Probability Distributions for Analog-To-Target Distances

P. PLATZER,[a,b,c] P. YIOU,[a] P. NAVEAU,[a] J.-F. FILIPOT,[c] M. THIÉBAUT,[c] AND P. TANDEO[b]

[a] *Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212 CNRS-CEA-UVSQ, Institut Pierre-Simon Laplace and Université Paris-Saclay, Gif-sur-Yvette, France*
[b] *Lab-STICC, UMR CNRS 6285, IMT Atlantique, Plouzané, France*
[c] *France Énergies Marines, Plouzané, France*

(Manuscript received 19 December 2020, in final form 27 July 2021)

ABSTRACT: Some properties of chaotic dynamical systems can be probed through features of recurrences, also called analogs. In practice, analogs are nearest neighbors of the state of a system, taken from a large database called the catalog. Analogs have been used in many atmospheric applications including forecasts, downscaling, predictability estimation, and attribution of extreme events. The distances of the analogs to the target state usually condition the performances of analog applications. These distances can be viewed as random variables, and their probability distributions can be related to the catalog size and properties of the system at stake. A few studies have focused on the first moments of return-time statistics for the closest analog, fixing an objective of maximum distance from this analog to the target state. However, for practical use and to reduce estimation variance, applications usually require not just one but many analogs. In this paper, we evaluate from a theoretical standpoint and with numerical experiments the probability distributions of the K shortest analog-to-target distances. We show that dimensionality plays a role on the size of the catalog needed to find good analogs and also on the relative means and variances of the K closest analogs. Our results are based on recently developed tools from dynamical systems theory. These findings are illustrated with numerical simulations of well-known chaotic dynamical systems and on 10-m wind reanalysis data in northwest France. Practical applications of our derivations are shown for forecasts of an idealized chaotic dynamical system and for objective-based dimension reduction using the 10-m wind reanalysis data.

KEYWORDS: Atmosphere; Statistics; Data science; Other artificial intelligence/machine learning

1. Introduction

Atmospheric analogs have been introduced by Lorenz (1969) in a study on atmospheric predictability. The faster one target state z and its closest analog a_1 diverge from one another, the harder it is to predict the evolution of z. In Lorenz's study, the state z was characterized by height values of the 200-, 500-, and 850-hPa isobaric surfaces at a grid of ≈1000 points over the Northern Hemisphere. The database of available analogs, called the catalog, contained five years of twice-daily values. In his abstract, Lorenz states that there are ''numerous mediocre analogues but no truly good ones.''

Since Lorenz's work, analogs have been used in many applications such as weather generators (Yiou 2014), data assimilation (Hamilton et al. 2016; Lguensat et al. 2017), kernel forecasting (Alexander et al. 2017), downscaling (Wetterhall et al. 2005), nonlinear bias correction (Hamill et al. 2015), climate reconstruction (Schenk and Zorita 2012; Fettweis et al. 2013; Yiou et al. 2013), and extreme event attribution (Cattiaux et al. 2010; Jézéquel et al. 2018).

The reason why Lorenz could not find any good analog was made clear later on by Van Den Dool (1994). It was shown that for high-dimensional systems, the mean return time of a good analog (used as a proxy for a minimum catalog size) grows exponentially with dimension. This result is a variant for analogs of the ''curse of dimensionality,'' well known in data sciences. With three pressure levels over the whole Northern Hemisphere, the dimension of Lorenz's study was very high, and only 5 years of twice-daily data was not enough to hope finding a good analog.

Nicolis (1998) added a dynamical systems' perspective to Van Den Dool's analysis. She showed that studying mean return times was not enough, as the relative standard deviation of this return time could be very high. Furthermore, it was shown that return-time statistics exhibit strong local variations in phase-space, so that certain target states may need a larger catalog size to find good analogs.

Accounting for Van Den Dool's findings, it is now usual to reduce as much as possible the feature-space dimension before searching for analogs. Also, the last decades have witnessed a proliferation of data from in situ and satellite observations, as well as outputs from numerical physics-based model. Such conditions allow one to find good analogs in many situations, and it has become standard to use not just one, but many analogs (usually a few tens). From a statistical perspective, using many analogs instead of one can increase estimation bias, but it reduces estimation variance, so that the estimation is less sensitive to noise. Using many analogs also allows us to perform local regression techniques on the analogs, such as local linear regression (Lguensat et al. 2017). This technique has proven efficient in analog forecasting applications (Ayet and Tandeo 2018), and it was shown that local linear regression allows analog forecasting to capture the local Jacobian of the dynamics of the real system (Platzer et al. 2021).

Corresponding author: Paul Platzer, paul.platzer@ifremer.fr

DOI: 10.1175/JAS-D-20-0382.1

This new context suggests focusing not only on the closest analog a_1, but also the kth closest analog, for k up to ~40. The number of analog used is usually the result of a trade-off between the number of available good analogs and the minimum number of analogs required to perform a given task (for instance, Yiou and Déandréis 2019; search for 20 analogs at each step to perform ensemble analog forecasts). Also, one can now reasonably hope to find good analogs using dimension reduction and a large amount of data. Thus, one is less interested in return times, but rather in analog distances. That is, for a given length of available data, how far will the closest analogs be? Performances of analog-based methods are usually conditioned by analog-to-target distances [see, for instance, the relationship between analog distances and forecast performance in Farmer and Sidorowichl (1988) and Platzer et al. (2021)]. In this work, we propose to evaluate the probability distribution of these distances. Our analytical probability distributions make the link between analog-to-target distances, catalog size, and local dimension. This brings new insight on the impact of dimensionality on analog methods.

Section 2 outlines the theoretical framework and findings. The section 3 shows implications of the findings and compares the present analysis with past studies. Section 4 shows results from numerical experiments of the three-variable Lorenz (1963) system, the variable-dimension Lorenz (1996) system, and from 10-m wind reanalysis data from the regional climate model AROME, further referred to as "the AROME reanalysis data." Detailed derivations of the results of section 2 can be found in appendixes B and C.

2. Theory

a. Analogs in dynamical systems and local dimensions

We assume a dynamical system with an attractor set $\mathcal{A}$, so that (almost) all trajectories in the basin of attraction of $\mathcal{A}$ converge to the attractor (Milnor 1985). For such systems, almost all trajectories starting from the attractor come back infinitely close to their initial condition after a sufficiently long time (Poincaré 1890). Analog methods are based on the idea that if one is provided with a long enough trajectory of the system of interest, one will find analog states close to any point z of the attractor $\mathcal{A}$.

The trajectory from which the analogs are taken is called the "catalog" $\mathcal{C}$ and can either come from numerical model output or reprocessed observational data. It can be seen either as a trajectory from a discrete dynamical system or as evenly spaced time samples from a continuous dynamical system. In any case, the catalog has a finite number of elements noted $L := \mathrm{card}(\mathcal{C})$. This catalog size may be divided by a typical correlation time scale so that elements of the catalog can be considered independent (Van Den Dool 1994). In fact, for the analogs of a given target z to be considered independent, it is enough that the maximum distance between any two analogs of z be smaller than the minimum distance between any analog of z and its neighbors in time (i.e., its successor and predecessor).

The structure of the attractor, expressed by the system's invariant measure μ, conditions the structure of the catalog and the ability to find analogs. In particular, Van Den Dool (1994) and Nicolis (1998) studied the role of the attractor's dimension that we will now introduce. Let $B_{z,r}$ the ball centered on $z \in \mathcal{A}$ and of radius r, then

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

defines the finite-resolution (r-resolution) local dimension at point z. As mentioned later in the text, this definition depends on the unit used to measure the distance r, although only lightly if r is small [see Eq. (13)]. For instance, relative temperature differences are higher if measured in degrees Fahrenheit rather than degrees Celsius, resulting in a smaller value of $d_{z,r}$ at fixed r. However, in practice we estimate here $d_{z,r}$ based on Eq. (2) which considers ratios of distances and is therefore unit independent. There are many other ways to estimate dimension, including ones that do not depend on the choice of unit [see, for instance, the more global estimates of Wang and Shen (1999)]; however, Eq. (1) is the most suited to our purpose and derivations, as appears clearly in appendix B.

Note that for ergodic measures, $\mu(B_{z,r})$ can be approximated by counting the number of times a given trajectory enters $B_{z,r}$ [this is the consequence of the ergodic theorem of Birkhoff (1931)]. In the following, we assume that μ is ergodic and stationary. This does not apply when nonstationary processes, such as climate change, break the stationarity of μ. Also, in practice, periodic forcings such as seasonality make the structure of the attractor of a system such as the atmosphere vary between winter and summer. Therefore, analogs must be searched within a given time window around the calendar date of the target z, so that the subsampling allows us to recover an invariant measure (see Lorenz 1969; Yiou and Déandréis 2019). For a discussion on the modification of the invariant measure due to seasonality and nonperiodic forcing, see Robin et al. (2017).

Assuming that μ is ergodic and that $\lim_{r\to 0} d_{z,r}$ exists, μ is said to be exact dimensional and the limit is independent of z (Young 1982). This typical value of the local dimension is the order-one Renyi dimension, also called information dimension, or attractor dimension, and is here noted D_1. It is a typical value in the sense that for every z and for small enough r, $d_{z,r}$ is close to D_1. Also, D_1 can be estimated by taking the average of the estimates of local dimension (see next section):

$$D_1 := \lim_{r\to 0} d_{z,r}.$$

The finite-resolution local dimension $d_{z,r}$, however, can deviate from the typical value D_1. More precisely, $d_{z,r}$ exhibits large deviations from its limit value. The amplitude of these deviations depends on $(-\log r)^{-1/2}$ and on the spectrum of fractal dimensions (for more details, see Caby et al. 2019).

These definitions of dimension correspond to the notion of attractor dimension, which comes from the field of dynamical systems. There are strong connections with other mathematical objects used to estimate dimensionality in computer science and machine learning. These include the doubling dimension (Gupta et al. 2003) and expansion dimension (Karger and Ruhl 2002) which are related to ratios of volume occupied by data, and the intrinsic dimension (Houle 2013), which is related to

the minimum number of variables needed to correctly represent a dataset. The local intrinsic dimension as defined by Houle (2017) is closely related to the local attractor dimension $d_{z,r}$ which is used in the present study.

The definitions of $B_{z,r}$, $d_{z,r}$, and D_1 depend on the metric that is used to evaluate distances. However, we show in appendix A that the limit value D_1 is independent of the choice of metric; therefore, $d_{z,r}$ is also expected to depend only lightly on the metric that is used. The theoretical results expressed in this paper in the limit of small distance $r \to 0$ (or, equivalently, of large catalog $L \to +\infty$) are valid whatever the metric used. Note that this does not apply to measures of similarity such as correlation or statistical divergence, that are not actual metrics (of which we recall the definition in appendix A).

All these definitions are valid in the limit of small distance r, which can be hard to achieve in high dimension due to the concentration of norms or "curse of dimensionality" (Verleysen and François 2005). The effect of the curse of dimensionality on the estimation of dimensions following Eq. (1) was studied analytically and numerically by Pons et al. (2020), with effects starting to be nonnegligible in dimension ≈40. In the numerical experiments presented here, we have checked empirically that the concentration of norms was small enough.

The distance from the kth analog $a_k(z) \in \mathcal{C}$ to the target state z is noted $r_k(z) := \mathrm{dist}[a_k(z), z]$. To lighten notations, we will often make the z dependency implicit, writing simply a_k and r_k rather than $a_k(z)$ and $r_k(z)$. Analog-to-target distances always depend on a target z, and the only way to remove this dependency would be through averaging, which is done only in section 4e. Distances are sorted so that $r_1(z) < r_2(z) < \cdots < r_K(z)$, and K is the total number of analogs considered. Empirical methods usually set K to a fixed value, reaching for a bias-variance trade-off. A small value of K typically increases the variance of the analog method, for instance, in the case of observation noise. Raising the value of K allows us to average out this variability. However, a too large value of K would include analogs that are too far from the target and not relevant, therefore raising bias. For an example of this bias-variance trade-off, see Platzer et al. (2021). This amounts to looking at a lower quantile of the function $x \mapsto \mathrm{dist}(z, x)$. Another possibility is to set a threshold R for the analog-to-target distances so that $r_K(z) < R < r_{K+1}(z)$. In this case, $K(z)$ depends on z. This is referred to as the epsilon nearest neighbor search. However, in the numerical experiments of this paper we always set K to a fixed value.

b. *Simple scaling of analog-to-target distance with local dimension*

Using extreme value theory and dynamical systems theory, Caby et al. (2019) showed that $d_{z,r}$ can be estimated using the empirical cumulative distribution function (CDF) of points inside a ball of exponentially decreasing radius:

$$\overline{F}_z(s) = \frac{\mu(B_{z,r_K e^{-s}})}{\mu(B_{z,r_K})},$$

where s takes values according to the available data, that is, for the kth analog of z, $s_k = -\log(r_k/r_K)$, and $\overline{F}_z(s) = k/K$. This empirical distribution is thus the CDF of the K closest available analogs. It follows from Caby et al. (2019) that, for regular enough measures, $\overline{F}_z(s) \approx e^{-ds}$, where $d = d_{z,r_K}$. Therefore, an estimate of d_{z,r_K} is given by

$$d_{z,r_K} \approx \left\{ \sum_{k=2}^{K} (s_k - s_{k-1}) \overline{F}_z(s_k) \right\}^{-1} = \left\{ \sum_{k=2}^{K} \frac{k}{K} \log\left(\frac{r_{k-1}}{r_k}\right) \right\}^{-1}. \tag{2}$$

In the following and unless otherwise noted, "the local dimension," or d, both refer to d_{z,r_K}, which is estimated using the above formula. Exceptions will arise in appendix B where a formal proof is given and d might refer to $d_{z,r}$ as defined in Eq. (1).

