency with dimension is hidden in the ratio $\Delta p(x)/p(x)$, as probability density functions also strongly

depend on dimension. For instance, the probability density function of a standard normal vector evaluated at 0 decreases with dimension n as $(2\pi)^{-n/2}$. Particular cases are outlined below to better understand these expressions.

3 Particular cases and numerical experiments

3.1 Double-well potential

First consider a one-dimensional example of a stochastic system emanating from the following stochastic differential equation (SDE, see e.g. [24]):

$$\mathrm{d}x = -\partial_x V(x)\mathrm{d}t + \sigma \mathrm{d}W \,, \tag{20}$$

where $x(t)$ is real-valued, t is time, $\sigma > 0$ and W is a Wiener-process of variance $\mathrm{d}t$, with the following potential:

$$V(x) = (1 - x^2)^2 \,, \tag{21}$$

which is the famous symmetric double-well. In particular, this potential has the following drift and second-derivative:

$$-\partial_x V = 4(1 - x^2)x \,, \tag{22}$$

$$\partial_x^2 V = 4(3x^2 - 1) \,. \tag{23}$$

This potential has two stable equilibrium at $x = \pm 1$ and one unstable equilibrium at $x = 0$. We have $\partial_x^2 V(\pm 1) = 8$, and $\partial_x^2 V(0) = -4$.

The Fokker-Planck Equations associated with the above SDE is:

$$\partial_t p(x,t) = \partial_x \left[p(x,t)\partial_x V(x) \right] + \partial_x^2 \left[\frac{\sigma^2}{2} p(x,t) \right] \,, \tag{24}$$

which has the static solution :

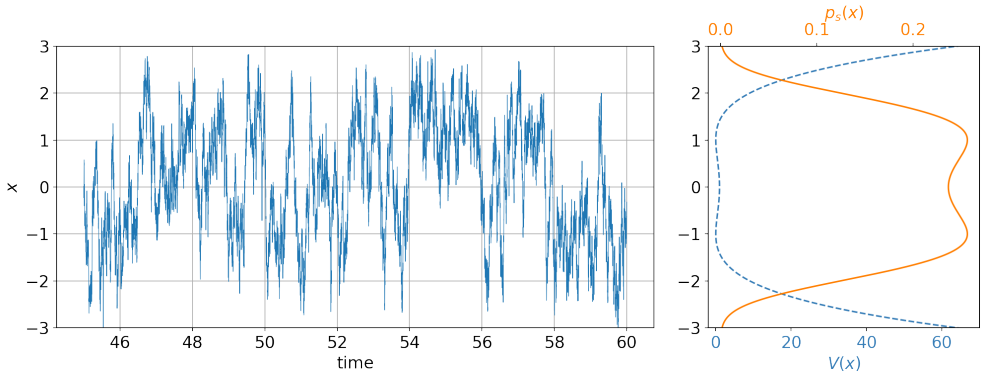

Fig. 1 Left: example of trajectory following Eq. (20) with potential (21), and noise amplitude $\sigma = 5$. Right: corresponding potential (dashed blue line) and static probability density (full orange line).

$$p_s(x) = \frac{\exp\left(-\frac{2V(x)}{\sigma^2}\right)}{\int_{-\infty}^{+\infty} \exp\left(-\frac{2V(u)}{\sigma^2}\right) \mathrm{d}u} \,. \tag{25}$$

One simulation of this stochastic system with the Euler–Maruyama method and a time step of 10^{-3} is shown in the left panel of Fig. 1. The right panel shows the corresponding potential and static probability density function. We can see the typical behaviour of this system, jumping randomly from one well to another.

From section 2.3, dimension estimates are expected to deviate from the true dimension $n = 1$, with lower dimensions around the wells of the potential, and higher dimensions not only at the centered unstable fixed point but also to the right and left of the wells. More precisely, combining $p_s(x)$ from Eq. (25), expressed from Eq. (21), with Eq. (13) gives:

$$\hat{d}(x, r) = 1 - \frac{8}{9\sigma^2}\left(3x^2 - 1 + \frac{4}{\sigma^2}(1 - x^2)^2 x^2\right) r^2 + \mathcal{O}(r^4) \,. \tag{26}$$

A numerical simulation of Eq. (20) is performed with time step 10^{-3}, running for 5×10^5 non-dimensional time. This numerical simulation will serve as a "catalog" from which the distances r_k are computed. The empirical local dimension is then estimated on a regular grid. The interval $-3 < x < 3$ is spanned, using Eq. (3) at

fixed radius $r = 0.3$ by choosing $K(x)$ at each position x so that $r_K < r < r_{K+1}$, and the $\{r_k\}_k$ are the distances between x and the elements of the catalog. These empirical estimates of $\hat{d}(x, r)$ are then compared with the approximate analytical expression Eq. (26), and shown in Fig. 2. The approximation appears to be valid for $-1.5 \lesssim x \lesssim 1.5$, and starts to break down for larger absolute values of the position x. Note that the approximation still captures an interesting feature, also present in the empirical estimates of dimension from the simulated catalog: the position of the minimum of dimension differs from that of the maximum of probability.

