Model selection: Using information measures from ordinal symbolic analysis to select model sub-grid scale parameterizations

Manuel Pulido *

Department of Physics, FaCENA,

Universidad Nacional del Nordeste (UNNE) and CONICET,

Corrientes, Argentina.

Osvaldo A. Rosso [1,2,3]

[1] *Instituto de Física, Universidade Federal de Alagoas (UFAL)*

Maceió, AL, Brazil.

[2] *Instituto Tecnológico de Buenos Aires (ITBA) and CONICET,*

Ciudad Autónoma de Buenos Aires, Argentina.

[3] *Complex Systems Group, Facultad de Ingeniería y Ciencias Aplicadas,*

Universidad de los Andes,

Las Condes, Santiago, Chile.

**Corresponding author address:* Manuel Pulido, Department of Physics, FaCENA, Universidad Nacional del Nordeste (UNNE), Av. Libertad 5400, Corrientes (3400), Argentina.

E-mail: pulido@exa.unne.edu.ar

ABSTRACT

The use of information measures for model selection is examined to diagnose model physical parameterizations. Although the resolved dynamical equations of atmospheric or oceanic global numerical models are well established, the development and evaluation of parameterizations that represent subgrid-scale effects pose a big challenge. For climate studies, the parameters or parameterizations are usually selected according to a root-mean-square error criterion, that measures the differences between the model state evolution and observations along the trajectory. However, inaccurate initial conditions and systematic model errors contaminate root-mean-square error measures. In this work, information theory quantifiers, in particular Shannon entropy, statistical complexity and Jensen-Shannon divergence, are evaluated as measures of the model dynamics. An ordinal analysis is conducted using the Bandt-Pompe symbolic data reduction in the signals. The proposed ordinal information measures are examined in the two-scale Lorenz'96 system. By comparing the two-scale Lorenz'96 system signals with a one-scale Lorenz'96 system with deterministic and stochastic parameterizations, we show that information measures are able to select the correct model and to distinguish the parameterizations including the degree of stochasticity that results in the closest model dynamics to the two-scale Lorenz'96 system.

1. Introduction

The numerical models for climate predictions and weather forecasts involve a set of dynamical equations which represents the atmospheric or oceanic motions in a grid. Coupled to the resolved dynamical equations of the models, there is a set of parameterizations which represents the subgrid-scale physical processes. The model parameterizations are responsible for a large fraction of model error and thus for the resultant uncertainty associated to climate predictions (see e.g. Stainforth et al. 2005). One major challenge in model development is to decrease model error by recovering aspects of the natural system evolution represented by the parameterizations in the model. However, the actual dynamics of the system is unknown; limited and sparse observations with associated measurement errors is the only source of information of the natural system evolution. The usual procedure for parameterization development and also for inferring unknown parameters is to tune the parameterization or the parameters in order to decrease root-mean-square errors between the model integrations and the observations starting from initial conditions that are close to the natural system state at a given time. For short times, the model state is close to the natural system state, so that model sensitivity should follow natural system sensitivity (Pulido 2014). However, systematic model errors drift the model state from the natural system trajectory for long times (from 5-days); therefore the model and the natural system differ substantially. In this context, observed natural system sensitivity is not useful to constrain model sensitivity, and root-mean-square errors give limited information for model improvement.

Data assimilation techniques have been proposed as a method for estimating model parameters (Ruiz et al. 2013a; Aksoy 2015) and for model development (Pulido et al. 2016; Lang et al. 2016). In a data assimilation system, the model state is recursively pushed towards the observations at the analysis times so that one expects that model sensitivity can be constrained from the observed natural system sensitivity. Under the presence of multiple sources of model errors in a realistic scenario, the estimation of model parameters with data assimilation techniques compensates not only model errors due to the physical process represented in that parameterization but also other sources of model errors. For instance, Ruiz and Pulido (2015) show that estimating the parameters associated with moist processes in an atmospheric general circulation model compensates not only errors from convection but also errors produced by an incorrect representation of boundary layer dynamics. Therefore, the estimated parameters are optimal for that particular combination of model errors and for that particular point of the model state. In other situations, that estimated set of parameters will not represent the natural system sensitivity.

Klinker, and Sardeshmukh (1992) examined the initial tendency errors, the differences between model sensitivity and observed sensitivity during the first time step from the initial conditions. Rodwell and Palmer (2007) show that systematic initial tendency errors can be useful to assess climate models. Errors from different sources should be decoupled at initial times and they should be localized close to the source locations. In a multi-scale system, the errors that dominate at initial times are produced by fast processes. The model sensitivity feedback interactions associated with slow processes are expected to be weak compared with fast processes so that they will not be easily captured by initial tendency errors (Rodwell and Palmer 2007).