A practical application of Eq. (2) with the system of Lorenz (1963) (see appendix D for a formal definition of this system) is given in Fig. 1. Another way to estimate d_{z,r_K} is not to use directly Eq. (2) but rather to make a least squares fit of the empirical CDF, $\overline{F}_z(s)$, assuming an exponential shape $\overline{F}_z(s) \approx e^{-s/\sigma}$ and returning the obtained value σ^{-1}. As can be seen in the example of Fig. 1, both methods give similar results.

Also following Caby et al. (2019), we can estimate the attractor dimension D_1 from the average of realizations of d_{z,r_K} inside the catalog:

$$D_1 \approx \frac{1}{L} \sum_{z \in \mathcal{C}} d_{z,r_K}, \tag{3}$$

where it is taken care of that the neighbors in time of $z \in \mathcal{C}$ are not included in the list of analogs $[a_k(z)]_{k=1,\ldots,K}$. Caby et al. (2019) use this approximation to estimate the attractor dimension of the system of Lorenz (1963, hereafter noted L63) as 2.06, which is in agreement with values found in the literature.

The approximation $\overline{F}_z(s) \approx e^{-ds}$ implies the scaling of $r_k(z)$ with k:

$$r_k(z) \sim k^{1/d}, \tag{4}$$

where again $d = d_{z,r_K}$ is the local dimension at finite resolution r_K.

Equation (4) already reveals an important point of our analysis, which is the scaling of r_k with k, and is approximately given by a power-law with exponent $1/d$. However, this formula comes from a work on local dimensions, not analog-to-target distances. It is therefore not surprising that some of the elements required for our study are missing. In particular, this scaling does not give the constant in front of $k^{1/d}$, in which resides the relation to the catalog size, a crucial point for analog applications. Also, it only gives a mean or typical value of $r_k(z)$, while our objective is to evaluate the probability distribution of $r_k(z)$ at fixed z and L, or at least the probability of departures from this mean scaling.

The next section gives the full probability distribution of $r_k(z)$ for a fixed target z as a function of the local dimension, the catalog size, and the analog number k.

c. *Full probability distribution of analog-to-target distance*

In appendix B we show the main result of this paper, which is that, assuming fixed and known values of L, k, and d_{z,r_K} the kth analog-to-target distance $r_k(z)$ follows the following probability density function:

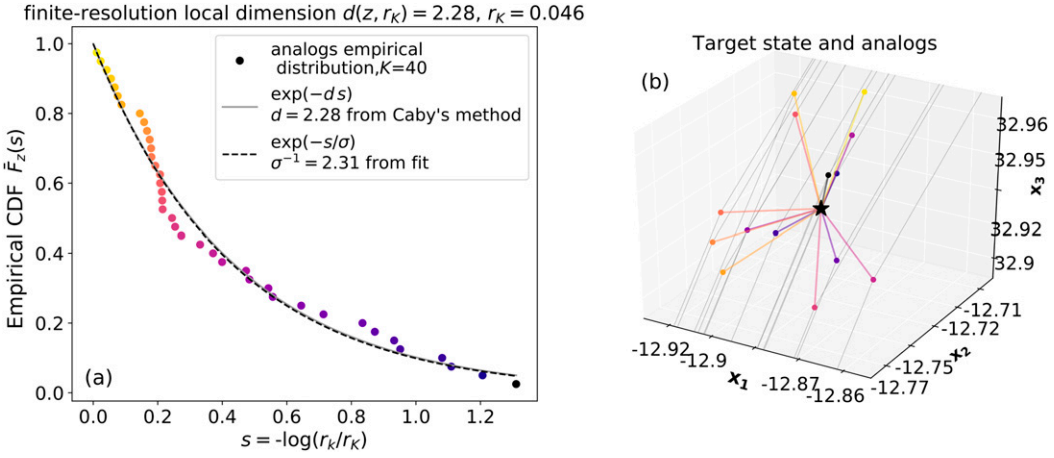

FIG. 1. Computing the finite-resolution local dimension $d = d_{z,r_K}$ at a point z of the three-variable L63 system, using $K = 40$ analogs. (a) Following from Caby et al. (2019), we evaluate d by taking the mean of the empirical CDF of analog distance in logarithmic scale. For this example, fitting the empirical CDF with an exponential $\exp(-s/\sigma)$ and taking the inverse of σ gives approximately the same value for d_{z,r_K}. (b) Target z (black star) and one in three analogs [colored dots matching (a)]. The trajectories from which the analogs are taken are in gray. In this example, the smallest analog-to-successor distance is much larger than the largest analog-to-target distance (the successors are not even visible in the figure).

$$p_k(r) = dLr^{d-1}\frac{(Lr^d)^{k-1}}{(k-1)!}e^{-Lr^d}, \tag{5}$$

where $p_k(r)$ is defined through $\mathbb{P}(r_k \in [r, r+\delta r)) = p_k(r)\delta r$, and the variables r_k and d both depend on z. Equation (5) was obtained neglecting the variations of $d_{z,r}$ with r, and therefore, in practice we assume that $d = d_{z,r_K}$ and that it is estimated from Eq. (2) with a fixed value of K. The value of d is thus assumed to be independent of k. This is consistent with practical applications where the limited available data do not allow us to evaluate fine variations of $d_{z,r}$ with r, but where clear variations of d_{z,r_K} with z are witnessed and reveal different dynamical situations (Faranda et al. 2017). Therefore, in this section d always refers to d_{z,r_K}. An alternative proof for Eq. (5) using K largest-order statistics from extreme value theory is given in appendix C.

Equation (5) then allows us to compute the mean and variance of r_k for fixed k, z, L, and d:

$$\langle r_k \rangle = \frac{\Gamma\left(k+\frac{1}{d}\right)}{L^{1/d}\Gamma(k)}, \tag{6a}$$

$$\langle r_k^2 \rangle - \langle r_k \rangle^2 = \frac{1}{L^{2/d}\Gamma(k)^2}\left\{\Gamma\left(k+\frac{2}{d}\right)\Gamma(k) - \Gamma\left(k+\frac{1}{d}\right)^2\right\}, \tag{6b}$$

where Γ is Euler's gamma function. These identities can be simplified through scalings of the gamma function $\Gamma(x+1) = \int_0^{+\infty} u^x e^{-u}\, du$ for large x, using Laplace's method up to second order to evaluate the integral (the first order gives Stirling's formula). This gives the following expressions for the mean and relative standard deviation:

$$\langle r_k \rangle \approx \left(\frac{k}{L}\right)^{1/d}, \tag{7a}$$

$$\frac{\left(\langle r_k^2 \rangle - \langle r_k \rangle^2\right)^{1/2}}{\langle r_k \rangle} \approx \frac{1}{dk^{1/2}}, \tag{7b}$$

where we recover the scaling $r_k \sim k^{1/d}$ of Eq. (4). These approximations are the result of Taylor expansions for large k from Eqs. (6a) and (6b), and will therefore be increasingly valid as k grows. However, even for $k = 1$, Eqs. (7a) and (7b) give a satisfactory numerical approximation of Eqs. (6a) and (6b).

If $kd > 1$, one can also compute r_k^*, the value of r for which p_k reaches a maximum:

$$r_k^* = \mathrm{argmax}_r[p_k(r)] = \left(\frac{k-\frac{1}{d}}{L}\right)^{1/d},$$

and when $kd \le 1$, $r_k^* = 0$ and $p_k(0) = +\infty$. Note that the three quantities $\langle r_k \rangle$, $(k/L)^{1/d}$ and r_k^* are equivalent as $k \to +\infty$.

Equation (5) calls for the rescaling of r_k by $L^{1/d}$, later on referred to as the catalog density. The probability distribution $\tilde{p}_k$ of the rescaled analog-to-target distance $u_k = L^{1/d} r_k$ can be computed by imposing the change of variable $\tilde{p}_k(u)du = p_k(r)dr$, giving

$$\tilde{p}_k(u) = du^{d-1}\frac{u^{d(k-1)}}{(k-1)!}e^{-u^d}, \tag{8}$$

which shows that after rescaling by the catalog density $L^{1/d}$, the probability density is independent of L.

Figure 2 shows plots of $\tilde{p}_k(u)$ against u for varying values of d and k. As a consequence of the scaling $u_k \sim k^{1/d}$, we observe large variations of $\langle u_k \rangle$ with k for small dimensions d, and very small variations of $\langle u_k \rangle$ with k for large dimensions d. Note that, in the limiting case $d \to \infty$, the random variables r_k are degenerate and all equal $L^{-1/d}$ almost surely. This can be witnessed through the different scales of the horizontal axis of

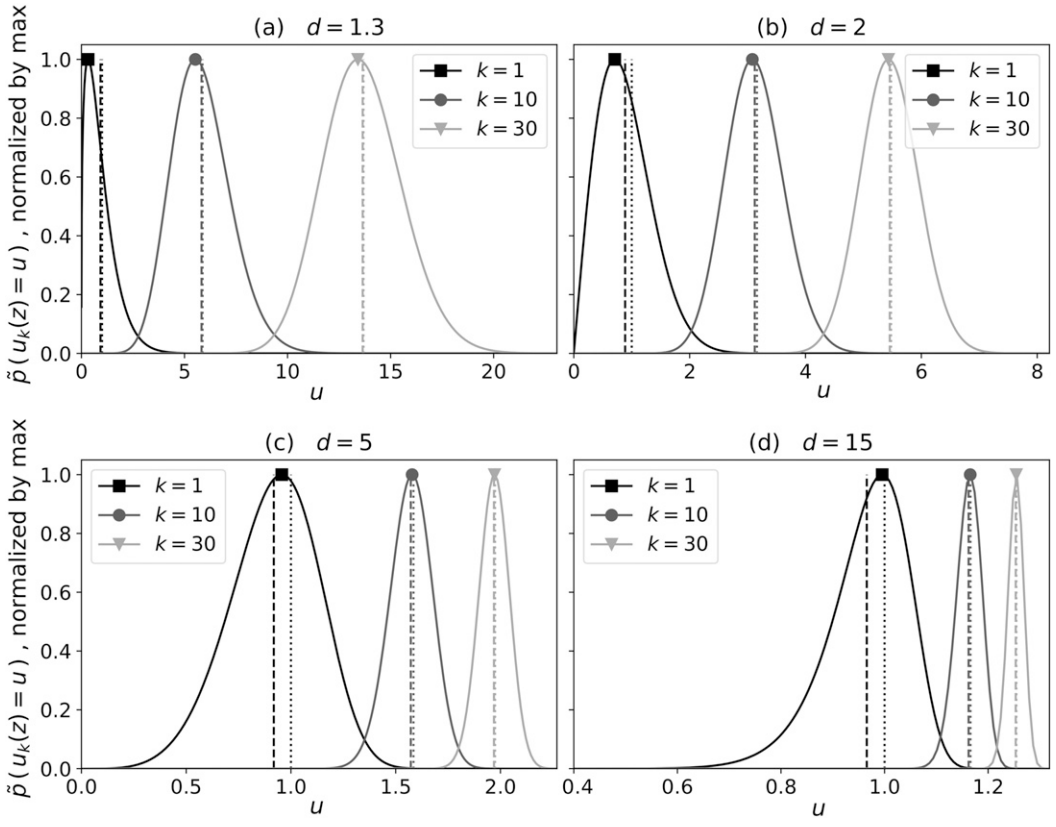

FIG. 2. Probability density functions of $u_k = L^{1/d} r_k$, the rescaled kth analog-to-target distance, for fixed values of k, and of the local dimension d, from Eq. (8). The dimension equals (a) 1.3, (b) 2, (c) 5, and (d) 15, and is assumed to be independent of k. All densities $\tilde{p}_k$ are normalized by their maximum value. Dashed vertical lines indicate the exact mean value $L^{1/d}\langle r_k \rangle$ from Eq. (6a), while dotted vertical lines indicate the approximate value $k^{1/d}$ from Eq. (7a). The argmax values of $\tilde{p}_1$, $\tilde{p}_{15}$, and $\tilde{p}_{30}$ are shown respectively with squares, circles, and triangles.

the plots. This result is the consequence of the contraction of norms in high dimension, which can cause the search for analogs to be meaningless. In particular, Beyer et al. (1999) showed that, under reasonable conditions, the ratio between the distance from a target state z to its nearest neighbor r_1 and the distance to the farthest point in a dataset r_L equals 1 for infinitely high-dimensional systems. Finally, it might seem counterintuitive that large values of the horizontal axis are observed in low dimension and not in high dimension, but this is only because the $L^{-1/d}$ factor was removed by rescaling. Figure 2 is still consistent with Eq. (7a) which shows that $\langle r_k \rangle$ is, at fixed L, a growing function of the local dimension d.