This can be important for weather regimes [see e.g. 6]. These regimes are usually defined through the fit of a Gaussian Mixture Model to the empirical probability density of atmospheric circulation data projected onto Empirical Orthogonal Functions. They are therefore defined through the maxima of density. These regimes are studied with dynamical features such as the local dimension, and some studies have shown that peaks of regime index coincide with troughs in dimension [8, 25], arguing that this strengthens the physical meaningfulness of weather regimes. The latter should in principle be associated with less complex dynamics and higher predictability, and therefore lower fractal dimension (as well as higher persistence). With the numerical example, variations of *estimated* local dimension are due to variations in density. It is an artifact and not a sign of a local modification of the true fractal dimension. These local dimension variations have a similar behaviour as the one depicted in [8, 25], with low dimension associated with high density (and therefore high regime index). However, there is a strong shift between the position of the peak density and the position of the trough of estimated dimension. Therefore, the effects depicted here might be candidates to explain the observed variations in dimension estimate observed in [8, 11, 25], at the peak of regimes and at transitions between regimes. However, the fact that the position of the stable fixed points is shifted with respect to the position of minimum dimension seems to indicate that reported variations in dimension estimate

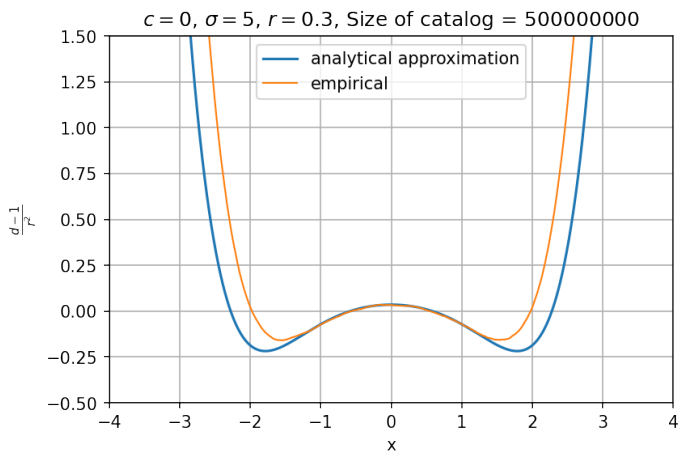

Fig. 2 Empirical dimension estimate versus analytical approximation from a long simulation of the double-well stochastic system.

may certainly not be solely due to the artifacts depicted here. In the next subsection, the case of two-dimensional Gaussian Mixture Models is considered to better assess the possible effect of this artifact.

The simple one-dimensional example can finally be used to investigate one property of Eq. (26), which is the scaling $\hat{d}(x,r) - 1 \sim r^2$. To do so, we take a closer look at a few points x between -1.3 and 0 for which the approximation seems to be valid, Fig. 2. We take regularly sampled values of r^2 between 0.0004 and 0.25, for which the local dimension is estimated with Eq. (3), and K defined as previously through $r_K < r < r_{K+1}$. These estimates are compared with the analytical expression of Eq. (26) in Fig. 3. The agreement is very good between the analytical approximation and the empirically estimated values, validating both the scaling of $\hat{d}(x,r) - 1$ with r^2 and the values of the slopes given by the analytical expression $\frac{\partial_x^2 p(x)}{9p(x)}$ in dimension 1.

3.2 Gaussian Mixture Model

The previous example showed that the position of minima of dimension differs from that of the maxima of probability density. To test this assertion, a Gaussian Mixture Model is considered (GMM, see e.g. [26]), of which k-means [27] are a particular case.

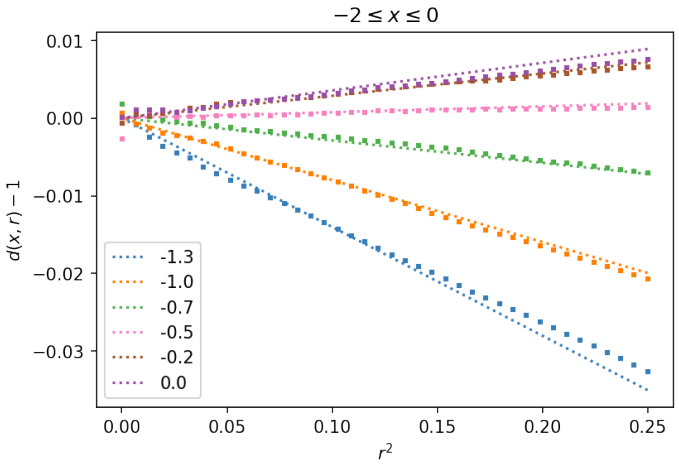

Fig. 3 Local dimension estimate as a function or r^2 for several locations x in the double-well stochastic system, estimated from numerical simulations compared to an analytical approximation. Squares: approximation from Eq. (26). Dotted lines: empirical values from numerical simulation.