The predictability of a dynamical system is quantified by the growth rate of errors as the system evolves. For chaotic systems, a small error in the initial conditions grows as the prediction range increases. The average long-term exponential separation between two trajectories which initially differ by an infinitesimal distance is given by the leading Lyapunov exponent. If the leading Lyapunov exponent is positive, the system is chaotic — errors grow with time. The leading Lyapunov exponent is a possible measure to quantify the predictability of the dynamical system. There is a strong relation between the Shannon entropy and the Lyapunov exponents. For a dynamical system which has a sufficiently smooth probability distribution the Pesin identity holds, the sum of the positive Lyapunov exponents is equal to the Kolmogorov-Sinai entropy (Pesin 1977; Eckmann and Ruelle 1985). In this way, the permutation Shannon entropy can be considered as an upper bound of the Lyapunov exponents (e.g. Bandt and Pompe 2002). Therefore, entropy is also a useful quantity to characterize the predictability in the climate system.

Leung and North (1990) introduce Shannon entropy as a measure of the uncertainty in a climate signal. They examine the similarities between a climate and a communication system. A state in the climate system with large entropy would be unpredictable. There are many possible states that are equally probable. Majda and Gershgoring (2011) propose to use information theory for measuring model fidelity and sensitivity. They use the relative entropy to measure the distance between the probability distribution functions (PDFs) of the natural system and of the numerical model, assuming that both PDFs are Gaussian. Tirabassi and Massoller (2016) use symbolic time-series analysis and mutual lag between time series at different grid points to identify communities in climate data, i.e. sets of nodes densely interconnected in the network.

In the present work, we examine information theory measures as a tool to evaluate numerical models. We extend the concepts introduced by Majda and Gershgoring (2011) to the use of Jensen–Shannon divergence (Grosse et al. 2002) computed with the ordinal symbolic PDFs. This ordinal analysis is conducted using the Bandt and Pompe (2002) symbolic data reduction in the signals, in particular, to determine the corresponding ordinal-based quantifiers, such as normalized Shannon entropy and statistical complexity. They can be used to distinguish different dynamical regimes and to discriminate clearly chaotic from stochastic signals (Rosso et al. 2007, 2012a,b). By comparing information measures from time series of variables of a set of imperfect models with information measures from observed time series, our aim is to find the imperfect numerical model that is closest to the information measures of the natural system.

Information measures of the two-scale Lorenz'96 system (Lorenz 1996) are evaluated using ordinal symbolic analysis as a function of the "physical" parameters of the system: the constant forcing and the interaction coefficient between the slow and fast dynamics. This two-scale system is then considered as the natural system evolution, while the numerical imperfect model is the one-scale Lorenz'96 (Lorenz 1996). We assume the small-scale processes cannot be represented explicitly in this imperfect model, so that the effects of small-scale processes are parameterized as a polynomial function which depends on large-scale variables. The information measures from ordinal symbolic analysis are used to find the most suitable parameterization of the small-scale processes. The information measures of the imperfect model should be as close as possible to the information measure of the "natural system", the two-scale Lorenz'96 system. We evaluate whether the measures are suitable for parameter selection, this is, whether parameter changes have enough sensitivity in the information measures, so that the optimal parameters could be properly inferred from information measures.

Physical parameterizations in atmospheric or oceanic numerical models represent the subgrid-scale physical processes, through functional dependences with the resolved variables. These resolved variables, that the parameterizations depend on, are slow large-scale variables; hence in general the models lack from small-scale variability. Palmer (2001) suggested the use of stochastic parameterizations to account for this lack of variability in the models. There are several works in the last decade that show that both weather forecasts and climate predictions appear to benefit from stochastic parameterizations. For instance, the ensemble prediction system of the European Center for Medium-range Weather Forecasts (ECMWF) uses a stochastic kinetic backscatter algorithm to improve the skill of ensemble forecasting (Shutts 2005). Convection processes have also been proposed to be represented through stochastic parameterizations (Christensen et al. 2015). Some climate features, such as the quasi-biennial oscillation, are better represented in models with stochastic parameterizations (Piani et al. 2004; Lott et al. 2012). Wilks (2005) showed that including a stochastic parameterization in the Lorenz'96 system produces improvements compared to deterministic parameterizations of both the model climatology and ensemble forecast verification measures. Here, we evaluate whether the use of information measures is sensitive to stochastic parameterizations and whether some of the noise variance parameters of stochastic parameterizations may be constrained by trying to reproduce with the model the information measures from the observed time series.

2. Information measures for characterizing model dynamics

Chaotic dynamical systems are sensitive to initial conditions. These manifest instability everywhere in the phase space and lead to non-periodic motion, i.e. chaotic time series (Abarbanel 1996). They are unpredictable in the long term despite the deterministic character of the temporal trajectory. In a system undergoing chaotic motion, two neighboring points in the phase space move away exponentially. Let $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ be two such points, located within a ball of radius R at time t. Further, assume that these two points cannot be resolved within the ball due to observational error. At some later time t' the distance between the points will typically grow to

$$|\mathbf{x}_1(t') - \mathbf{x}_2(t')| \approx |\mathbf{x}_1(t) - \mathbf{x}_2(t)| \; \exp(\Lambda \, |t' - t|), \tag{1}$$

with $\Lambda > 0$ for chaotic dynamics, being Λ the leading Lyapunov exponent. When this distance at time t' exceeds R, the points become observationally distinguishable. This implies that instability reveals some information about the phase-space population that was not available at earlier times (Abarbanel 1996). Thus, under the above considerations chaos can be thought as an *information source*.