Also, as a consequence of Eqs. (7), we have that the standard deviation of r_k is a growing function of k for $d < 2$, while it is constant for $d = 2$ and decreasing for $d > 2$. However, the relative standard deviation of r_k is always a decreasing function of k and d according to Eq. (7b).

d. *Normalization and convergence to the standard normal distribution*

In this section, we go further from the rescaling u_k, and propose a normalization of the variable r_k (at fixed z) that depends on the local dimension $d = d_{z,r_K}$, on the value of $k < K$, and on the catalog size L. Equations (7a) and (7b) suggest the change of variables from r_k to v_k as

$$v_k := dk^{1/2}\left[\left(\frac{L}{k}\right)^{1/d} r_k - 1\right].$$

Then one can define the probability density function $h_k(v)$ of the normalized kth analog-to-target distance, so that $v = dk^{1/2}[(L/k)^{1/d}(r-1)]$ and $h_k(v)dv = p_k(r)dr$. This gives

$$h_k(v) = \frac{k^{k-1/2}}{(k-1)!}\left(1 + \frac{v}{dk^{1/2}}\right)^{dk-1} \exp\left[-k\left(1 + \frac{v}{dk^{1/2}}\right)^{d}\right], \quad (9)$$

and simple asymptotic analysis gives

$$\lim_{k\to+\infty} h_k(v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right),$$

which shows that the distribution of the normalized random variable v_k approaches the standard normal distribution for large k. Note, however, that this limit cannot be fully observed in practice, as the distribution of Eq. (5) is valid only in the limit of large catalog size and with $k \ll L$. In practice, as the convergence is in $k^{-1/2}$, the relative difference

between h_k and the standard normal distribution is of ≈15% for $k = 40$.

e. Distances in observation space

In practice, one is very rarely able to observe the full state z, but rather an observable $y = f(z)$ defined through a vector-valued function $f: \mathcal{A} \mapsto \mathbb{R}^n$. In this case, the appropriate measure on the space of observations is $\mu \circ f^{-1}$, where f^{-1} is the inverse image of f that acts on sets (not vectors), and can therefore be defined even when f is not invertible. This allows us to define an observation-based dimension:

$$d^f_{z,r} = \frac{\log \mu \circ f^{-1}(B_{f(z),r})}{\log r}.$$

The limit $\lim_{r \to 0} d^f_{z,r}$, when it exists, is a function of D_1 and of properties of f. For instance, if f is differentiable and its Jacobian matrix at z is of rank $m > 0$, then $\lim_{r \to 0} d^f_{z,r} = \min(m, D_1)$. Also, it is easy to find examples where f is quadratic, and its Jacobian matrix at z is zero, and therefore, $\lim_{r \to 0} d^f_{z,r} = D_1/2$. This shows that there are a variety of ways in which the observed dimension can be lower than the actual attractor dimension. For more details, see Caby et al. (2020).

However, if we keep the hypothesis that μ is ergodic and z is a nonperiodic point, we can conduct the same analysis as in appendix B but replacing μ by $\mu \circ f^{-1}$, z by $y = f(z)$, and $d_{z,r}$ by $d^f_{z,r}$. Adding the hypothesis that f is C^∞ and that $d^f_{z,r}$ exists and has a finite limit as $r \to 0$, we recover Eq. (5), only replacing d by d^f.

Therefore, the statistics of analog-to-target distances in observation space also follow Eq. (5), this time with a dimension that depends not only on the dynamical system, but also on properties of the observable.

3. Consequences for applications of analogs

a. Comparison with previous studies

The pioneering work of Van Den Dool (1994) focuses on the minimum length of catalog needed to have a 95% chance to find at least one analog with a distance below a low threshold ε. With our notations, this condition can be written

$$L|\mathbb{P}(r_1 < \varepsilon) > 0.95.$$

Van Den Dool (1994) uses a Gaussian approximation for the difference between two states, which is reasonable in high dimensions. Then $\mathbb{P}(r_1 < \varepsilon) = 1 - (1 - \alpha^{D_1})^L$, where α is the probability that the distance between two arbitrarily chosen states is less then ε and can be expressed as the integral of a Gaussian probability density function. For small ε, $\alpha = \mathcal{O}(\varepsilon)$ and $\alpha^{D_1} \ll 1$. This finally suggests

$$L > \frac{\log 0.05}{\log(1 - \alpha^{D_1})} \approx \frac{-\log 0.05}{\alpha^{D_1}}. \tag{10}$$

Similar results can be found from Eq. (5). Indeed, one has $\mathbb{P}(r_1 < \varepsilon) = \int_0^\varepsilon p_1(r)\, dr = 1 - [\exp(-\varepsilon^d)]^L$, so that $\alpha \approx \varepsilon$. Here, D_1 is replaced by the local finite-resolution dimension $d = d_{z,r_K}$. Thus, our analysis encompasses the one of Van Den Dool (1994).

Nicolis (1998) extended the work of Van Den Dool (1994). Interpreting Eq. (10) in terms of mean return times and using the formula from Kac (1959), she found an expression of mean return times using the identity $\mu_{z,r} \approx r^{D_1}$ and a mean velocity. This theoretical analysis includes neither variations in phase space of the return time, nor variability of the return time due to the variability of the catalog for fixed L. However, Nicolis (1998) performed empirical estimates of such variations of the return time, shading light on the pitfalls of an analysis limited to mean return times.

In the present paper, the point of view switches from statistics of return times to statistics of analog-to-target distance, and is extended to the K closest analogs rather than just the first one. The full probability distribution of Eq. (5) gives a detailed view of the variability of the process of searching for analogs.

Note that our work has many connections to the one of Houle (2017), who also studied probability distributions of distance functions. However, we are not aware of any published work giving probability distributions of analog distances such as in Eq. (5).

b. Searching for analogs: Consequences

The full probability distribution of Eq. (5) has many consequences for the practical search of analogs.

For very low-dimensional systems ($D_1 < 2$), the first analog-to-target distance has a lower variability than the next ones, so that a given value of r_1 will be more representative of the next values of r_1 than a given value of r_{10} would be of the next values of r_{10}. The inverse phenomenon happens for higher dimensional systems ($D_1 > 2$). This can be taken into account to evaluate the expected performances of analog methods.

Also, the scaling $r_k \sim k^{1/d}$ implies that the growth with k of the mean analog-to-target distance is much faster for low-dimensional systems ($D_1 \lesssim 2$), so that the thirtieth analog would be much farther from z than the first one. Again, this is consistent with the work of Beyer et al. (1999) on the concentration of norms in high dimensions. In low dimension, the sensitivity of analog-to-target distances to the choice of K (i.e., the number of analogs used) is thus higher than in high dimension. In practice, in the case of a sparse catalog (i.e., if the density of points $L^{1/d}$ is not large enough to ensure finding very close analogs), a low value of K might be preferred in order to avoid using analogs too far away from the target. Conversely, in high dimension and with a similar density $L^{1/d}$, using a small or a large number of analogs should not play an important role on analog-to-target distances. However, note that the most important factor driving analog-to-target distances remains the catalog density $L^{1/d}$, which is higher in low dimension if the catalog size L is fixed. Therefore, our analysis is still consistent with the fact that, for a given size of dataset, better analogs will be found in low attractor dimension than in high attractor dimension. The higher sensitivity of analog-to-target distances to K in low dimension is only true if $L^{1/d}$ is fixed, which means that we are comparing the case of a low dimension d and a small catalog size L to the case of a high dimension and a large catalog size.

For instance, Lguensat et al. (2017) use analogs to produce forecasts of several well-known dynamical systems, setting

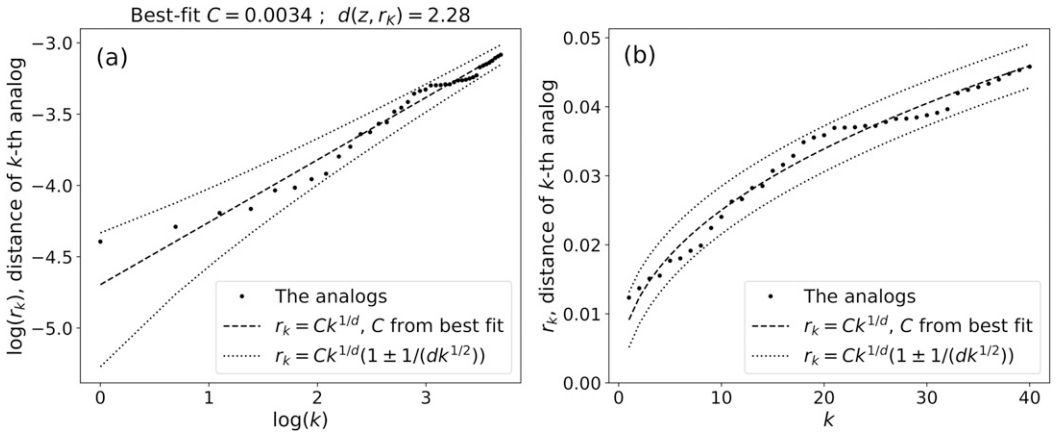

FIG. 3. Analog-to-target distance r_k, against analog number k at the same point z as in Fig. 1. (a) Log scale and (b) linear scale. Full circles are the empirical points given by the analogs. The dashed dark line is the best fit from Eq. (11), where d is fixed (from Caby's method) and C is estimated with least squares in log scale. Assuming that this fit gives an estimation of the mean, the dotted lines represent approximate standard deviation around this mean, assuming that the relative standard deviation is given by Eq. (7b).

$K = 40$, while the use of Gaussian kernels with a variable bandwidth equal to $\lambda_z = \text{median}_k r_k$ allows us to give a very low weight to analogs at distance $r_k > \lambda_z$. One might think that the filtering out of analogs with $r_k > \lambda_z$ makes the forecast procedure relatively insensitive to the choice of K. Conversely, assuming that $\lambda_z \approx \langle r_{[K/2]} \rangle$, where $[K/2]$ is the integer part of $K/2$, we have that λ_z grows with K as $\lambda_z \sim K^{1/d}$. Thus, for low-dimensional systems such as the one of L63 for which $D_1 \approx 2.06$, our results suggest that in the case of a low sampling density, high values of K might have detrimental effects on the efficiency of analog methods. This affirmation is tested in section 4b.

However, note that here we focus on analog-to-target distances assuming that they are an important driving factor of the efficiency of analog methods, but in practice many other parameters come into play, such as the choice of the proper metric, or the choice of the feature space. The tuning of analog methods does not reduce to the objective of minimizing analog-to-target distances. Nevertheless, our results can be used, with caution, to indicate tendencies and general behaviors of analog methods.

In particular, the scaling $\langle r_k \rangle \sim (k/L)^{1/d}$ can be used in the context of dimension reduction. Assume that one wants to perform a statistical task that necessitates K analogs (for instance, an ensemble forecast). Then assume that one wants to reduce the dimension in order to have $\langle r_K \rangle < \varepsilon$. From the scaling $\langle r_k \rangle \sim (k/L)^{1/d}$, we find that the dimension must be reduced to at least $d_{\max,K} = \{1 - [\log(K)]/[\log(L)]\} d_{\max,1}$. Detailed arguments and a practical example are given in section 4e. Thus, for instance, if the criterion $\langle r_1 \rangle < \varepsilon$ is met for $d_{\max,1} = 10$ and if $L = 10^4$, then the criterion $\langle r_{25} \rangle < \varepsilon$ will be met only for $d_{\max,25} = 6$. This shows that any dimension reduction performed with the objective of decreasing analog distances strongly depends on how many analogs are required.

Finally, the joint distribution of analog-to-target distances from appendix C theoretically allows us to express the probability distributions of any random variable of the form $\sum_k \omega_k r_k^p$, where $(\omega_k)_k$ are weights and p is a positive integer. Such quantities can give error bounds for analog methods [see Platzer et al. (2021) for the case of analog forecasting]. However, a closed form for the distribution of such variables is yet to be derived.

4. Numerical experiments

a. *Three-variable Lorenz system*

Using the procedure of Caby et al. (2019), one estimates the local finite-resolution dimension $d = d_{z,r_K}$ for any point z using the K closest analogs in the system of L63. This procedure is illustrated in Fig. 1. Then the scaling of Eq. (7a) is used to make a least squares fit from the data:

$$r_k(z) \overset{\text{LS}}{\approx} C(z) k^{1/d}, \tag{11}$$

where $r_k(z)$ is the observed kth analog-to-target distance and $\overset{\text{LS}}{\approx}$ means that the constant $C(z)$ is evaluated with least squares from Eq. (11). Figure 3 shows an application of this procedure for a given z of the L63, plotting the real values of $r_k(z)$, and using $C(z)k^{1/d}$ as an approximation for $\langle r_k(z) \rangle$ and dotted lines show the standard deviation around the mean from the approximate relative standard deviation given in Eq. (7b).

From Eqs. (11) and (7) one expects to find

$$C(z) \approx L^{-1/d}; \tag{12}$$

however, as L takes large values (from 10^5 to 10^7 or more), a small estimation error for d results in a large estimation error for $L^{-1/d}$. Another way to look at this estimation issue is that d is relatively insensitive to a rescaling of distances or a change of unit. Let

$$d'_{z,r} = \frac{\log \mu_{z,r}}{\log(r/\rho)}, \tag{13}$$

where ρ is a scalar value and r/ρ is a rescaled version of r, or equivalently r expressed in a different unit system. Note that we use $\mu_{z,r}$ and not $\mu_{z,r/\rho}$ as we only changed the unit of r, not the actual distance it represents. Then $d'_{z,r} \sim d_{z,r}$ as long as $|\log\rho| \ll |\log r|$. In particular, the method of Caby et al. (2019) is insensitive to a change of unit, as it involves only ratios of distances [see Eq. (2)]. Thus, Eq. (12) does not hold when C and d are determined as explained above. This is why $C(z)$ is rather evaluated through Eq. (11), which allows one to find the scaling factor $\rho(z)$ defined through

$$C(z) = \frac{\rho(z)}{L^{1/d}}. \tag{14}$$

Note that similar issues are raised by Faranda et al. (2011) regarding the continuity of $\mu_{z,r}$ with respect to r and its limiting behavior for small r, which motivates Lucarini et al. (2014) to postulate that $\mu_{z,r}$ is the product of r^{D_1} and a slowly varying function of r, which is in some sense equivalent to our hypothesis that $C(z)$ has to be rescaled with $\rho(z)$ when the local dimension is estimated from the method of Caby et al. (2019).

The fact that $\rho(z)$ varies with z (and is thus not exactly a change of unit) can be explained by the possibility for two points z_1 and z_2 to have the same local dimension $d_{z_1,r_K} = d_{z_2,r_K}$, but not to be visited at the same frequency by the system. A simple example of such a situation is any nonuniform, one-dimensional, continuous random variable. For such a variable Z, there exists values z_1 and z_2 such that the probability for Z to lie in the vicinity of z_1 is higher than in the vicinity of z_2, and yet $d_{z_1,r_K} = d_{z_2,r_K} = 1$.