Such a model allows to define a random variable as stemming from several components (of the "mixture"), each component being defined by a Gaussian distribution with its own characteristics (mean and covariance matrix).

These statistical models are also typically used to assign atmospheric circulation data to weather regimes. For instance, in [28], four weather regimes are defined through the fitting of a GMM to atmospheric circulation data, smoothed in time and projected on two empirical orthogonal functions, giving a two-dimensional space. As a reminiscence of this configuration, we take here interest in a two-dimensional random variable defined by a GMM with four components. This random variable X has the following distribution p_X:

$$p_X = \sum_{i=1}^{4} \phi_i \mathcal{N}(\boldsymbol{m_i}, \boldsymbol{\Sigma_i}) \,, \tag{27}$$

where $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$ stands for the probability density function of a Gaussian distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$, and each ϕ_i corresponds to the probability of a given component to be selected. We choose to use diagonal covariance matrices for

Parameter \Component	Upper-right	Lower-right	Lower-left	Upper-left
$\boldsymbol{m_i}$	[1.5 , 1.5]	[1 , -1]	[-1 , -0.9]	[-1 , 0.9]
$\boldsymbol{\Sigma_i}$	1.3	0.9	1.2	0.9
ϕ_i	0.25	0.25	0.25	0.25

Table 1 Parameters used for the two-dimensional Gaussian Mixture Model. Covariance matrices are proportional to the identity matrix and therefore only one coefficient is given. The components are given names related to their position in phase-space, as shown in the plots of Fig. 4.

simplicity. The values set for the means $\boldsymbol{m_i}$, covariances $\boldsymbol{\Sigma_i}$ and probabilities ϕ_i are listed in Table 1.

Although feasible, there is no interest in giving the analytical expression for the approximate analytical expression of $\hat{d}(x, r)$ from Eq. () using the expression for the probability density function of this random variable. However, we can visually check the agreement between this analytical expression and the true dimension estimated from numerical experiments. To do so, we draw 10^7 samples of the GMM, and for each two-dimensional position x on a regular grid of 200×200 points ranging from -3.5 to $+3.5$ in both dimensions, we compute the empirical dimension at radius $r = 0.5$ using this randomly sampled data and Eq. (3). The result of this procedure is shown in Fig 4(c), and compared to our approximate analytical expression in Fig 4(b), while the GMM model density is shown in Fig 4(a). A very good agreement between our approximation and the empirical estimates in terms of the position of the minima of estimated dimension (see in particular the position of the minima of dimension on the top-right), as well as the general behaviour (including rising dimension in areas of low density, far from the GMM components). Again, the approximation overestimates the amplitude of these variations, here by a factor of ~ 3. However, the relative intensities of the empirical dimension minima also agree with previsions from our estimates: the bottom-right and top-left troughs of dimension are stronger than the one on the top-right and bottom-left.

This example shows that our approximation captures anomalous variations in estimated dimension for random systems stemming from a GMM. In particular, troughs

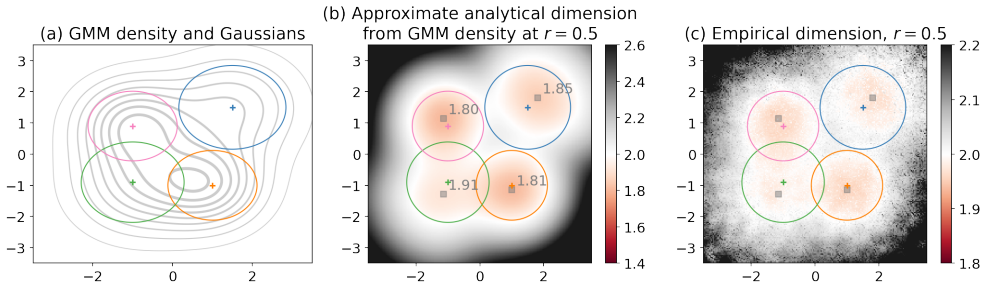

Fig. 4 (a) GMM density (contours) and Gaussian distributions' averages (crosses) and radius at which the probability of the gaussian is twice smaller then its maximum probability (circles). (b) Approximate analytical expression for the dimension estimates from Eq. (3.2) and true GMM density, setting radius $r = 0.5$. Crosses and circles from (a) are repeated for comparison. Squares indicate the position of the minima of dimension, and the corresponding minimum values of dimension are written next to the squares. (c) Same as (b) but for the empirical dimension estimates from numerically sampled GMM and with Eq. (3). The position of the squares of (b) are repeated for comparison.

of dimension are witnessed near the weather regime means, so that the observations of [8, 11, 25] may indeed be associated to such effects of density variation rather than true changes in the fractal properties of the attractor of the underlying atmospheric dynamical system. However, the amplitude of the variations in dimension estimate observed in this example are small, and more investigations are needed to understand how these effects depend on phase-space dimension. This is the subject of the following particular case.