The information content of a system is typically evaluated via a PDF, P, describing the characteristic behavior of some measurable or observable quantity, generally a time series $\mathscr{X}(t)$. Quantifying the information content of a given observable quantity is therefore largely equivalent to characterizing its probability distribution. This is often done with the wide family of measures called information theory quantifiers (Gray 1990). We can define information theory quantifiers as measures able to characterize relevant properties of the PDF associated with the time series which can be generated from observations of a dynamical system or from model integrations.

a. Ordinal symbolic analysis

The evaluation of quantifiers derived from information theory, like Shannon entropy and statistical complexity, supposes some prior knowledge about the system; specifically, a probability distribution associated to the time series under analysis should be provided beforehand. Although for a physical quantum system, the concept of probability is uniquely defined; there are several ways to define a probability distribution for a dynamical system. The traditional is the histogram, the state space is partitioned into bins and by counting the number of times N_i that the trajectories of an ensemble pass through the i-bin at a given time, the probability is, in this way, defined as $p_i = N_i/N$, where N is the total number of trajectories. This symbolic sequence can be regarded to as a non causal coarse-grained description of the time series under consideration.

An alternative definition is given with time sequences. Suppose we use a sequence of L time steps and we label the bins, then in L time steps the trajectory passes through L bins, and we can form a *symbolic sequence* of length L. In the symbolic sequence, each symbol from a finite alphabet represents a bin, and the pattern is formed by the sequences of bins, which visits the trajectory in the L time steps. Counting the occurrence of each pattern, over the total number of sequences we determine the probability distribution. If we diminish the size of the bins, in the limit we can derive from this probability the Kolmogorov-Sinai entropy (Schuster and Just 2006).

For some dynamical systems, the information measures determined from bin-symbolic analysis are sensitive to the way the bins are generated (Bollt et al. 2000). Bandt and Pompe (2002) introduced a simple and robust symbolic methodology that takes into account time causality of the time series —a causal coarse-grained methodology— by comparing neighboring values in a time series. In this work, we refer as ordinal symbolic analysis to the Bandt and Pompe methodology. The symbolic data are: *(i)* created by ranking the values of the series; and *(ii)* defined by reordering the embedded data in ascending order, which is equivalent to a phase-space reconstruction with embedding dimension (pattern length) D. In this way, the diversity of the ordering symbols (patterns) derived from a scalar time series is quantified.

The appropriated symbolic sequence arises naturally from the time series, and no system-based assumptions are needed in Bandt and Pompe methodology. In fact, the necessary "partitions" are devised by comparing the order of neighboring relative values rather than by apportioning amplitudes according to different levels. This technique, as opposed to most of those in current practice, takes into account the temporal structure of the time series generated by the physical process under consideration. As such, it allows us to uncover important details concerning the ordinal structure of the time series (Rosso et al. 2007) and can also yield information about temporal correlation (Rosso and Masoller 2009a,b).

The "ordinal patterns" of order (length) D in the Bandt and Pompe methodology are generated by

$$(s) \mapsto \left(x_{s-(D-1)}, x_{s-(D-2)}, \ldots, x_{s-1}, x_s\right) \ , \tag{2}$$

which assigns to each time s the D-dimensional vector of values at times $s-(D-1), \ldots, s-1, s$. By "ordinal pattern" related to the time (s), we mean the permutation $\pi = (r_0, r_1, \ldots, r_{D-1})$ of

$[0,1,\ldots,D-1]$ defined by

$$x_{s-r_{D-1}} \leq x_{s-r_{D-2}} \leq \cdots \leq x_{s-r_1} \leq x_{s-r_0} \,. \tag{3}$$

In this way the vector defined by (2) is converted into a unique symbol π. We set $r_i < r_{i-1}$ if $x_{s-r_i} = x_{s-r_{i-1}}$ for uniqueness, although ties in samples from continuous distributions have null probability.

Then, the occurrence of each symbolic pattern is counted in the whole time series. The probability of each symbol, π_i, is the number of occurrences of the pattern over the total number of analyzed sequences in the time series. The Bandt and Pompe PDF (BP-PDF) is given by $P = \{p(\pi_i), i = 1,\ldots,D!\}$, with

$$p(\pi_i) = \frac{\#\{s | s \leq M-(D-1);\ (s) \text{ is of type } \pi_i\}}{M-(D-1)} \,, \tag{4}$$

where # denotes cardinality and M is the time series length.