Equations (11) and (14) are tested in numerical experiments using the system of L63, with results reported in Fig. 4. Analogs of a fixed target point z are sought for in 3×600 independent catalogs, with three different catalog sizes. Each catalog is built from a random draw without replacement of L points inside a (common) trajectory of 10^9 points, generated using a Runge–Kutta numerical scheme with a time step of 0.01 in usual nondimensional notations. The dimension is calculated using $K = 150$ points, where this number is justified by a bias-variance trade-off: using this number and testing the procedure on 100 points picked from the measure μ, one finds a mean dimension D_1 from Eq. (3) between 2.03 and 2.04, which is coherent with values reported by Caby et al. (2019), and a standard deviation of ~0.26. Using a lower value of K results in a higher variance, and using higher values results in biases that are dependent on the value of L used in this study. For more details on the distribution of local dimensions in the system of L63 the reader is referred to Faranda et al. (2017).

The consistency of empirical densities of ρ across varying values of L validates the scaling of C with L and d. Empirical probability densities of rescaled analog-to-target distances, also consistent across varying catalog sizes, are coherent with the theoretical probability densities from Eq. (5). The values of the rescaling parameter ρ are not surprising, as typical values of distances between points in the attractor are ~16 and maximum distances are ~28. Note that Nicolis (1998) uses a rescaling in studying analog return times with Lorenz's three-variable system, dividing all distances by the maximum distance between two points on the attractor. The fact that $\rho(z)$ exhibits seemingly large values is only the result of the choice of variables in the system of L63. For instance, it is possible to make a change of variables that would result in a system having the same chaotic properties, the same dimension, defined by almost the same dynamical equations, but with variables spanning smaller ranges, which would give numerical values of $\rho(z)$ close to 1 (see appendix D).

Repeating this experiment for different target points z gives similar results. Values of ρ are on the same order of magnitude as the ones reported in Fig. 4. The consistency across varying values of L is almost always recovered, except for some points that have slightly higher dimensions $d \gtrsim 2.15$ (not shown). We expect this to come from a bad choice of K when estimating the dimension and the rescaling factor: the choice of $K = 150$ is relevant for most points, but should be adapted to the local dimension. Moreover, the use of other metrics (Manhattan, order-8 Minkowski, Chebyshev) has a very small influence on the results presented in Fig. 4.

Finally, we have conducted the same experiments but using observations of the first coordinate of the Lorenz system. The results are shown in Fig. 5. Again, the numerical data fit the theory, with an observed dimension close to 1 as expected. These last numerical experiments confirm the fact that our theory can be applied to observables of dynamical systems.

b. *N-variable system of Lorenz (1996)*

In sections 2c and 3b, we state that for a fixed catalog local density $L^{1/d}$, the sensitivity of analog-to-target distances with k is stronger in low dimension. We also make the link between this sensitivity and the choice of K to be made for the efficiency of analog methods. Here we propose a simple illustration with analog forecasting on the system of Lorenz (1969) (see appendix D for a description of the system).

We use a time step of 0.05 (nondimensional units) to generate catalogs. We perform one forecast experiment with $N = 12$ variables and another with $N = 20$ variables. Dimensions D_1 were estimated from Eq. (3) on an independent trajectory of 10^5 points, and with full, perfect observation catalogs of size 10^5 for each value of N (these were not the catalogs used to perform forecasts). This gives values of $D_1 \approx 8$ when $N = 12$ and $D_1 \approx 12$ when $N = 20$.

For the forecast experiment, we set the mean attractor density to $L^{1/D_1} \approx 3.5$. This number is intentionally low, placing ourselves in a situation where using too many analogs can be detrimental to the efficiency of analog forecasts. The catalog sizes were then $L = 10^3$ time units for the $N = 12$-variables system, and $L = 10^5$ time units for the $N = 20$-variable system. We used catalogs of noisy observations, adding independent and identically distributed (i.i.d.), zero-mean Gaussian white noises to a trajectory of full observations. The standard deviation of the noise was set to 1% of the root-mean-square distance (RMSD) between two states picked randomly in the attractor:

$$\mathrm{RMSD} = \left[\frac{1}{L(L-1)}\sum_{i\neq j}\mathrm{dist}(z_i, z_j)^2\right]^{1/2} \approx \left[\iint \mathrm{dist}(z, z')^2\, d\mu(z)\, d\mu(z')\right]^{1/2}.$$

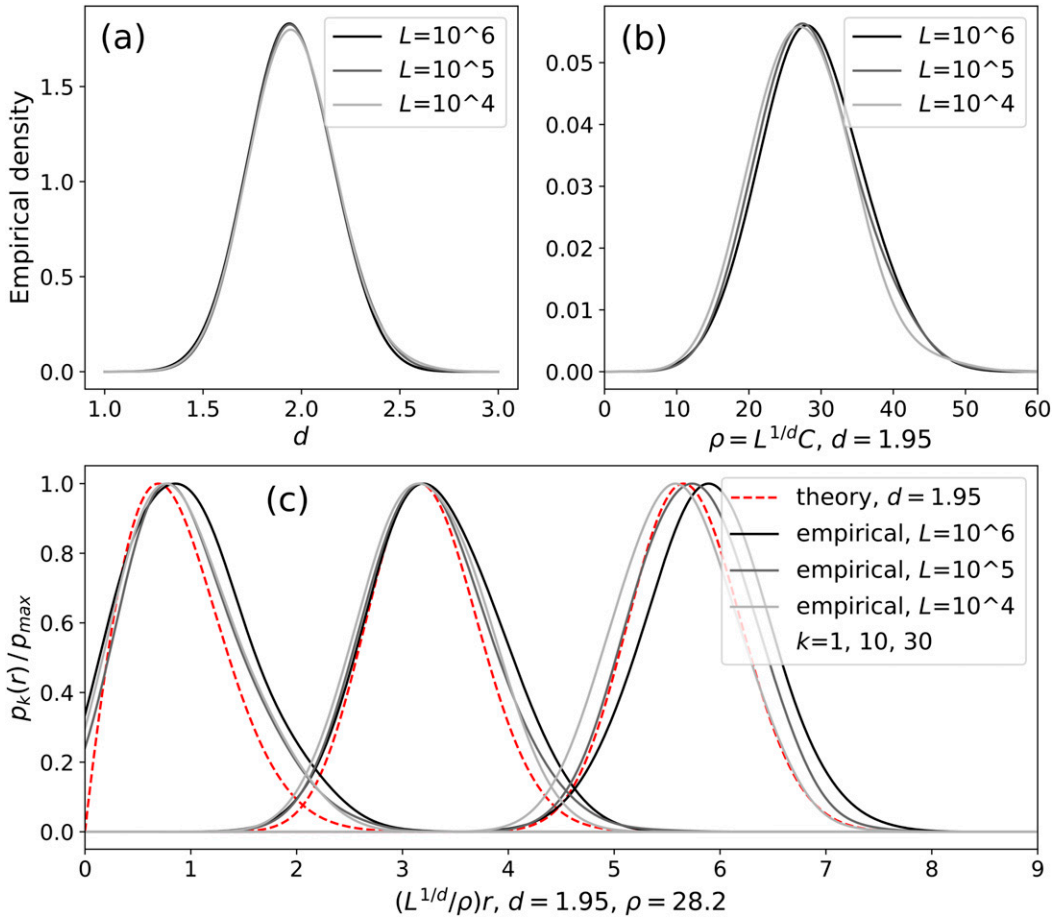

FIG. 4. Numerical experiments of the system of L63, for a fixed target point z, using catalogs of various sizes L, repeating the experiment 600 times for each catalog to obtain empirical probability densities. (a) Empirical density of the local dimension d, obtained with the method of Fig. 1 and with 150 analogs, (b) empirical density of $\rho(z)$ obtained from Eqs. (11) and (14), setting d to the mean value of its empirical densities, which is $d = 1.95$ here, and (c) normalized empirical probability densities of rescaled distances $(L^{1/d}/\rho)r$, setting ρ and d to the mean value of their empirical densities, that is $\rho = 28.2$ and $d = 1.95$ and normalized theoretical probability densities using the same value of d. The probability densities are estimated using Gaussian kernels with bandwidths of 0.15 (for d), 4 (for ρ), and 0.3 (for rescaled r).

The analog forecast was simply done with a weighted mean of the successors of the K closest analogs, and weights defined by Gaussian kernels $\omega_k \propto \exp(-r_k^2/2\lambda_k^2)$, where $\lambda_k(z)$ is defined as the median over k of the values $r_k(z)$ as explained in section 3b and used in Lguensat et al. (2017) and Platzer et al. (2021). Values of $K = 5$, 15, 25, 50, and 75 were tested for the total number of analogs. Distances were evaluated using the Euclidean metric. The analog forecast error was computed as the Euclidean distance between the analog forecast and the true future state, divided by the RMSD.

Figure 6 shows medians of analog forecast errors from this numerical experiment as a function of forecast horizon. First, it can be seen that the errors are very similar in magnitude, confirming that analog forecast errors strongly depend on analog-to-target distances (Platzer et al. 2021), which are largely determined by catalog density as we have seen. These errors are between 15% and 40% of the RMSD, which is the mean error of a climatological forecast that estimates the future state as a constant equal to the average over all states in the catalog. Therefore, the analog forecast errors from Fig. 6 appear to be relatively high, which was expected since the catalog density is quite low.

In higher dimension $D_1 \approx 12$ and for small forecast horizon ($\leq$0.15), using five analogs results in the highest forecast error, because for this system averaging through a large number of analogs helps the forecast and reduces observational noise (Platzer et al. 2021). Then, still for small forecast horizon ($\leq$0.15) and attractor dimension $D_1 \approx 12$, using 15, 25, 50, or 75 analogs does not make a significant difference. This is consistent with the fact that analog-to-target distances grow slowly with k in high dimension. Now, for the same system, the same catalog density L^{1/D_1}, the same time horizon ($\leq$0.15), but a lower attractor dimension D_1, the worst forecast is still witnessed with a low number of analogs $K = 5$, but values of K above 25 (i.e., $K = 50, 75$) increase forecast error, since analog-to-target distances grow faster with k in lower dimension. For

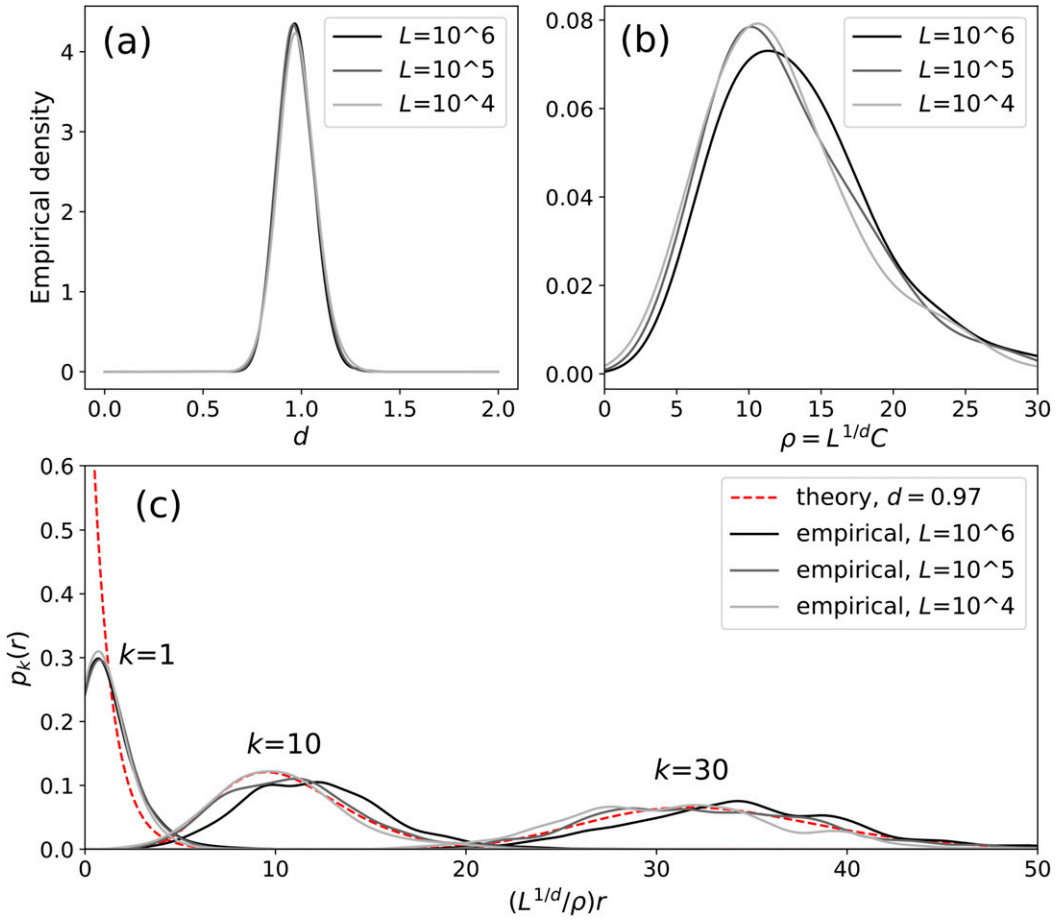

FIG. 5. As in Fig. 4, but only using observations of the first coordinate of the system of L63. The mean value of d that is used to produce (b) and (c) is $d = 0.97$. The mean value of ρ that is used to produce (c) is $\rho = 12.5$. In (c), the probability densities are not normalized, as $p_1(r)$ has no maximum value since $k < 1/d$. The empirical probability densities are estimated using Gaussian kernels with bandwidths of 0.15 (for d), 4 (for ρ), and 1 (for rescaled r).

larger forecast horizons ($\geq$0.15), the error is increased due to the chaotic dynamics of the system, and this growth is stronger for large values of K which correspond to larger analog-to-target distances. For these larger time horizons and in dimension $D_1 \approx 8$, using $K = 5$, 15, or 25 analogs results in lower forecast errors than using $K = 50$ or 75 analogs.