3.3 Multivariate Gaussian

A multivariate Gaussian system is now considered, to more particularly explore the effect of dimensionality on our claims. For a standard multivariate Gaussian, the probability density function is given by:

$$p(x) = \frac{\exp(-\frac{|x|^2}{2})}{(2\pi)^{\frac{n}{2}}} \,, \tag{28}$$

where x is a n-dimensional vector. The Laplacian of $p(x)$ reads:

$$\Delta p(x) = \left(|x|^2 - n\right) \frac{\exp(-\frac{|x|^2}{2})}{(2\pi)^{\frac{n}{2}}} \,. \tag{29}$$

Substituting this into Eq. (19) gives:

$$\hat{d}(x,q) \approx n \left\{ 1 + 2\sqrt{\pi} \left(|x|^2 - n \right) \exp\left(\frac{|x|^2}{n} \right) \frac{(\Gamma(\frac{n}{2}+1)q)^{2/n}}{(n+2)^2} \right\} . \tag{30}$$

Again, the witnessed behaviour is similar to those depicted in the previous experiments, with a decreased dimension towards the area of high probability density (here $x = 0$), and an increased dimension with respect to the theoretical value n in areas of low density (here for large values of $|x|$).

Mentioned above, these formulas are only approximations of the true behaviour of local dimension estimates for data generated from the standard multivariate Gaussian. To test their validity using numerical experiments, we first consider five different positions $x = (0, \ldots, 0)$, $x = (1, 0, \ldots, 0)$, $x = (2, 0, \ldots, 0)$, $x = (3, 0, \ldots, 0)$, $x = (4, 0, \ldots, 0)$; as well as three values of the proportion of data used to compute the local dimensions $q = 10^{-3}$, $q = 10^{-4}$, $q = 10^{-5}$; and finally three values for the exact dimension $n = 2$, $n = 5$, $n = 8$. To each triplet of values (x, q, n) can be associated an approximate value of $\hat{d}(x, q)$ from Eq. (30), which is reported in Fig. 5(b). To compare this to real dimension estimates from numerical experiments, we generate 100 independent datasets for each pair (q, n), each dataset containing $10^3/q$ samples of the standard multivariate Gaussian of exact dimension n. Then, with each dataset we compute the local dimension estimate using Eq. (3) at all five positions x listed above. For each triplet (x, q, n), we therefore obtain 100 values for $\hat{d}(x, q)$. Taking the average over all 100 realisations, we obtain the results shown in Fig. 5(a).

Numerical experiments confirm the same tendency as the ones given by our approximate Eq. (30): slightly lower than n for $x = 0$, and growing with $|x|$, eventually exceeding n. The deviation at $x = 0$ from the exact value n is a growing function of n for the values considered here, both according to our approximation and to the numerical experiments. At fixed position $x = 4$, the opposite behaviour is observed:

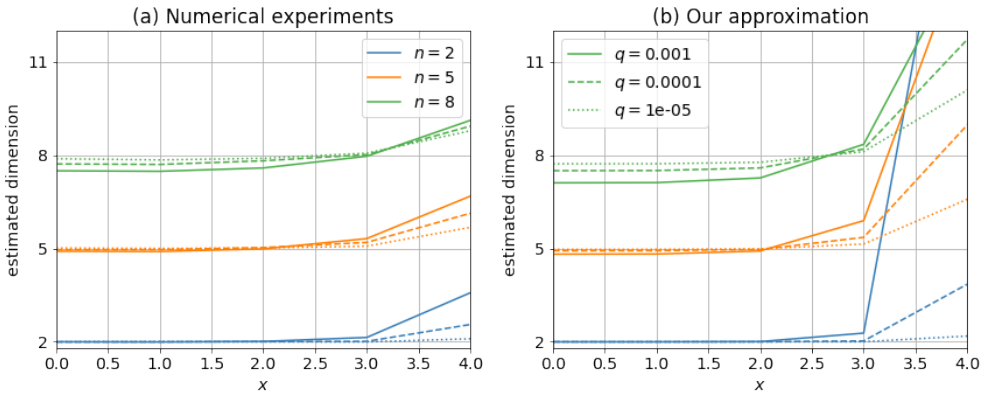

Fig. 5 Estimated local dimension for the multivariate normal distribution, at positions $x = (0, 0, \ldots, 0)$, $x = (1, 0, \ldots, 0)$, up to $x = (4, 0, \ldots, 0)$, for three values of q, the proportion of data used to compute the local dimensions, and three values of n, the exact dimension. (a) Values obtained from numerical experiments, averaged over 100 realisations for each triplet (x, q, n). (b) Approximation from Eq. (30).

the deviation is stronger for small values of n. Note also that, as in the previous experiments, our approximation strongly departs from the numerical values in areas of very small density (here, large values of $|x|$). Our approximation also overestimates the deviation at $x = 0$. However, these numerical experiments display the same behaviour as one would expect from our approximation, suggesting that the latter adequately represents the effect of changing density on variations of numerical estimates of fractal dimension.