In order to illustrate ordinal symbolic analysis, let us consider a simple example: a time series with seven ($M = 7$) values $\mathscr{X} = \{4,7,9,10,6,11,3\}$ and compute the BP-PDF for $D = 3$. In this case, the state space is divided into 3! partitions so that 6 mutually exclusive permutation symbols are considered. The triplets $(4,7,9)$ and $(7,9,10)$ represent the permutation pattern $\{012\}$, since they are in increasing order. On the other hand, $(9,10,6)$ and $(6,11,3)$ correspond to the permutation pattern $\{201\}$ since $x_{t+2} < x_t < x_{t+1}$, while $(10,6,11)$ has the permutation pattern $\{102\}$ with $x_{t+1} < x_t < x_{t+2}$. Then, the associated probabilities to the 6 patterns are: $p(\{012\}) = p(\{201\}) = 2/5$; $p(\{102\}) = 1/5$; $p(\{021\}) = p(\{120\}) = p(\{210\}) = 0$.

The existence of an attractor in the D-dimensional phase space is not required in the ordinal symbolic analysis. The only condition for the applicability of the method is a very weak stationary assumption. For $k \leq D$, the probability for $x_t \leq x_{t+k}$ should not depend on t.

b. Entropy, statistical complexity and Jensen-Shannon divergence

Entropy is a basic quantity with multiple field-specific interpretations. For instance, it has been associated with disorder, state-space volume, and lack of information (Brissaud 2005). When dealing with information content, the Shannon entropy is often considered as the foundational and most natural one (Shannon 1948; Shannon and Weaver 1949). It is a positive quantity that increases with increasing uncertainty and is additive for independent components of a system. From a mathematical point of view, Shannon entropy is the only information measure that satisfies the Kinchin axioms (Kinchin 1957).

Let $P = \{p_i; i = 1, \ldots, N\}$ with $\sum_{i=1}^{N} p_i = 1$, be a discrete probability distribution, with N the number of possible states of the system under study. The "Shannon" logarithmic information measure is defined by

$$\mathrm{S}[P] \; = \; -\sum_{i=1}^{N} p_i \ln(p_i) \; . \tag{5}$$

This can be regarded to as a measure of the uncertainty(lack of information) associated to the physical process described by P. For instance, if $\mathrm{S}[P] = \mathrm{S}_{\min} = 0$, we are in a position to predict with complete certainty which of the possible outcomes i, whose probabilities are given by p_i, will actually take place. Our knowledge of the underlying process described by the probability distribution is maximal in this instance. In contrast, our knowledge is minimal for a uniform distribution $P_e \equiv \{p_i = 1/N, i = 1, \ldots, N\}$ since every outcome exhibits the same probability of occurrence. Thus, the uncertainty is maximal, i.e., $\mathrm{S}[P_e] = \mathrm{S}_{\max} = \ln N$. In the discrete case, we define a "normalized" Shannon entropy, $0 \leq \mathscr{H} \leq 1$, as

$$\mathscr{H}[P] \; = \; \mathrm{S}[P]/\mathrm{S}_{\max} \; . \tag{6}$$

Statistical complexity is often characterized by a complicated dynamics generated from relatively simple systems. Obviously, if the system itself is already involved enough and is constituted by many different parts, it may clearly support a rather intricate dynamics, but perhaps without the emergence of typical characteristic patterns (Kantz et al. 1998). Therefore, a complex system does not necessarily generate a complex output. Statistical complexity is therefore related to structures hidden in the dynamics, emerging from a system which itself can be much simpler than the dynamics it generates (Kantz et al. 1998).

We follow the original idea for statistical complexity introduced by López-Ruiz et al. (1995). A suitable complexity measure should vanish both for completely ordered and for completely random systems and it cannot only rely on the concept of information (which are maximal and minimal for the above mentioned systems). It can be defined as the product of a measure of information and a measure of disequilibrium, i.e. some kind of distance from the equiprobable distribution of the accessible states of a system (López-Ruiz et al. 1995; Lamberti et al. 2004).

The statistical complexity measure to be used here (Lamberti et al. 2004; Rosso et al. 2007) is defined through the functional product form

$$\mathscr{C}[P] \;=\; Q_{JS}[P,P_e] \cdot \mathscr{H}[P] \tag{7}$$

of the normalized Shannon entropy $\mathscr{H}$, see (6), and the disequilibrium Q_{JS}. It is defined in terms of the Jensen-Shannon divergence $D_{JS}[P,P_e]$,

$$Q_{JS}[P,P_e] \;=\; Q_0 \cdot D_{JS}[P,P_e] \;=\; Q_0 \cdot \{\mathrm{S}[(P+P_e)/2] - \mathrm{S}[P]/2 - \mathrm{S}[P_e]/2\}, \tag{8}$$

where Q_0 is equal to the inverse of the maximum of $D_{JS}[P,P_e]$ which is obtained when one of the components of P is one and the remaining are zero. Therefore, the disequilibrium Q_{JS} measures the normalized distance of the probability distribution of the system under study P and the uniform distribution P_e which is the equilibrium PDF.

For a given value of $\mathscr{H}$, the range of possible $\mathscr{C}$ values varies between a minimum $\mathscr{C}_{min}$ and a maximum $\mathscr{C}_{max}$, restricting the possible values of the statistical complexity measure (Martín et al. 2006). The planar representation entropy-complexity plane, $\mathscr{H} \times \mathscr{C}$, is an efficient tool to distinguish between the deterministic chaotic and stochastic nature of a time series since the permutation quantifiers have distinctive behaviors for different types of dynamics (Rosso et al. 2007). This tool has also been used for visualization and for a characterization of different dynamical regimes when the system parameters vary (Zanin et al. 2012).