This example illustrates the higher sensitivity of analog methods to the choice of K in low dimension, at fixed catalog density L^{1/D_1}. However, it also shows that the main driver of analog-to-target distances is the catalog density, which is a rapidly decreasing function of dimension. Indeed, in this example, keeping a constant catalog density amounts to multiplying by 100 the catalog size while only multiplying by 1.5 the attractor dimension. Therefore, we stress again that at fixed catalog size L, reducing the dimension (through any dimension-reduction technique) allows us to find more analogs close to the target.

c. *AROME reanalysis data: Dimensionality*

To further appreciate the applicability of our results to high-dimensional, real geophysical systems, the theoretical developments from section 2 are tested on five years (2015–19) of hourly 10-m wind output from the physical model AROME (Ducrocq et al. 2005) coupled with satellite, radar, and in situ observations through a variational data assimilation scheme (similar to the one of Fischer et al. 2005). The spatial domain is an evenly spaced grid above Brittany, with latitudes ranging from 47.075° to 49.3° and longitudes from −5.7° to −2.575°, and a spacing of 0.025°. To focus on wind at sea, land points are removed from the data resulting in a domain of 8190 grid points.

Note that this dataset is not comprised of state vectors, but of partial observations (10-m wind, over a finite-width, evenly spaced grid) of the state of the atmosphere. Projections of the state z would be noted $y = f(z)$ classically. However, we keep the notations z, r_k, d, D_1, when referring to quantities computed directly from the 10-m wind data. As stated in section 2e, our analytical derivations are still valid for observational data, only that the dimension d obtained when searching analogs of observables can be different from the dimension obtained when searching analogs of the system state.

From these data, one can compute local dimensions with the method of Caby et al. (2019). As the data are limited ($\sim 3 \times 10^4$

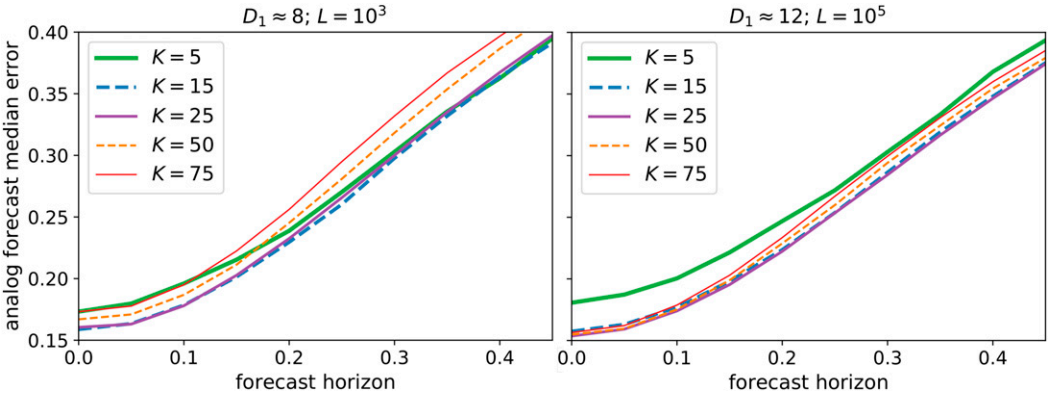

FIG. 6. Sensitivity of analog forecasting to the choice of K, for the same catalog density $L^{1/D_1} \approx 3.5$, but different attractor dimensions (and thus, catalog sizes), using the N-variable system of Lorenz (1996). (left) Lower attractor dimension D_1 and catalog size L, $N = 12$. (right) Higher attractor dimension D_1 and catalogs size L, $N = 20$. The catalogs are simulated from noisy observations of long trajectories. Analog forecasts are performed as weighted means of successors of the K-closest analogs.

time points), K is set to 40. Note also that, as elements of the catalog are only one hour away from each other, they cannot be assumed to be independent. Therefore, if several analogs are neighbors in time, only one analog is retained, and it is selected randomly in the set of time-neighboring analogs. Also, analogs that are less than one-and-a-half days away from the target z are discarded. Usually, analog are searched for in a time window of fixed length around the calendar date of the target z. However, in this example, searching for analogs with or without calendar-date restriction resulted in similar results for estimates of dimension and analog-to-target distances, indicating that the closest analogs naturally lied in similar seasons than their targets z.

Histograms of local dimensions d_{z,r_K} are plotted in Fig. 7a. These indicate that the (observed) system lives in an attractor of dimension approximately between 7 and 19, with some local dimensions likely to exceed 25. The average of these local dimensions d_{z,r_K}, noted D_1 here, is 16. Our local dimension histogram is similar in shape to the one of Faranda et al. (2017), who also focused on North Atlantic circulation (in their study, the local dimension is called "instantaneous dimension"). However, our histogram shows slightly higher average dimensions and a higher variability. Note that we focus on two components of horizontal wind velocity, on a dense grid of $\sim 10^4$ grid points, while Faranda et al. (2017) focus on sea level pressure (SLP) at $\sim 10^3$ grid points. Therefore, it is not surprising that we find higher average values of the local dimension. The fact that we observe a higher variability in the local dimension could be due to an intrinsic higher variability of this dynamical indicator, but also to a higher variability in the process of estimating d caused by a lack of data. Indeed, we have slightly less data than Faranda et al. (2017), for a system of slightly higher dimension, so that we can find fewer good analogs to estimate d than Faranda et al. (2017). Faranda et al. (2017) use $L \sim 2 \times 10^4$ days of historical data. We use $\sim 4 \times 10^4$ hours of data, which must be divided by the typical correlation time scale in hours. If we assume that the latter is between 12 and 24 h, we find that our L is between 1.5×10^3 and 3×10^3. Faranda et al. (2017) found a seasonality in the local dimension of SLP fields, with higher dimensions and a higher variability in winter. In our case, no seasonal trend for the mean or median dimension is observed, but the weekly variability of local dimensions is higher in winter, as witnessed in Fig. 7b. Also, a diurnal cycle can be seen in Fig. 7c, with dimension increasing in daytime and decreasing in nighttime. As diurnal variability is mixed with other sources of variability, it cannot always be identified by eye (see the three first days of Fig. 7c). Histograms of dimension restricted to daytime are similar to histograms restricted to nighttime, so that diurnal cycle does not appear to be the main driver of dimension variability.

We repeated the experiments leading to the histograms of Fig. 7a, but using different metrics (the Manhattan distance, order-8 Minkowski metric, and Chebyshev distance). This did not result in significant change, only that the dimension estimates were slightly larger when using the order-8 Minkowski and Chebychev metrics (not shown). This further demonstrates the robustness of our results to a change of metric.

d. *AROME reanalysis data: Analog distances*

An example of target state and analogs is shown in Fig. 8. The chosen target state is a classical winter situation in Brittany, with strong eastward wind coming from the sea. Thus, good analogs are found in the catalog. It is hard to discriminate which analog is closest: for such a high-dimensional system, the first analog-to-target distances are very similar.

Note that for this moderately high dimensional system, the concentration of norms might make the search for analogs meaningless as pointed out by Beyer et al. (1999). For very high-dimensional systems, the ratio between the distance to the nearest analog, r_1, and the distance to the furthest point in the catalog, r_L, is close to one, making the search for analog irrelevant. Moreover, Hinneburg et al. (2000) showed that for order-p Minkowski metrics the difference between the distance to the furthest point and to the nearest neighbor scales as $d^{(1/p)\,-\,(1/2)}$, indicating that for different types of distances the

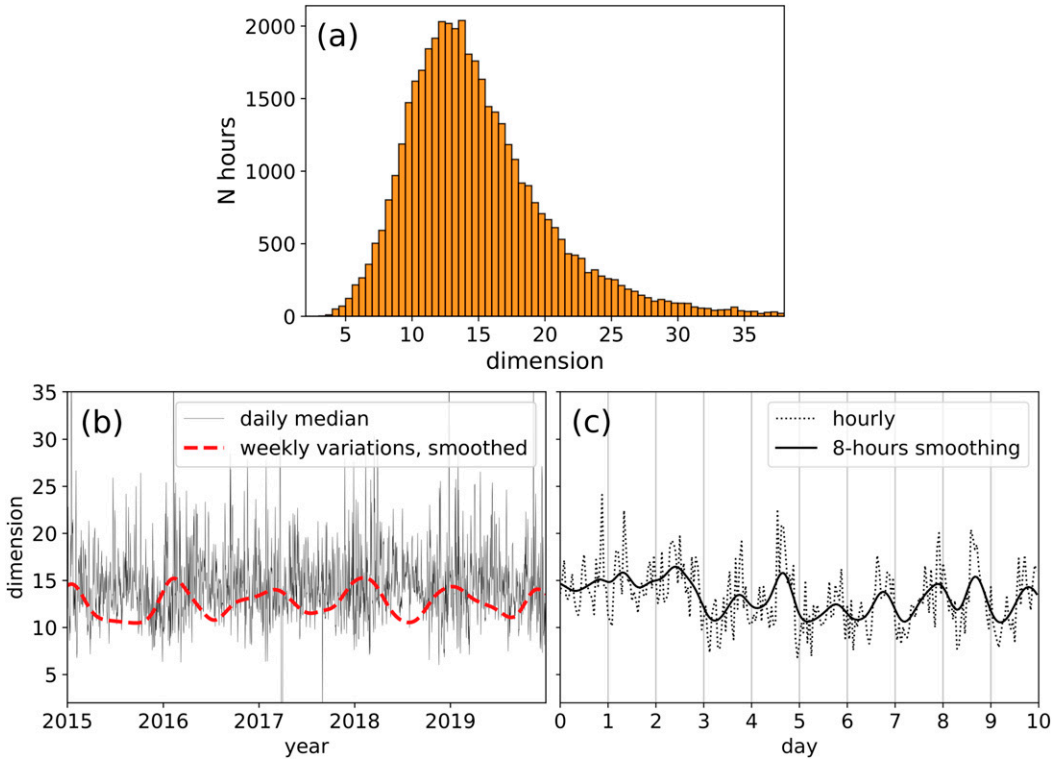

FIG. 7. Statistics of local dimensions estimated using the method of Caby et al. (2019), as in Fig. 1, with $K = 40$. (a) Histogram of dimension from 10-m wind data off the Brittany coast, (b) 5 years of dimension daily averages, and weekly variations defined as the difference between the 90% and 10% quantiles of hourly dimension over a week. This last quantity is smoothed over an ~80-day window using convolution and Gaussian kernels, and (c) 14 days of hourly local dimension, and an 8-h smoothing using convolution and Gaussian kernels.

concentration of norm might behave differently. To ensure that this concentration of norm was not an issue, we computed r_1/r_L for every point in the catalog (again, omitting neighbors in time to compute r_1), and for Minkowski metrics of order 1 (also called Manhattan distance), 2 (also called Euclidean distance), 8, and infinity (also called Chebyshev distance or infinity norm). This allowed us to compute histograms of r_1/r_L (not shown), which showed a very low probability for r_1/r_L to exceed 0.3 whatever the distance used. This shows that the curse of dimensionality is not a severe issue for our example of 10-m wind reanalysis, and that looking for analogs is still meaningful.

Using the estimated values of $d = d_{z,r_K}$ and $C(z)$ (through the least squares approximation introduced in the previous section), it is possible to approximate the rescaled theoretical variable υ_k (introduced in section 2d) through

$$\tilde{\upsilon}_k = dk^{1/2}\left(\frac{r_k}{Ck^{1/d}} - 1\right), \tag{15}$$

so that $\tilde{\upsilon}_k$ should be close to υ_k, especially for large values of k. However, due to the small catalog size, only probability densities up to $k = 8$ will be studied; otherwise, the expressions obtained theoretically in the limit $L \to +\infty$ are likely not to hold.

To obtain these distributions, analogs of each hourly $z \in \mathcal{C}$ (where $\mathcal{C}$ is the catalog) are sought for in the catalog, omitting analogs that are neighbors in time as explained previously. For each z, $C(z)$ is computed from Eq. (11), and the distances are rescaled following Eq. (15) and then stored. Finally, the stored values of each $\tilde{\upsilon}_k$ are used to estimate probability density functions using Gaussian kernels with a bandwidth of 0.3. Figure 9 shows the outcome of this procedure. For comparison, a similar procedure is applied on data from the model of L63, using a catalog of $L = 10^6$ points and testing the procedure on 10^5 target points that are taken from a trajectory independent from the catalog. Also, the theoretical density functions υ_k from Eq. (9) are shown for similar (fixed) dimensions. Note that to obtain distributions $\tilde{\upsilon}_k$ we are combining values obtained at different points and therefore different values of d_{z,r_K}. However, we should find $\langle\tilde{\upsilon}_k\rangle \approx 0$ and $\langle\tilde{\upsilon}_k^2\rangle \approx 1$.