A next question is then the following: what is the typical variation of fractal dimension estimate that is only due to density variations, and how does this typical variation depend on the exact dimension n ? To test this, we define the radius $r^*(n, \tau)$, where $0 < \tau < 1$ is the probability that x lies in a ball of radius r^* centered on $x = 0$. This radius r^* is thus defined implicitly through the following equation:

$$\tau = (r^*)^{n-1} \exp\left(\frac{(r^*)^2}{2}\right) , \tag{31}$$

where the right-hand side of this equation is obtained by integrating the probability density function of a standard Gaussian distribution from $r = 0$ to $r = r^*$. We can

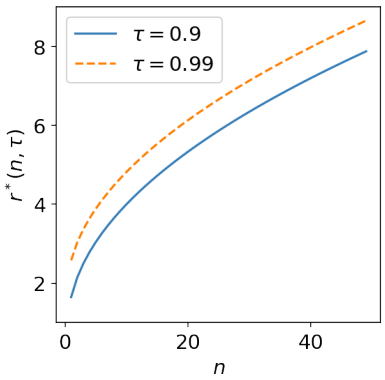

Fig. 6 Plot of radius $r^*(n,\tau)$ versus phase-space dimension n defined through Eq. (31). This is the radius for which the probability of a standard multivariate Gaussian to lie in a 0-centered ball of radius r^* equals τ.

find an approximate value for $r^*(n,\tau)$ by solving this equation numerically for each desired values of n and τ: see Fig. 6 for the behaviour of r^* with n.

Then, our objective is to estimate the following quantity:

$$\frac{\Delta\hat{d}}{n} := \frac{1}{n}\left(\hat{d}(|x| = r^*(n,\tau), q) - \hat{d}(|x| = 0, q)\right) . \tag{32}$$

This last quantity is representative of the typical variations of fractal dimension estimate that would be observed roughly 1-τ times on average. These variations are only due to changes in probability density, and they do not not represent variations in fractal properties. Since there is no analytical expression for r^*, we cannot give an explicit expression for $\frac{\Delta\hat{d}}{n}$ using our approximation (30), however we can plot it numerically. This is shown in Fig. 7. Our approximation (30) predicts that $n \mapsto \frac{\Delta\hat{d}}{n}$ is a growing function of n for $\tau = 0.9$ and values of q below 0.001, while it reaches a maximum for moderate values of n if $\tau = 0.9$ and $q = 0.01$, or if $\tau = 0.99$ and for all considered values of q. This maximum value of $n \mapsto \frac{\Delta\hat{d}}{n}$ depends on q and τ, as well as the value of n for which the maximum is reached. Fig. 7 indicates very strong values for $\Delta\hat{d}$, up to 4 times the exact phase-space dimension n. On the one hand, noting the discrepancy between our approximation and numerical experiments from

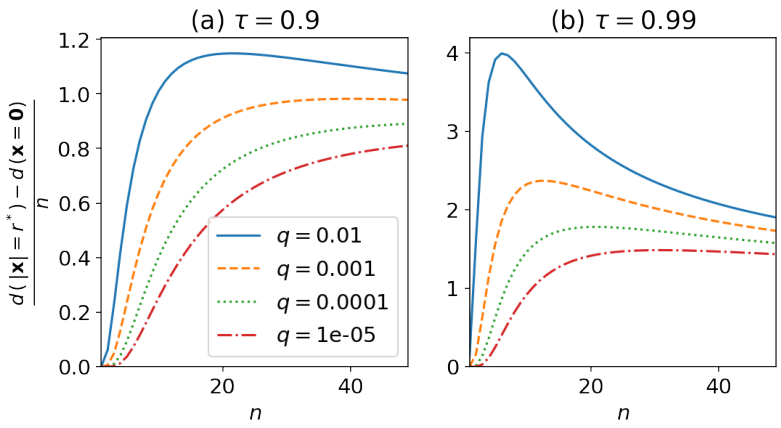

Fig. 7 Typical variations of dimension estimates for the multivariate Gaussian, as predicted from our approximation Eq. (30), as a function of phase-space dimension n, and for various values of the ratio q of total data used to compute the local dimensions. (a) Probability $\tau = 0.9$ of being in a centered ball of radius r^*. (b) Same with $\tau = 0.99$.