Finally, we consider a measure for model evaluation against the observed time series. A measure of the distance between the probabilities from the model and observed time series. This concept has been used earlier by Majda and Gershgoring (2011) who called it model fidelity. They use the Kullback-Leibler relative entropy to measure the distance between the two probabilities. Arnold et al. (2013) evaluated the use of Hellinger distance. They found similar results using the Hellinger distance and Kullback-Leibler distance in the Lorenz'96 system. We use the Jensen-Shannon divergence to measure the distance between the probabilities to be coherent with the information theory quantifiers used in this work and because it is a symmetric positive-definite quantity. The square-root of the Jensen-Shannon divergence satisfies metric properties and triangle inequality (Lin 1991).

Assuming P_M and P_O are the corresponding BP-PDFs from the model time series and from the observed time series respectively, the Jensen-Shannon divergence is defined as a symmetric measure of the Kullback-Leibler divergence,

$$D_{JS}[P_M, P_O] \; = \; \sum p_i^M \ln(p_i^M/p_i^O) + p_i^O \ln(p_i^O/p_i^M) \; = \; \sum (p_i^M - p_i^O) \ln(p_i^M/p_i^O), \qquad (9)$$

it vanishes when $p_i^M = p_i^O$ for all i. It can also be expressed in terms of the Shannon entropy (5):

$$D_{JS}[P_M, P_O] \; = \; S[(P_M + P_O)/2] - S[P_M]/2 - S[P_O]/2\,. \qquad (10)$$

To evaluate (10), we determine the probability of the observed time series P_O and of the different model time series P_M using ordinal symbolic analysis. The Jensen-Shannon divergence is a measure of distance between two PDFs, P_M and P_O, so that a small Jensen-Shannon divergence indicates a model PDF close to the observed PDF. The best model or the optimal parameters are the ones whose the time series gives the smallest Jensen-Shannon divergence.

3. Description of the numerical experiments

In the numerical experiments, we evaluate the potential of ordinal symbolic analysis to select subgrid-scale parameterizations using the integration of the two-scale Lorenz'96 system (Lorenz 1996) as the natural system evolution. The equations of this system are given by a set of N equations of large-scale variables X_n,

$$\frac{\mathrm{d}X_n}{\mathrm{d}t} + X_{n-1}(X_{n-2} - X_{n+1}) + X_n \; = \; F - \frac{h\,c}{b} \sum_{j=(M/N)(n-1)+1}^{nM/N} Y_j \; ; \tag{11}$$

where $n = 1, \ldots, N$; and a set of M equations of small-scale variables Y_m, given by

$$\frac{\mathrm{d}Y_m}{\mathrm{d}t} + c\,b\,Y_{m+1}(Y_{m+2} - Y_{m-1}) + c\,Y_m \; = \; \frac{h\,c}{b}\, X_{\mathrm{int}[(m-1)/(M/N)]+1} \; ; \tag{12}$$

where $m = 1, \ldots, M$. Note that both sets of equations (Eqs. (11) and (12)) are in a periodic domain, that is $X_0 = X_N$, $X_{-1} = X_{N-1}$ and $Y_0 = Y_M$, $Y_1 = Y_{M+1}$, $Y_2 = Y_{M+2}$.

Equations (11) and (12) are essentially the same but with different scales. They have coupling terms between them, the equations of small-scale variables, (12), are forced by the local (closest) large-scale variable. The equations of large-scale variables, (11), are forced by the external forcing F, and by the averaged small-scale variables which are located around the large-scale variable in consideration.

Lorenz (1996) suggested this simple model as a one-dimensional atmospheric model with two distinct time scales in a latitudinal circle with interactions between the two scales and he used it to illustrate atmospheric predictability issues. In the experiments, we use the standard set of constants: $N = 8$, $M = 256$, coupling constant $h = 1$, time-scale ratio $c = 10$, and spatial-scale ratio $b = 10$ (unless stated otherwise). Note that setting $h = 0$ in (11), we recover the one-scale Lorenz'96 system.

In reality, the atmospheric numerical models cannot represent the small-scale variables associated with convection processes, small-scale waves, etc., so that the effects of the small-scale variables on the large-scale equations must be parameterized in the numerical models through forcing terms with functional dependencies of only the large-scale variables and a set of free parameters. Thus, the equations of the *imperfect model* are

$$\frac{\mathrm{d}X_n^M}{\mathrm{d}t} + X_{n-1}^M(X_{n-2}^M - X_{n+1}^M) + X_n^M \;=\; G_n(X_n^M, a_0, \cdots, a_J)\;; \tag{13}$$

where $n = 1, \ldots, N$ and X_n^M represents the variables of the imperfect model. The function $G_n(X_n^M, a_0, \cdots, a_J)$ is a parameterization of the small-scale processes and the forcing term, it seeks to mimic the right hand side term of (11). The a_j are free parameters.