Figure 9 shows a relatively good agreement between theoretical and empirical distributions, especially for the Lorenz data. Indeed, the curves of Figs. 9b and 9d are similar in shape, especially the asymmetry for $k = 1$. As k grows, the variance of the empirical data (Fig. 9b) becomes smaller than expected in theory (Fig. 9d). This can be explained by the fact that the assumption $L \to +\infty$ (or equivalently $r_k \to 0$) is better satisfied for low values of k. High values of r_k are associated with a low variability. This also explains the lower variance of the empirical curves (Fig. 9a) compared to the theoretical curves

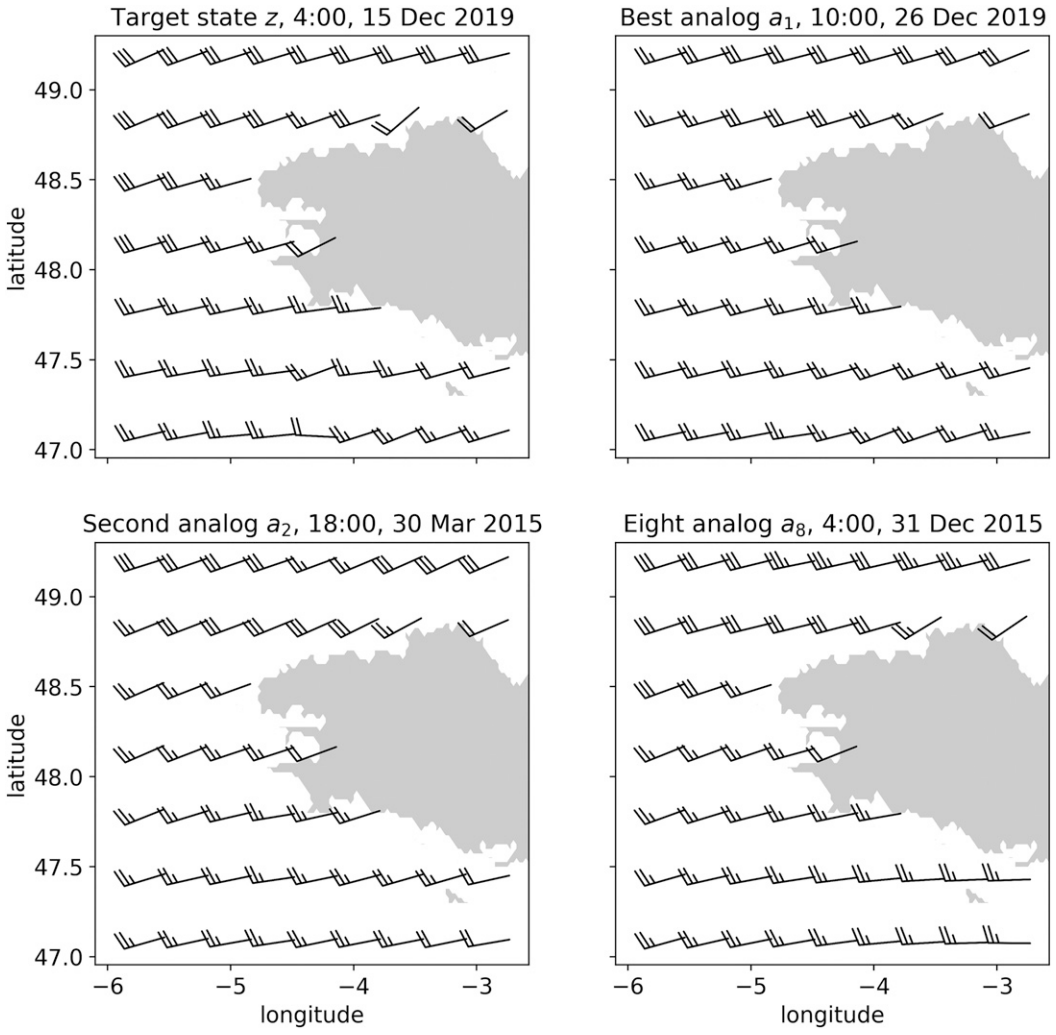

FIG. 8. An example of (top left) target state z and the (top right) first, (bottom left) second, and (bottom right) eighth analogs, using 10-m wind data off the coast of Brittany from the AROME reanalysis. Standard station model notations are used, with wind speed in knots and point-centered flags.

(Fig. 9c), using the wind data. Again, the asymmetry in the shape of the curves for $k = 1$ is respected, and the estimation of the mean fits our theory.

This experiment shows that the present theory, which was derived assuming a large catalog density, is also partially applicable to limited catalog densities (here $L^{1/D_1} \approx 1.6$, which is even lower than the example of section 4b). Although we overestimate the standard deviation of r_k at fixed k, z, and L, our estimates of the mean $\langle r_k(z) \rangle$ are satisfying even for low catalog densities. Therefore, most of our theory seems to be applicable to partial observations of real, moderately high-dimensional systems, with limited catalog size (here, only 5 years of data, for a system of observed dimension $D_1 \approx 16$). The fact that our theory could eventually break down for even lower values of the catalog density is not worrying, as it would mean that analog-to-target distances would probably be too large for analogs to be used.

e. *AROME reanalysis data: Objective-based dimension reduction*

In this section, we apply a dimension reduction technique to the AROME reanalysis data in order to achieve the following criterion:

$$\frac{\overline{r_k}}{\mathrm{RMSD}} < \varepsilon, \tag{16}$$

where $\overline{r_k}$ is the mean over all target points of the kth analog-to-target distance, RMSD is the root-mean-squared distance between two points randomly taken from the dataset, and ε is a user-defined threshold. $\overline{r_k}$ is thus different from $\langle r_k(z) \rangle$, which is the mean over all possible realizations of the kth analog-to-target distance at fixed target z and catalog size L. The average $\overline{r_k}$ does not depend on z, while in the rest of this document $r_k(z)$ depends on z, and so does $\langle r_k(z) \rangle$.

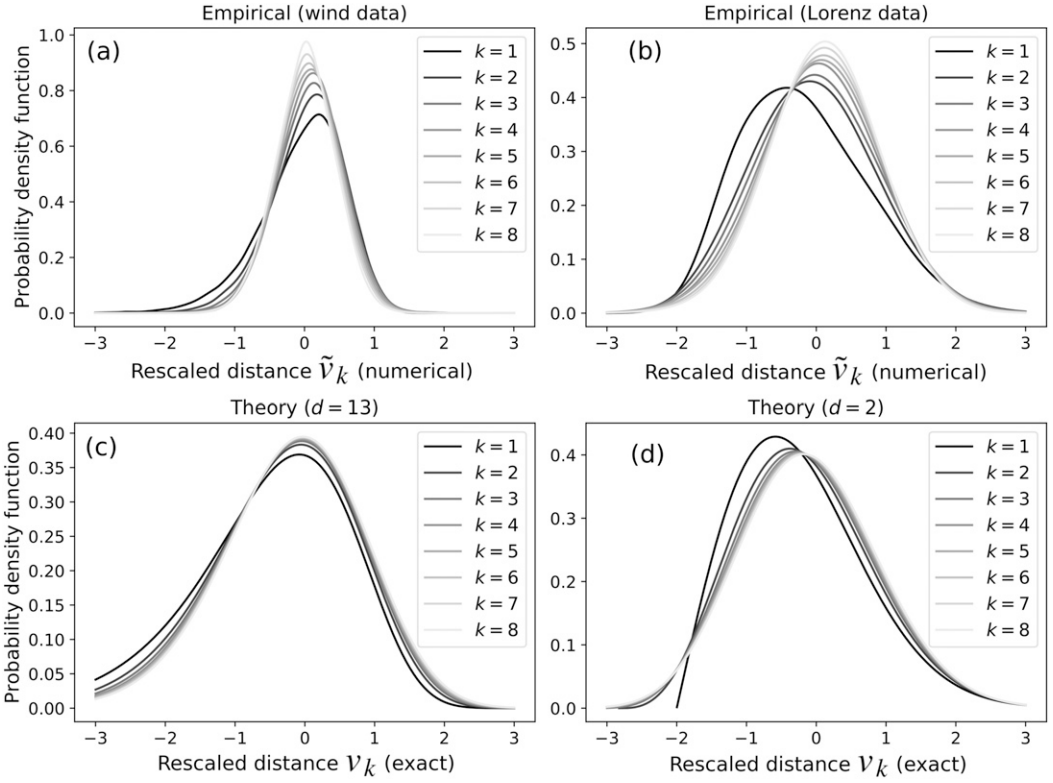

FIG. 9. Probability densities of rescaled analog-to-target distances r_k (a) from 10-m wind data off the Brittany coast and (b) from numerical experiments of the L63 system, compared to theoretical distributions from Eq. (9) for a local dimension of (c) 13 and (d) 2. Empirical probability densities are estimated using Gaussian kernels with a bandwidth of 0.3. Empirical values of d_{z,r_K} are estimated with $K = 40$.

We reduce dimension using EOFs, which allows us to reduce $\overline{r_k}$/RMSD. However, one might not want to reduce dimension too much, in order to keep enough information on the state of the system. In this scenario, the practical question is, What is the maximum number of EOFs that can be used in order to meet Eq. (16)?

We use the notation $d^{\text{eof}} = d^{\text{eof}}_{z,r_K}$ for the local dimension estimated as previously but after applying the projection on a limited number of EOFs noted N^{eof}. We note $D_1^{\text{eof}} = (1/L)\sum_i d^{\text{eof}}_{z_i,r_K(z_i)}$ where the sum is over all elements of the catalog. D_1^{eof} is thus the average dimension of the dataset after projection onto the N^{eof} first EOFs.

According to the theoretical study of Caby et al. (2020), we expect D_1^{eof} to be inferior to both N^{eof} and the attractor dimension of the dynamical system under study. For large enough N^{eof} we should find $D_1^{\text{eof}} \approx D_1$ (where D_1 is the dimension found using the original dataset). For small N^{eof}, in principle, $D_1^{\text{eof}} \approx N^{\text{eof}}$. The numerical experiments presented below show that the behavior is more complex when N^{eof} is close to D_1.

Following from the theoretical results of this paper, we assume that, for each target point z,

$$\frac{\langle r_k(z)\rangle}{\text{RMSD}} = \rho(z)\left(\frac{k}{L}\right)^{1/d^{\text{eof}}},$$

where $\rho(z)$ is on the order of one. When using the method described in the previous sections to compute d^{eof}_{z,r_K} and $C(z)$, we find that $\rho(z)$ is typically between 0.4 and 0.7. Then we make the following ergodicity hypothesis:

$$\overline{r_k(z)} = \overline{\langle r_k(z)\rangle},$$

neglecting the variations of d^{eof}_{z,r_K} with z, we finally find the approximate scaling:

$$\frac{\overline{r_k}}{\text{RMSD}} \approx \overline{\rho}\left(\frac{k}{L}\right)^{1/D_1^{\text{eof}}},$$

which gives, combined with Eq. (16):

$$D_1^{\text{eof}} < D_{\max,k} := \frac{\log(L/k)}{-\log(\varepsilon/\overline{\rho})}.$$

From this formula, it appears that $D_{\max,k}$ is a linear function of $\log(k)$. This can be rearranged to give

$$D_{\max,k} = D_{\max,1}\left[1 - \frac{\log(k)}{\log(L)}\right]. \tag{17}$$

This last expression shows how $D_{\max,k}$ strongly depends on k. On a practical example, assume that $D_{\max,1} \approx 10$ and that $L = 10^4$, then $D_{\max,25} \approx 6$. In this experiment, we assume that

the number of required analogs is fixed. Reducing dimension in order to decrease analog distances thus strongly depends on how many analogs are needed for the analog method. For instance, if an ensemble of analogs is used to estimate the full probability density function of a one-dimensional variable (say, the day after tomorrow's accumulated rainfall over the city of Paris), then one might need at least 100 analogs. Yet 10 analogs might be enough to simply estimate the mean of the distribution. As another example, if one wants to estimate the covariance associated with the forecast error of 5 independent variables, one needs at the very least 5 analogs, but 50 analogs might be necessary, especially in the presence of observational noise. Also, the complexity of the system under study might vary according to phase space location, so that the number of required analogs could depend on the state z. In practice, the number of required analogs is a complex function of the quantity to be estimated, the quality of the data, the method that is used, and properties of the system at stake.

Figure 10 shows comparison of this scaling with numerical experiments performed on the AROME reanalysis data. Upper and lower bounds for $D_{\max,k}$ were derived from estimations of D_1^{eof} and by checking whether the criterion $\overline{r_k}/\text{RMSD} < \varepsilon$ was met. For low values of N^{eof} we find that $D_1^{\text{eof}} \approx N^{\text{eof}}$, and for high values of N^{eof} we find that $D_1^{\text{eof}} < D_1$ while we expected $D_1^{\text{eof}} \approx D_1$. This calls for more theoretical studies on the dimension of observables. However, considering only the applicability of Eq. (17), Fig. 10 shows a satisfying agreement between our theoretical scaling and the numerical experiments, especially given the number of approximations that we have taken.

The way to use these equations for $D_{\max,k}$ in practice depends on the particular application. For instance, if one wants to perform a statistical task such as downscaling, one might impose a fixed minimum number of K samples to correctly represent a statistical distribution. One might, at the same time, ask that the analog-to-target distance does not exceed a given threshold to ensure a good quality of analogs (assuming that this "quality" is correctly estimated by the chosen distance). Then our formulas can be used to estimate how much of dimension reduction is needed to fulfil these criteria by choosing a number of EOFs close to the theoretical value of $D_{\max,k}$.

Another possibility is that the required number of samples K varies with D_1. This is the case in ensemble forecast where one wants to use successors to estimate the covariance matrix of the future state. If the local dimension is d, we can assume that the data have been projected on some $\lceil d \rceil$-dimensional space, where $\lceil d \rceil$ is the ceiling function of d (i.e., the smallest integer i such that $i \geq d$). In this case, the covariance matrix of the future state is of size $\lceil d \rceil(\lceil d \rceil + 1)/2$, and each successor is a $\lceil d \rceil$-dimensional vector. Therefore, one needs to have at least $K \geq (\lceil d \rceil + 1)/2$ for the successors to be able to estimate the covariance matrix using the estimation formulas of Lguensat et al. (2017). Identifying d with D_1, this last inequality can be rewritten in the form $D_1 \leq D'_{\max,K}$, where $D'_{\max,K}$ is a growing function of K (in this covariance example, $D'_{\max,K} = 2K - 1$). Since $D_{\max,K}$ from Eq. (17) is a decreasing function of K, the intersection D^* between $D_{\max,K}$ and $D'_{\max,K}$ (i.e., the dimension D^* so that there is a value K^* for which $D_{\max,K^*} = D'_{\max,K^*}$) gives a maximum value for D_1 that is independent of K. This maximum value is fixed by the threshold ε and the required relationship between the minimum value of K and the dimension D_1. Knowing D^*, one can estimate the optimal number of EOFs to use.