Fig. 5, we expect these numbers to greatly overestimate the true value of $\frac{\Delta\hat{d}}{n}$. On the other hand, the good agreement shown in Fig. 5 between experiments and analytical approximation in terms of behaviour with n and q suggests that the same kind of qualitative agreement could be found for $\frac{\Delta\hat{d}}{n}$.

To test the validity of these approximations, numerical experiments are again performed. This time, we use two values for $q = 10^{-3}, 10^{-4}$; and ten values for $n = 2, 5, 8, \ldots, 29$; and finally two values for $\tau = 0.9, 0.99$. As in the previous experiment, for each pair (q, n) 100 independent datasets are generated, each containing $10^3/q$ samples of the standard multivariate Gaussian of exact dimension n. For each dataset, Eq. (3) is used to estimate the dimension at $x = (0, 0, \ldots, 0)$ and at $x = (r^*, 0, \ldots, 0)$. For each triplet (τ, q, n), 100 values are therefore obtained for $\frac{\Delta\hat{d}}{n}$. Taking the average over all 100 realisations, we obtain the empty circles and full stars shown in Fig. 8, and compared against the semi-analytical curves of the previous figure.

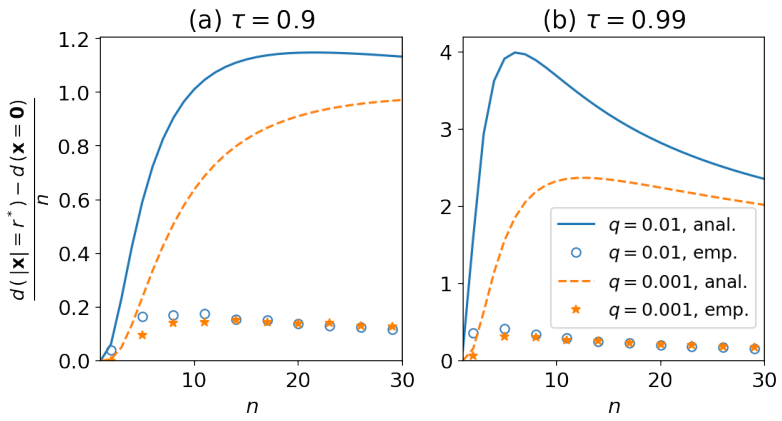

Fig. 8 Estimated variations of local dimension $\frac{\Delta \hat{d}}{n}$ for the multivariate normal distribution, between position $x = (0, 0, \ldots, 0)$, and $x = (r^*, 0, \ldots, 0)$, where r^* is such that the probability to be within distance r^* from $x = 0$ is $\tau = 0.9$, for two values of q, the proportion of data used to compute the local dimensions, and 10 values of n, the exact dimension. Circles and stars: values obtained from numerical experiments, averaged over 100 realisations for each triplet (τ, q, n). Full lines: approximation from Eq. (30), exactly as in Fig. 7.

The comparison with the approximation from (30) confirms that the latter strongly overestimates the amplitude of these variations, by approximately one order of magnitude. However, as in the previous experiment, the behaviour is quite the same between our theoretical approximation and the numerical experiments. For $n = 2$, $\frac{\Delta \hat{d}}{n}$ is negligible. The numerical estimates of $n \mapsto \frac{\Delta \hat{d}}{n}$ at fixed (q, τ) seem to reach a maximum between $n = 5$ and $n = 8$ for $q = 10^{-2}$, while the maximum is reached for slightly higher values of n in the case $q = 10^{-3}$. The empirical values of the maxima are ~ 0.15 for $\tau = 0.9$ (one event out of 10) and between ~ 0.3 and ~ 0.4 for $\tau = 0.99$ (one event out of 100) depending on the value of q. Although our approximation overestimates these effects, they are still non-negligible in practice. According to this experiment, for a system of dimension $n = 14$, and using 1 millionth of the data to compute local fractal dimensions, the difference in estimated dimension between the most probable position ($x = 0$) and a position which is visited one time out of 10 (respectively 100) is of the order of 15% (respectively 30%) of the true dimension, that is ~ 2 (respectively ~ 4).

To recall, a dimension of 14 is typical of continental-scale atmospheric circulation systems [6]. This experiment suggests that variations of local dimension estimates of the order of 2-4 might be due to changes in local density, and not to true changes in the fractal dimension. Note that such an amplitude of local variations of dimension is usually interpreted as variations in the local fractal nature of the attractor [6]. The results shown here suggest that these variations may be more difficult to interpret, possibly embedded with changes associated to uneven sampling of the phase-space caused by local changes in density.