Two representations of the forcing term are examined in this work: *a)* a *deterministic parameterization* given by a polynomial function,

$$G_n(X_n^M,\; a_0, \cdots, a_J) \;=\; \sum_{j=0}^{J} a_j \cdot (X_n^M)^j\;; \tag{14}$$

and *b)* a *stochastic parameterization* defined in Wilks (2005) by a polynomial function and a stochastic component given by realizations of a first-order autoregressive process

$$G_n(X_n^M,\; a_0, \cdots, a_J, \sigma, \phi) \;=\; \sum_{j=0}^{J} a_j \cdot (X_n^M)^j \;+\; \eta_n(t)\;; \tag{15}$$

where

$$\eta_n(t) \;=\; \phi\; \eta_n\,(t - \Delta t) \;+\; \sigma\,(1 - \phi^2)^{1/2}\; \nu_k(t)\;, \tag{16}$$

ϕ is the autoregressive parameter, ν_k is a realization of a normal distribution with zero mean and unit variance, and σ is the standard deviation of the process. Both ϕ and σ, apart from a_j, are free parameters.

The Lorenz'96 system was integrated using a Runge-Kutta of fourth order, with an integration step of $\delta = 0.001$. In what follows the time resolution of the time series or the observational time resolution is taken to be $\delta = 0.05$, which considering the growth rates of the system, it represents 6 hours in the atmosphere and so it is able to capture the instability growth (Lorenz 1996). To avoid spin-up behavior, the state is started from a random initial condition and it is integrated by 10^5 observational times (this corresponds to $5 \cdot 10^6$ time steps). The resulting state is used as the initial condition and it is integrated further by $N_d = 10^5$ observational times (i.e. N_d is the time series length) which are used to compute the information measures.

In order to evaluate the imperfect model, we use an "observed" time series of a single large-scale variable from the natural system evolution, the two-scale Lorenz'96 system. This is, we assume that the large-scale is the only information observed so that signals from a single large-scale variable are used in the ordinal symbolic analysis. The small-scale dynamics is neither modeled nor observed, except in the "true" state integration which is conducted with the two-scale Lorenz'96 and considered as the natural system trajectory.

In all the experiments, we use the ordinal symbolic analysis to determine BP-PDFs associated with the time series of the dynamical system and then the information quantifiers, normalized Shannon entropy (6), statistical complexity (7) and Jensen-Shannon divergence (10), are computed. The length of the pattern for the ordinal analysis is taken to be $D = 6$. This gives a total of $D! = 6! = 720$ possible ordinal symbolic patterns, which clearly satisfy the condition $N_d \gg D!$ for robust statistics (Rosso et al. 2007). The choice of the length of the pattern is a compromise decision, a longer D gives a more casual and higher resolution PDF, but it requires a longer time series for accurate statistics. We took $D = 6$ as in Rosso et al. (2007); Serinaldi et al. (2014). However, note that because of the short climate time series available, Tirabassi and Massoller (2016) used $D = 3$ for monthly climate time series with meaningful results.

In a first set of experiments, we explore the two-scale Lorenz'96 system with the information quantifiers: Shannon entropy (6) and statistical complexity (7). Different dynamical regimes are uncovered as the forcing and the coupling coefficient are varied.

A second set of experiments focuses on model fidelity, in which we determine the BP-PDFs of the observed time series P_O and of the modeled time series P_M, and so (10) is evaluated. Observed and modeled time series are completely independent including the initial condition. They are both assumed to be on the attractor of the dynamical system (after the spin-up integration). The synthetic observed time series is in the second set of experiments generated with an integration of the one-scale Lorenz '96 system and a set of prescribed parameter values. Then we can evaluate the sensitivity of the information quantifiers to the model parameters for integration of the one-scale Lorenz '96 system with different parameter values. In particular, we expect a minimum in the Jensen-Shannon divergence when the model parameters are set at the "true" values (the ones used to generate the observations). The evaluated parameterizations in this perfect model framework are a deterministic parameterization, which consists of a quadratic polynomial function (14), and a stochastic parameterization, which consists of a quadratic polynomial function and a first-order autoregressive process (15).

To estimate the optimal parameter values, a genetic algorithm was implemented (Charbonneau 2002; Pulido et al. 2012). As in Pulido et al. (2012), the genetic algorithm is able to find the global minimum even in the presence of multiple local minima, however it presents slow convergence. Therefore, only 5 generations were evolved. Then, the newUOA optimization (Powell 2006) was applied, using as initial guess parameters the ones estimated with the genetic algorithm. It is an unconstrained minimization algorithm which does not require derivatives. Both, the genetic algorithm and newUOA are suitable for control spaces of up to a few hundred dimensions. The Jensen-Shannon divergence is used as the minimization function in the optimization method.

The third set of experiments explores the Jensen-Shannon divergence for imperfect models. In this case the observed time series is obtained from a 'nature' integration of the two-scale Lorenz '96 system, and we seek to reproduce the dynamics of the system with integrations of imperfect models generated from one-scale Lorenz '96 systems with deterministic and stochastic parameterizations. From these experiments we determine a set of optimal values using the mentioned optimization method for a deterministic and stochastic parameterization that seek to represent the small-scale dynamical effects of the two-scale Lorenz '96 system. These optimal parameter values are used in long-term climate prediction experiments to examine whether the optimal parameters have a positive impact on climate measures.