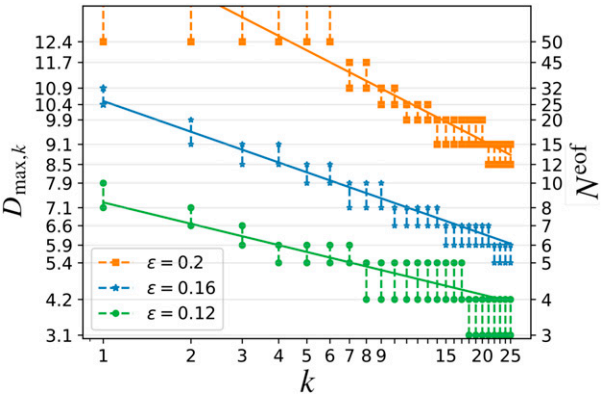

FIG. 10. Maximum dimension (or number of EOF) to fulfill the criterion $(1/\text{RMSD})\overline{r_k} < \varepsilon$, where $\overline{r_k}$ is the mean over all target points of the kth analog-to-target distance, RMSD is the root-mean-squared distance between two random points from the dataset, and ε is a user-defined threshold. We use the 10-m wind data, and we project both component simultaneously on N^{eof} basis functions. For a given value of N^{eof}, the dimension D_1^{eof} is computed as the mean of dimensions estimated from the method of Caby et al. (2019). Then $(1/\text{RMSD})\overline{r_k}$ is computed empirically and compared to ε, giving upper and lower bounds for the maximum dimension $D_{\max,k}$. Full lines show the theoretical scaling $D_{\max,k} = D_{\max,1}[1 - \log(k)/\log(L)]$. The values of $D_{\max,1}$ were set by hand in order to fit visually the so-obtained upper and lower bounds for $D_{\max,k}$, and L was set to $\sim 2 \times 10^3$, which corresponds to a correlation time scale of 24 h.

However, note that our formulas do not reveal how much information is left behind when reducing the dimension. For instance, in the case of forecast, the maximum dimension $D_{\max,K}$ might be too low to represent accurately the dynamics of the system. In such a case, one is bound to either raising the value of L (which can rarely be done) or increasing the value of ε (which might decrease the efficiency of the analog method).

5. Conclusions

We combined extreme value theory and dynamical systems theory to derive analytical joint probability distributions of analog-to-target distances in the limit of large catalog density. Those distributions shed new light on the influence of dimension in practical use of analog. In particular, we found that analog-to-target distances are more sensitive to the number of analogs used in low dimension than in high dimension, at fixed catalog density. Contrarily to previous works on the probability to find good analogs, this study focuses on distances rather than return times, gives whole probability distributions rather than first moments, and looks at the K closest analogs rather that only the closest one. Numerical simulations of the three-variable Lorenz system confirm the theoretical findings. An example of practical consequence of our theory on the

sensitivity of analog forecasts to the number of analogs used, depending on dimension, is given using the system of Lorenz (1996). The 10-m wind reanalysis data from the AROME physical model show that our analysis is also relevant for observations of real systems. Our investigation indicates that the studied wind fields lie in an attractor of moderately high dimension ~16. In this situation of moderate dimensionality, the analog-to-target distances of the first analogs are all very similar and have a low variability. Our theoretical derivations can be used to find optimal dimension reduction for the purpose of decreasing analog distances, which we demonstrate on an example using the AROME reanalysis data. These examples reveal the applicability of the derived probability distributions even to relatively low catalog densities.

Acknowledgments. The work was financially supported by ERC Grant 338965-A2C2 and ANR Grant 10-IEED-0006-26 (CARAVELE project). This piece of work took its origins in discussion with Théophile Caby, to whom we express our gratitude. In particular, appendix A is an adaptation of a derivation by Théophile Caby. The theoretical derivations of the probability density functions shown in this paper are the result of several exchanges with Benoît Saussol, who we must thank here. We are indebted to Fabrice Collard, Bertrand Chapron, and Caio Stringari, for fruitful insights and discussions about the exploration and interpretation of the AROME reanalysis data. Finally, the last version of this manuscript owes much to the meticulous work of three anonymous reviewers who we thank again here.

APPENDIX A

Proof that D_1 is Independent of Metric Choice in Finite Dimension

A metric dist$(\cdot, \cdot)$ associates a real positive number to any two vectors $\mathbf{z}_1 \neq \mathbf{z}_2$, and must verify dist$(\mathbf{z}_1, \mathbf{z}_1) = 0$, dist$(\mathbf{z}_1, \mathbf{z}_2)$ = dist$(\mathbf{z}_2, \mathbf{z}_1)$, and for any third vector $\mathbf{z}_3$, dist$(\mathbf{z}_1, \mathbf{z}_3) \leq$ dist$(\mathbf{z}_1, \mathbf{z}_2)$ + dist$(\mathbf{z}_2, \mathbf{z}_3)$.

Let dist$(\cdot, \cdot)$ and dist$'(\cdot, \cdot)$ be two metric acting on a finite-dimensional space. We note $B'_{z,r}$ the ball of radius r around the point z, defined with the distance dist$'(\cdot, \cdot)$, such that $B'(z, r) = \{a | \text{dist}'(z, a) < r\}$. This allows us to define the finite-resolution local dimension:

$$d'_{z,r} = \frac{\log\mu(B'_{z,r})}{\log r},$$

and the attractor dimension $D'_1 = \lim_{r\to 0} d'_{z,r}$. Quantities without primes $'$ are defined using the regular distance dist$(\cdot, \cdot)$.

Here, we will prove that $D'_1 = D_1$. The finite-dimension hypothesis implies strong equivalence of metrics; therefore, there exists two real positive numbers q and Q such that for all points z and a:

$$q\,\text{dist}'(z,a) \leq \text{dist}(z,a) \leq Q\,\text{dist}'(z,a). \tag{A1}$$

It is easy to check that this implies the double inclusion $B_{z,qr} \subseteq B'_{z,r} \subseteq B_{z,Qr}$, for all points z and all positive real number r. Taking the logarithm of the measure of this double inclusion, we find

$$\log\mu(B_{z,qr}) \leq \log\mu(B'_{z,r}) \leq \log\mu(B_{z,Qr}),$$

and, dividing by log r,

$$\frac{\log\mu(B_{z,qr})}{\log(qr) - \log q} \leq \frac{\log\mu(B'_{z,r})}{\log r} \leq \frac{\log\mu(B_{z,Qr})}{\log(Qr) - \log Q}.$$

Taking the limit of this last inequality when $r \to 0$ gives $D'_1 = D_1$.

This means that for small values of r, $d'_{z,r}$,, and $d_{z,r}$ approach the same limit, and are therefore close to each other. However, this proof does not give the convergence rate. In particular, it is possible to find a metric dist$'(\cdot, \cdot)$ such that the rate of convergence of $d'_{z,r}$ toward D_1 is arbitrarily slow. Therefore, for a given dataset, it is always possible to find a specific metric such that dimension estimates are far from the real limit value D_1. Nevertheless, this peculiar behavior is not expected for usual metrics, such as the order-p Minkowski metrics used in the numerical examples of the present paper.

APPENDIX B

Direct Proof for $p_k(r)$

In this appendix, we give the proof of Eq. (5) by evaluating directly the probability that analogs lie between the sphere of radius r and the sphere of radius $r + \delta r$.

a. Poisson distribution of the number of analogs in a ball

Haydn and Vaienti (2019) have shown that, for dynamical systems having rare event Perron–Frobenius operator properties, and for nonperiodic points z, the number of visits $V(z, r)$ of a trajectory of size L into the ball $B_{z,r}$ follows a Poisson distribution with mean $L\mu(B_{z,r})$:

$$\mathbb{P}[V(z,r) = k] = \frac{[L\mu(B_{z,r})]^k}{k!} e^{-L\mu(B_{z,r})}, \tag{B1}$$

where $k!$ is k factorial. In the context of analogs, this is the probability to find k analogs with distances to z below the radius r. In machine learning, this is called the epsilon nearest neighbor search. In the following we write $\mu_{z,r} := \mu(B_{z,r})$.

b. Distribution of analogs close to the sphere

Now we will use μ to evaluate $\mathbb{P}(r_k \in [r, r + \delta r))$, the probability that the kth analog-to-target distance is between r and $r + \delta r$, for fixed k and z and where δr is small compared to r.

The event "$r_k \in [r, r + \delta r)$" is the intersection of the event "there are $k - 1$ analogs in the ball $B_{z,r}$" and the event "there is one analog in $B_{z,r+\delta r} \cap \overline{B_{z,r}}$." For a Poisson point process these two events are independent (Daley and Vere-Jones 2003), so that

$$\begin{aligned}\mathbb{P}(r_k \in [r, r+\delta r)) &= \mathbb{P}[V(z,r) = k-1 \wedge \exists x \in \mathcal{C} \cap B_{z,r+\delta r} \cap \overline{B_{z,r}}] \\ &= \mathbb{P}[V(z,r) = k-1]\mathbb{P}(\exists x \in \mathcal{C} \cap B_{z,r+\delta r} \cap \overline{B_{z,r}}) \\ &= \frac{(L\mu_{z,r})^{k-1}}{(k-1)!} e^{-L\mu_{z,r}} \mathbb{P}(\exists x \in \mathcal{C} \cap B_{z,r+\delta r} \cap \overline{B_{z,r}}).\end{aligned} \tag{B2}$$

Then it follows from Haydn and Vaienti (2019) that the event that strictly one element of the catalog lies between $B_{z,r}$ and $B_{z,r+\delta r}$ has a probability of the same form as Eq. (B1) but replacing k by 1 and $\mu_{z,r}$ by $\delta\mu_{z,r} := \mu_{z,r+\delta r} - \mu_{z,r}$:

$$\mathbb{P}(\exists! x \in \mathcal{C} \cap B_{z,r+\delta r} \cap \overline{B_{z,r}}) = L\delta\mu_{z,r} e^{-L\delta\mu_{z,r}}. \tag{B3}$$

If the invariant measure μ is regular enough so that $\lim_{\delta r \to 0} \delta\mu_{z,r} = 0$ we then have $e^{-L\delta\mu_{z,r}} \approx 1$. Also, the probability to find more than one element of the catalog between $B_{z,r}$ and $B_{z,r+\delta r}$ has a probability of $\mathcal{O}(\delta\mu_{z,r})^2$. This justifies the approximation $\mathbb{P}(\exists x \in \mathcal{C} \cap B_{z,r+\delta r} \cap \overline{B_{z,r}}) \approx \mathbb{P}(\exists! x \in \mathcal{C} \cap B_{z,r+\delta r} \cap \overline{B_{z,r}})$. Finally, combining Eqs. (B2) and (B3), one finds

$$\mathbb{P}(r_k \in [r, r+\delta r)) = L\delta\mu_{z,r} \frac{(L\mu_{z,r})^{k-1}}{(k-1)!} e^{-L\mu_{z,r}}. \tag{B4}$$

This last equation is a more general form of our main result which is given in the next section. Here, the probability is expressed in terms of the invariant measure, which is usually not known analytically. The next section expresses the same probability in terms of the analog-to-target distance r.

c. *Distribution of analog-to-target distances*

The link between $\mu_{z,r}$ and r is given by the definition of the finite-resolution local dimension in Eq. (1):

$$\mu_{z,r} = r^d, \tag{B5}$$

where $d = d_{z,r}$. In this section, we first acknowledge the variations of $d_{z,r}$ with r, to better justify why they are neglected in the rest of the paper. Therefore, in this section $d = d_{z,r}$ for varying values of z and r, while in the rest of the paper d usually refers to a value at fixed distance r_K, d_{z,r_K}.

The link between $\delta\mu_{z,r}$ and δr involves variations of the local dimension with r. Let $\Delta = d_{z,r+\delta r} - d_{z,r}$, we have $\mu_{z,r+\delta r} = (r + \delta r)^{d+\Delta} = \mu_{z,r} r^{\Delta}(1 + \delta r/r)^{d+\Delta}$, which gives

$$\frac{\delta\mu_{z,r}}{\mu_{z,r}} = \left(1 + \frac{\delta r}{r}\right)^{d+\Delta} e^{\Delta \log r} - 1. \tag{B6}$$

Using the regularity hypothesis $\Delta \ll d$, and keeping only lower-order terms, we find

$$\frac{\delta\mu_{z,r}}{\mu_{z,r}} \approx d\frac{\delta r}{r} + \Delta \log r. \tag{B7}$$

The term $d(\delta r/r)$ represents an almost steady increase in $\mu_{z,r}$ when r grows. The term $\Delta \log r$ represents fluctuations in this increase given by the fluctuations in $d_{z,r}$. In practice, the method described in section 2b to evaluate d_{z,r_K} should catch a mean local dimension over the analogs and not catch the fluctuations of $d_{z,r}$ with r at scales smaller than r_K. Thus, the approximation

$$\frac{\delta\mu_{z,r}}{\mu_{z,r}} \approx d\frac{\delta r}{r}, \tag{B8}$$

which is not valid in theory, should be relevant in practice for finite catalog size and regular enough measures. For small enough δr, one can then define p_k, the probability density function of r_k through the identity $\mathbb{P}(r_k \in [r, r + \delta r)) = p_k(r)\delta r$. Combining Eqs. (B4), (B5), and (B8), we find

$$p_k(r) = dLr^{d-1} \frac{(Lr^d)^{k-1}}{(k-1)!} e^{-Lr^d},$$

which is the main result of this paper.

APPENDIX C

Alternative Proof for $p_k(r)$ and Joint Probability Distribution Using K Largest-Order Statistics

Lucarini et al. (2016) give a detailed analysis of the map from $\mathcal{A}$ to $\mathbb{R}$, $x \mapsto -\log \mathrm{dist}(z, x)$, using tools from dynamical systems theory and extreme value theory (EVT). For our purpose, it is interesting to look at the simpler distance map $x \mapsto \mathrm{dist}(z, x)$.