4 Conclusion and perspectives

Approximate analytical expressions have been derived to anticipate the variations of local dimension estimate of random variables possessing an absolutely continuous measure (i.e., a continuous probability density function) without zeros or singularities. Such variables should not display variations of the local dimension according to the multi-fractal formalism of dynamical systems. These variations are therefore not related to the local fractal properties of the attractor. Rather, they are consequences of uneven sampling of the phase-space due to local changes in density of the underlying system. The derived approximate analytical expressions are compared to numerical experiments, proving relevant for a one-dimensional double-well stochastic system, a two-dimensional Gaussian Mixture Model, and finally standard multivariate Gaussian random variables. Although the given approximations overestimate these anomalous variations, good qualitative agreements are found between the behaviour expected from our approximations and that observed in the numerical simulations.

The issue tackled in this work is related to that of [20], who showed that the attractor dimension, obtained by averaging local dimensions on the attractor estimated as in Eq. (3), differs from the true phase-space dimension for random variables with absolutely continuous measures. Here, we focused not on the average of the local

dimension but on the variations of this local dimension. [20] showed that the deviation from the true phase-space dimension n of the averaged local dimension is strongest in high-dimension and with low values of q, the proportion of data used to compute the local dimension. Here, studying the *relative* variations in phase-space of local dimension for a multivariate Gaussian (see Eq. 30), we find similar results for the dependency with q. However, since we focus on the relative variations of dimension estimates, we expect the variations of local dimension to be strongest for moderate values of the phase-space dimension n, around $n \sim 11$ (see empirical values of Fig. 8). For atmospheric circulation data with typical local dimensions between 8 and 13 [6], our results suggest that such density-related effects could, *in principle*, be the prominent drivers of dimension estimate variability for these studies.

Furthermore, tests on simulated Gaussian Mixture Model data also suggest that the effect of lower dimension around regime peaks (as observed by [8] and [25]), and higher dimension around transitions between regimes (observed by [11]), is also obtained for purely random systems that should not, in principle, exhibit local variations of the local attractor dimension. However, note that this is only true if the regime peak happens close to the center of the regimes. On the contrary, [25] showed that the effect of lowered dimension around regime peaks is strongest for high value of the peak weather-regime index, *i.e.* far from the regime centers, where the density of data is low. According to our work, such a behaviour is not expected for random variables with absolutely continuous measures, because the latter would witness an increase of dimension far from regime centers due to the lower data density. This last fact strengthens, on the contrary, the idea that the observed diminution of local dimension around peak weather regime index is dominated by effects of change in the multi-fractal nature of the attractor, rather then the density-based effects studied here.

These elements suggest that more investigations are needed to establish the relevance of these results to real atmospheric circulation from realistic model simulations

and observations. Taking these inquiries further would allow to assess the relative importance of two concurring views of weather regimes: the statistics-based description which views atmospheric circulation as a random system subject to fluctuations between different metastable states, and the dynamical systems-based description where local variations in the fractal properties of the attractor drive the dynamics of the system.

More broadly speaking, this study suggests that at least a part of the variability of dimension fluctuations is due to changes in density, and not solely changes in fractal properties. Being able to discriminate the part of dimension variability related to each of these two sources would allow one to interpret better the notion of dimensionality from such estimates. In particular, with the objective of building a low-order model, one would be interested in knowing if the largest values of estimated dimension are due to changes in fractal properties (in which case a large number of variables would be needed in a low-order model) or to changes in density (in which case one could rely on a number of variables lower than the largest estimated dimension). Again, further developments are needed in order to separate these two sources.

Acknowledgments. This work was financially supported by the ERC project 856408-STUOD. We are thankful to Théophile Caby for fruitful discussions on this work.

Competing interests. The authors declare they have no competing interests.

Code availability. The code used to generate data and produce figures for this article is accessible upon request.

Author's contribution. Manuscript first writing, equations derivation, methodology development, numerical experiments: PP. Scientific guidance, manuscript modifications and final approval: PP, BC. Funding acquisition: BC.

References

[1] Grassberger, P., Procaccia, I.: Characterization of strange attractors. Physical review letters **50**(5), 346 (1983)

[2] Kantz, H., Schreiber, T.: Nonlinear Time Series Analysis vol. 7. Cambridge university press, ??? (2004)

[3] Badii, R., Broggi, G.: Measurement of the dimension spectrum f (α): Fixed-mass approach. Physics Letters A **131**(6), 339–343 (1988)

[4] Faranda, D., Masato, G., Moloney, N., Sato, Y., Daviaud, F., Dubrulle, B., Yiou, P.: The switching between zonal and blocked mid-latitude atmospheric circulation: a dynamical system perspective. Climate Dynamics **47**, 1587–1599 (2016)

[5] Messori, G., Caballero, R., Faranda, D.: A dynamical systems approach to studying midlatitude weather extremes. Geophysical Research Letters **44**(7), 3346–3354 (2017)