4. Results and discussion

a. Experiments with the two-scale Lorenz '96 system

First, the ordinal symbolic analysis is applied to the integration of what we consider as the natural system evolution, the two-scale Lorenz '96 system. Integrations varying the forcing F were conducted with a resolution of $\delta F = 0.01$, and the ordinal symbolic analysis is applied to each integration (i.e. time series of the Lorenz '96 variable X_1). Figure 1 shows the information quantifiers: permutation entropy ($\mathscr{H}$, Fig. 1a), permutation statistical complexity ($\mathscr{C}$, Fig. 1b). From Figs. 1a and 1b, four regions with different dynamical regimes are found (which are delimited by vertical dotted lines): *i)* For small external forcing, $0 \leq F \leq 3.75$, the system is dissipative and so after the spinup time the entropy goes to zero. *ii)* A narrow region, between $3.75 < F < 4.5$, with high permutation entropy and high permutation statistical complexity. *iii)* An intermediate region, between $5 < F < 12$, with small entropy $\mathscr{H} \approx 0.2 - 0.23$ and similar complexity. *v)* Finally a region for larger F, $F > 13$, which has large entropy $\mathscr{H} > 0.4$ but relatively small complexity ($\mathscr{C} < 0.4$).

Figure 1c shows the causal entropy-complexity plane ($\mathscr{H} \times \mathscr{C}$) which combines the entropy and statistical complexity measures. In this plane, the statistical complexity has a minimum and maximum value as a function of entropy ($\mathscr{C}_{min}$ and $\mathscr{C}_{max}$ respectively), which are the upper and lower continuous curves in Fig. 1c, so that all the possible dynamical regimes are limited to the area between these curves. The four dynamical regimes can be clearly distinguished in the entropy-complexity plane. The dissipative regime is located at the extreme of null entropy and complexity. The quasi-periodic dynamical regime (iii) with low entropy and maximal statistical complexity is denoted by the triangles that are close to the upper curve which represents the maximal statistical complexity. The large F chaotic regime (iv) which has large entropy and relatively small complexity is represented with gray circles. The gray triangles correspond to the narrow region between $3.75 < F < 4.5$ with large entropy and maximal complexity (at the $\mathscr{C}_{max}$ curve). Since the system is purely deterministic, there are no dynamical regimes in the large entropy region, close to $\mathscr{H} = 1$, which would represent a purely stochastic system (Rosso et al. 2007).

Figure 2 shows the time series, resulting from the dynamical regimes obtained from the two-scale Lorenz '96 dynamical system (except the dissipative regime) identified using the information quantifiers, for $F = 4$ (Figure 2a), $F = 7$ (Figure 2b) and $F = 18$ (Figure 2c). These represent quasi-periodic motion with high entropy, quasi-periodic motion with low entropy and chaotic motion, respectively.

Figure 3 shows the information quantifiers from integrations of the two-scale Lorenz '96 system varying the coupling constant h. The external forcing is fixed to $F = 4, 6, 18$. For $h \to 0$ we recover the measures for the one-scale Lorenz '96 system since the two sets of equations, (11) and (12), are uncoupled. In that case, the permutation entropy and the permutation statistical complexity scales with the forcing. For $F = 4$, there is a peak of entropy and complexity when the coupling constant h is close to 1, which was the regime already found in Fig. 1 with complexity close to $\mathscr{C}_{max}$ (note that in those integrations $h = 1$). As we increase F to 6, the large complexity regime is found for larger coupling between the two scales, for h between 1.4 and 2. On the other hand, small entropy and complexity is found for the $F = 18$ for coupling constants between 1 and 2. For larger coupling constants, a regime with high disordered patterns is found (small complexity and large entropy). For coupling constants close to 5, a regime with high statistical complexity appears to emerge for the $F = 18$ but we did not explore integrations for larger coupling constants. Some of the dynamical regimes that appear to emerge from the Lorenz '96 system varying the coupling constant and varying stochastic noise will be investigated further in a follow-up work.

b. Perfect-model experiments

To evaluate the potential of information quantifiers to distinguish between time series generated with different parameterizations, we conducted a so-called twin experiment. We consider the one-scale Lorenz '96 system, (13), with a known parameterization as the natural system evolution to generate the observed time series and then we evaluate the information measures for integrations of the one-scale Lorenz '96 system with varying parameters using the hybrid optimization algorithm, with genetic algorithm and newUOA methods. This is an experiment where the model is assumed perfect, and a set of prescribed parameters are used to generate the observations. Then, the optimization method is used to estimate the parameters through the differences in the observed and modeled time series. In this way, we can evaluate whether the Jensen-Shannon divergence measure determined with the ordinal symbolic analysis is able to estimate the "true" parameters.