The minimum of this map over the catalog is achieved for the closest analog of z, a_1. The minimum is thus r_1. EVT tells (see Coles et al. 2001) that in the limit of large catalog, the minimum of this lower-bounded distance map on a finite sample of the attractor (a catalog of size L) follows a Weibull distribution, after rescaling. The Poisson law from Eq. (B1) with $k = 1$ actually gives the scaling and the exact form of the Weibull distribution:

$$\mathbb{P}(r_1 > r) = e^{-Lr^d},$$

for positive r; otherwise, the probability is 1.

The K largest-order statistics of this function then correspond to the K analogs of the point z. Again, in the limit of large catalog and for small enough K, EVT provides the limit law (see Coles et al. 2001) for the kth minima of this distance function when $L \to \infty$:

$$\mathbb{P}(r_k > r) = e^{-Lr^d} \sum_{s=0}^{k-1} \frac{(Lr^d)^s}{s!}.$$

Differentiating and with a bit of rearrangement, one finds back the formula of Eq. (5):

$$\begin{aligned} p_k(r) &= -\frac{\partial}{\partial r}\mathbb{P}(r_k > r) \\ &= dLr^{d-1} \frac{(Lr^d)^{k-1}}{(k-1)!} e^{-Lr^d}. \end{aligned}$$

From a broader perspective, extremal process theory (Lamperti 1964) gives the joint distribution of analog-to-target distances $p_{1:K}$ in the limit $L \to \infty$:

$$p_{1:K}(r_1, \ldots, r_K) = (dL)^K \left(\prod_{k=1}^{K} r_k\right)^{d-1} e^{-Lr_K^d},$$

where the function is nonzero only when $0 < r_1 < r_2 < \cdots < r_K$. For notation convenience and only in this formula, the random variables r_k are noted identically as the values they can possibly take.

APPENDIX D

Three-Variable Lorenz System

The three-variable L63 system of equations is

$$\begin{cases} \dfrac{dx_1}{dt} = \beta_1(x_2 - x_1), \\ \dfrac{dx_2}{dt} = x_1(\beta_2 - x_3) - x_2, \\ \dfrac{dx_3}{dt} = x_1x_2 - \beta_3x_3, \end{cases} \tag{D1}$$

with usual parameters $\beta_1 = 10$, $\beta_2 = 28$, and $\beta_3 = 8/3$. In this case, the variables X_1, X_2, and X_3 span values between approximately $[-20, 20]$, $[-20, 20]$, and $[0, 40]$, respectively. If we now make the following change of variables,

$$\begin{cases} x_1 \to X_1 = \dfrac{x_1}{\beta_2}, \\ x_2 \to X_2 = \dfrac{x_2}{\beta_2}, \\ x_3 \to X_3 = \dfrac{x_3}{\beta_2}, \end{cases}$$

amounts to changing the units of all variables by the same amount. In this case, the new set of governing equation becomes

$$\begin{cases} \dfrac{dX_1}{dt} + \beta_1X_1 = \beta_1X_2, \\ \dfrac{dX_2}{dt} + X_2 = \beta_2X_1(1 - X_3), \\ \dfrac{dX_3}{dt} + \beta_3X_3 = \beta_2X_1X_2, \end{cases}$$

which is very similar to the usual set of equation. Setting the same values for the parameters gives the same chaotic patterns, only in different units. The local dimensions of the system are the same, but now X_1, X_2, and X_3 span values between approximately $[-2/3, 2/3]$, $[-2/3, 2/3]$, and $[0, 4/3]$, respectively. For this new system, the values of $\rho(z)$ calculated as in section 4a of the present paper would be close to 1 and not to 28.

Finally, the N-variable system of Lorenz (1996) is defined by the following equations:

$$\forall i \in [1, N], \quad \frac{dx_i}{dt} = -(x_{i-2} + x_{i+1})x_{i-1} - x_i + \theta, \tag{D2}$$

where θ is the forcing parameter. In our numerical experiments we use the value $\theta = 8$ and two different values of $N = 12$ and $N = 20$, with periodic boundary conditions $x_{i+n} = x_i$.

REFERENCES

Alexander, R., Z. Zhao, E. Székely, and D. Giannakis, 2017: Kernel analog forecasting of tropical intraseasonal oscillations. *J. Atmos. Sci.*, **74**, 1321–1342, https://doi.org/10.1175/JAS-D-16-0147.1.

Ayet, A., and P. Tandeo, 2018: Nowcasting solar irradiance using an analog method and geostationary satellite images. *Sol. Energy*, **164**, 301–315, https://doi.org/10.1016/j.solener.2018.02.068.

Beyer, K., J. Goldstein, R. Ramakrishnan, and U. Shaft, 1999: When is "nearest neighbor" meaningful? *Int. Conf. on Database Theory*, Jerusalem, Israel, ICDT, 217–235.

Birkhoff, G. D., 1931: Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. USA*, **17**, 656–660, https://doi.org/10.1073/pnas.17.2.656.

Caby, T., D. Faranda, G. Mantica, S. Vaienti, and P. Yiou, 2019: Generalized dimensions, large deviations and the distribution of rare events. *Physica D*, **400**, 132143, https://doi.org/10.1016/j.physd.2019.06.009.

——, ——, S. Vaienti, and P. Yiou, 2020: Extreme value distributions of observation recurrences. *Nonlinearity*, **34**, 118–163, https://doi.org/10.1088/1361-6544/abaff1.

Cattiaux, J., R. Vautard, C. Cassou, P. Yiou, V. Masson-Delmotte, and F. Codron, 2010: Winter 2010 in Europe: A cold extreme in a warming climate. *Geophys. Res. Lett.*, **37**, L20704, https://doi.org/10.1029/2010GL044613.

Coles, S., J. Bawa, L. Trenner, and P. Dorazio, 2001: *An Introduction to Statistical Modeling of Extreme Values*. Vol. 208. Springer, 208 pp.

Daley, D. J., and D. Vere-Jones, 2003: *Elementary Theory and Methods*. Vol. I, *An Introduction to the Theory of Point Processes*, Springer, 471 pp.

Ducrocq, V., F. Bouttier, S. Malardel, T. Montmerle, and Y. Seity, 2005: Le projet AROME. *Houille Blanche*, **91**, 39–43, https://doi.org/10.1051/lhb:200502004.

Faranda, D., V. Lucarini, G. Turchetti, and S. Vaienti, 2011: Extreme value distribution for singular measures. arXiv, https://arxiv.org/abs/1106.2299.

——, G. Messori, and P. Yiou, 2017: Dynamical proxies of North Atlantic predictability and extremes. *Sci. Rep.*, **7**, 41278, https://doi.org/10.1038/srep41278.

Farmer, J. D., and J. J. Sidorowichl, 1988: Exploiting chaos to predict the future and reduce noise. *Evolution, Learning and Cognition*, Y. C. Lee, Ed., World Scientific, 277–330.

Fettweis, X., E. Hanna, C. Lang, A. Belleflamme, M. Erpicum, and H. Gallée, 2013: Important role of the mid-tropospheric atmospheric circulation in the recent surface melt increase over the Greenland ice sheet. *Cryosphere*, **7**, 241–248, https://doi.org/10.5194/tc-7-241-2013.

Fischer, C., T. Montmerle, L. Berre, L. Auger, and S. E. Ştefănescu, 2005: An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system. *Quart. J. Roy. Meteor. Soc.*, **131**, 3477–3492, https://doi.org/10.1256/qj.05.115.

Gupta, A., R. Krauthgamer, and J. R. Lee, 2003: Bounded geometries, fractals, and low-distortion embeddings. *44th Annual IEEE Symp. on Foundations of Computer Science*, Cambridge, MA, IEEE, 534–543, https://doi.org/10.1109/SFCS.2003.1238226.

Hamill, T. M., M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1.

Hamilton, F., T. Berry, and T. Sauer, 2016: Ensemble Kalman filtering without a model. *Phys. Rev. X*, **6**, 011021, https://doi.org/10.1103/PhysRevX.6.011021.

Haydn, N., and S. Vaienti, 2019: Limiting entry times distribution for arbitrary null sets. arXiv, https://arxiv.org/abs/1904.08733.

Hinneburg, A., C. C. Aggarwal, and D. A. Keim, 2000: What is the nearest neighbor in high dimensional spaces? *26th Int. Conf.*

on Very Large Databases, Cairo, Egypt, VLDB, 506–515, https://www.vldb.org/dblp/db/conf/vldb/HinneburgAK00.html.

Houle, M. E., 2013: Dimensionality, discriminability, density and distance distributions. *2013 IEEE 13th Int. Conf. on Data Mining Workshops*, Dallas, TX, IEEE, 468–473, https://doi.org/10.1109/ICDMW.2013.139.

——, 2017: Local intrinsic dimensionality I: An extreme-value-theoretic foundation for similarity applications. *Int. Conf. on Similarity Search and Applications*, Munich, Germany, SISAP, 64–79.

Jézéquel, A., P. Yiou, and S. Radanovics, 2018: Role of circulation in European heatwaves using flow analogues. *Climate Dyn.*, **50**, 1145–1159, https://doi.org/10.1007/s00382-017-3667-0.

Kac, M., 1959: *Probability and Related Topics in Physical Sciences*. Vol. 1. Interscience Publishers, 266 pp.

Karger, D. R., and M. Ruhl, 2002: Finding nearest neighbors in growth-restricted metrics. *Proc. 34th Annual ACM Symp. on Theory of Computing*, Montreal, QC, Canada, ACM, 741–750, https://doi.org/10.1145/509907.510013.

Lamperti, J., 1964: On extreme order statistics. *Ann. Math. Stat.*, **35**, 1726–1737, https://doi.org/10.1214/aoms/1177700395.

Lguensat, R., P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, 2017: The analog data assimilation. *Mon. Wea. Rev.*, **145**, 4093–4107, https://doi.org/10.1175/MWR-D-16-0441.1.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

——, 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

——, 1996: Predictability: A problem partly solved. *Proc. Seminar on Predictability*, Reading, United Kingdom, ECMWF, https://www.ecmwf.int/en/elibrary/10829-predictability-problem-partly-solved.

Lucarini, V., D. Faranda, J. Wouters, and T. Kuna, 2014: Towards a general theory of extremes for observables of chaotic dynamical systems. *J. Stat. Phys.*, **154**, 723–750, https://doi.org/10.1007/s10955-013-0914-6.

——, and Coauthors, 2016: *Extremes and Recurrence in Dynamical Systems*. John Wiley and Sons, 312 pp.

Milnor, J., 1985: On the concept of attractor. *The Theory of Chaotic Attractors*, B. R. Hunt et al., Eds., Springer, 243–264.

Nicolis, C., 1998: Atmospheric analogs and recurrence time statistics: Toward a dynamical formulation. *J. Atmos. Sci.*, **55**, 465–475, https://doi.org/10.1175/1520-0469(1998)055<0465:AAARTS>2.0.CO;2.

Platzer, P., P. Yiou, P. Naveau, P. Tandeo, Y. Zhen, P. Ailliot, and J.-F. Filipot, 2021: Using local dynamics to explain analog forecasting of chaotic systems. *J. Atmos. Sci.*, **78**, 2117–2133, https://doi.org/10.1175/JAS-D-20-0204.1.

Poincaré, H., 1890: Sur le problème des trois corps et les équations de la dynamique. *Acta Math.*, **13**, A3–A270, https://doi.org/10.1007/BF02392506.

Pons, F. M. E., G. Messori, M. C. Alvarez-Castro, and D. Faranda, 2020: Sampling hyperspheres via extreme value theory: Implications for measuring attractor dimensions. *J. Stat. Phys.*, **179**, 1698–1717, https://doi.org/10.1007/s10955-020-02573-5.

Robin, Y., P. Yiou, and P. Naveau, 2017: Detecting changes in forced climate attractors with Wasserstein distance. *Nonlinear Processes Geophys.*, **24**, 393–405, https://doi.org/10.5194/npg-24-393-2017.

Schenk, F., and E. Zorita, 2012: Reconstruction of high resolution atmospheric fields for northern Europe using analog-upscaling. *Climate Past*, **8**, 1681–1703, https://doi.org/10.5194/cp-8-1681-2012.

Van Den Dool, H. M., 1994: Searching for analogues, how long must we wait? *Tellus*, **46A**, 314–324, https://doi.org/10.3402/tellusa.v46i3.15481.

Verleysen, M., and D. François, 2005: The curse of dimensionality in data mining and time series prediction. *Int. Work-Conf. on Artificial Neural Networks*, Warsaw, Poland, ICANN, 758–770.

Wang, X., and S. S. Shen, 1999: Estimation of spatial degrees of freedom of a climate field. *J. Climate*, **12**, 1280–1291, https://doi.org/10.1175/1520-0442(1999)012<1280:EOSDOF>2.0.CO;2.

Wetterhall, F., S. Halldin, and C.-Y. Xu, 2005: Statistical precipitation downscaling in central Sweden with the analogue method. *J. Hydrol.*, **306**, 174–190, https://doi.org/10.1016/j.jhydrol.2004.09.008.

Yiou, P., 2014: AnaWEGE: A weather generator based on analogues of atmospheric circulation. *Geosci. Model Dev.*, **7**, 531–543, https://doi.org/10.5194/gmd-7-531-2014.

——, and C. Déandréis, 2019: Stochastic ensemble climate forecast with an analogue model. *Geosci. Model Dev.*, **12**, 723–734, https://doi.org/10.5194/gmd-12-723-2019.

——, T. Salameh, P. Drobinski, L. Menut, R. Vautard, and M. Vrac, 2013: Ensemble reconstruction of the atmospheric column from surface pressure using analogues. *Climate Dyn.*, **41**, 1333–1344, https://doi.org/10.1007/s00382-012-1626-3.

Young, L.-S., 1982: Dimension, entropy and Lyapunov exponents. *Ergodic Theory Dyn. Syst.*, **2**, 109–124, https://doi.org/10.1017/S0143385700009615.