[6] Faranda, D., Messori, G., Yiou, P.: Dynamical proxies of north atlantic predictability and extremes. Scientific reports **7**(1), 1–10 (2017)

[7] Faranda, D., Messori, G., Alvarez-Castro, M.C., Yiou, P.: Dynamical properties and extremes of northern hemisphere climate fields over the past 60 years. Nonlinear Processes in Geophysics **24**(4), 713–725 (2017)

[8] Hochman, A., Messori, G., Quinting, J.F., Pinto, J.G., Grams, C.M.: Do atlantic-european weather regimes physically exist? Geophysical Research Letters **48**(20), 2021–095574 (2021)

[9] Nabizadeh, E., Lubis, S.W., Hassanzadeh, P.: The summertime pacific-north

american weather regimes and their predictability. Geophysical Research Letters **49**(16), 2022–099401 (2022)

[10] Faranda, D., Messori, G., Yiou, P., Thao, S., Pons, F., Dubrulle, B.: Dynamical footprints of hurricanes in the tropical dynamics. Chaos: An Interdisciplinary Journal of Nonlinear Science **33**(1) (2023)

[11] Platzer, P., Chapron, B., Tandeo, P.: Dynamical properties of weather regime transitions. Stochastic Transport in Upper Ocean Dynamics, 223 (2023)

[12] Holmberg, E., Messori, G., Caballero, R., Faranda, D.: The link between european warm-temperature extremes and atmospheric persistence. Earth System Dynamics **14**(4), 737–765 (2023)

[13] Lucarini, V., Faranda, D., Freitas, J.M.M., Holland, M., Kuna, T., Nicol, M., Todd, M., Vaienti, S., *et al.*: Extremes and Recurrence in Dynamical Systems. John Wiley & Sons, ??? (2016)

[14] Lorenz, E.N.: Atmospheric predictability as revealed by naturally occurring analogues. Journal of Atmospheric Sciences **26**(4), 636–646 (1969)

[15] Yiou, P.: Anawege: a weather generator based on analogues of atmospheric circulation. Geoscientific Model Development **7**(2), 531–543 (2014)

[16] Tandeo, P., Ailliot, P., Ruiz, J., Hannart, A., Chapron, B., Cuzol, A., Monbet, V., Easton, R., Fablet, R.: Combining analog method and ensemble data assimilation: application to the lorenz-63 chaotic system. In: Machine Learning and Data Mining Approaches to Climate Science: Proceedings of the 4th International Workshop on Climate Informatics, pp. 3–12 (2015). Springer

[17] Platzer, P., Yiou, P., Naveau, P., Tandeo, P., Filipot, J.-F., Ailliot, P., Zhen, Y.:

Using local dynamics to explain analog forecasting of chaotic systems. Journal of the Atmospheric Sciences **78**(7), 2117–2133 (2021)

[18] Platzer, P., Yiou, P., Naveau, P., Filipot, J.-F., Thiébaut, M., Tandeo, P.: Probability distributions for analog-to-target distances. Journal of the Atmospheric Sciences **78**(10), 3317–3335 (2021)

[19] Caby, T., Faranda, D., Mantica, G., Vaienti, S., Yiou, P.: Generalized dimensions, large deviations and the distribution of rare events. Physica D: Nonlinear Phenomena **400**, 132143 (2019)

[20] Pons, F.M.E., Messori, G., Alvarez-Castro, M.C., Faranda, D.: Sampling hyperspheres via extreme value theory: implications for measuring attractor dimensions. Journal of statistical physics **179**(5-6), 1698–1717 (2020)

[21] Lorenz, E.N.: Deterministic nonperiodic flow. Journal of the atmospheric sciences **20**(2), 130–141 (1963)

[22] Young, L.-S.: Dimension, entropy and lyapunov exponents. Ergodic theory and dynamical systems **2**(1), 109–124 (1982)

[23] Caby, T., Gianfelice, M., Saussol, B., Vaienti, S.: Topological synchronisation or a simple attractor? Nonlinearity **36**(7), 3603 (2023)

[24] Kloeden, P.E., Platen, E., Kloeden, P.E., Platen, E.: Stochastic Differential Equations. Springer, ??? (1992)

[25] Lee, S.H., Messori, G.: The dynamical footprint of year-round north american weather regimes. Geophysical Research Letters **51**(2), 2023–107161 (2024)

[26] Reynolds, D.A., et al.: Gaussian mixture models. Encyclopedia of biometrics **741**(659-663) (2009)

[27] Hartigan, J.A., Wong, M.A., *et al.*: A k-means clustering algorithm. Applied statistics **28**(1), 100–108 (1979)

[28] Kondrashov, D., Ide, K., Ghil, M.: Weather regimes and preferred transition paths in a three-level quasigeostrophic model. Journal of the atmospheric sciences **61**(5), 568–587 (2004)