The first perfect model experiment uses a deterministic quadratic parameterization, (14). The true parameter values are set to $a_0^t = 17.0$, $a_1^t = -1.20$, $a_2^t = 0.035$ (t superscript denotes true values). These values are expected to be a representative deterministic parameterization of the two-scale model (Pulido et al. 2016). The integration with the true parameters is considered as the observational time series. The Jensen-Shannon divergence, (10), and Hellinger distance (Arnold et al. 2013) are minimized through the hybrid optimization algorithm which seek for the optimal model parameter values. The symbolic ordinal analysis is applied to each model and observational time series to evaluate the Jensen-Shannon divergence, (10), and Hellinger distance. The optimal parameter values obtained with the hybrid optimization algorithm were $a_0 = 17.1$, $a_1 = -1.18$ and $a_2 = 0.032$ using the Jensen-Shannon divergence and $a_0 = 16.9$, $a_1 = -1.17$ and $a_2 = 0.031$ using the Hellinger distance. Both distance measures give accurate estimates. In all preliminary experiments both distance measures performed similarly well, so that in what follows we only show the experiments with Jensen-Shannon divergence. This twin experiment shows that the information measures can be used to determine optimal parameters, the estimated optimal values are very close to the true parameter values.

The sensitivity in the Jensen-Shannon divergence to the parameters is shown in Fig. 4 varying each of the parameters and the other two parameters are fixed to the optimal values (which were obtained with the hybrid optimization method using Jensen-Shannon divergence). The optimal parameter is very well defined in the three parameters. The minimum of the Jensen-Shannon divergence is clearly located at the true parameters. One weak point of the measure is that it presents noise, including several local extremes. This affects the convergence speed of optimization methods.

A second perfect-model experiment takes a stochastic parameterization, (15), for the polynomial coefficients we use the same true values as in the previous experiment, $a_0^t = 17$, $a_1^t = -1.2$, $a_2^t = 0.035$ but we now include a noise forcing term with standard deviation $\sigma^t = 1$. Two optimization experiments with autoregressive parameters $\phi^t = 0$ and $\phi^t = 0.984$ were conducted. These two extreme values were taken by Wilks (2005) to represent serially independent and serially persistent stochastic forcing, respectively. The resulting optimal parameter values of the hybrid optimization algorithm are shown in Table 1. The combined estimation of deterministic parameters and the stochastic parameter σ gives rather good estimates. The stochastic parameter is slightly underestimated 10-20% in the two optimization experiments.

Once the optimal parameters for the stochastic parameterization are estimated, we then evaluate the sensitivity of Jensen-Shannon divergence measure with respect to this observational time series varying σ values in the model integrations. Figure 5 depicts the Jensen-Shannon divergence as a function of σ parameter for autoregressive parameters of $\phi^t = 0$ and $\phi^t = 0.984$ (the other parameters are fixed to the optimal values that were estimated with the hybrid optimization algorithm). A rather narrow negative peak is found in Fig. 5 close to the true parameter values. The $\phi^t = 0.984$ case (Fig. 5b) appears to be better conditioned.

c. *Imperfect-model experiments*

The usual procedure to infer unknown parameters of a parameterization scheme in an imperfect, coarse-grained, model is to tune the unknown parameters and to evaluate the response of the changes in the parameters on the root-mean-square error, which measures the differences between the evolution of some representative variables and the corresponding observed variables (or reanalysis data). The optimal parameters are the ones that minimize the root-mean-square error.

We conducted a similar experiment with synthetic observations but using information measures, i.e. Jensen-Shannon divergence, instead of root-mean-square error measures. The advantage of the ordinal symbolic analysis is that as it does not depend on the amplitude but on the "shape" of the patterns, it is not sensitive to possible systematic model errors. The analysis is performed in a sufficiently long trajectory (10^5 observational times). The probability of all the possible patterns is composed by a large number of cases and it is expected to be independent of the initial condition (the spin-up time is not considered in the statistics). The observed time series corresponds to a single variable taken from a model integration of the two-scale Lorenz '96 system which is started from random initial conditions and the spinup period is removed. The model time series is also generated from random initial conditions and integrating the one-scale Lorenz '96 system. Therefore, the two time series are completely independent—they do not have a common initial condition. In this sense, the Jensen-Shannon divergence is a global measure of the system dynamics.

Since we deal with an imperfect model, which does not represent explicitly the small-scale dynamics, the parameter estimation is not a twin experiment in which we know the "true" optimal parameters, so that the existence of a single set of optimal parameters is not a priori ensured.

We conducted two extreme experiments, one with the natural system evolution set for an external forcing of $F = 7$, which results in quasi-periodic motion, and the other for a forcing of $F = 18$, which results in chaotic dynamical behavior. As mentioned, the ordinal symbolic analysis may be applied to chaotic and quasi-periodic time series as long as the weak stationary assumption is satisfied.

The hybrid optimization algorithm was applied to the two observed time series. The genetic algorithm restricts the search for optimal values to the region delimited by the maximum and minimum values stated in Table 2. The parameter limits (maximum and minimum values) of the search region were taken according to the values obtained by Pulido et al. (2016). In the case that the resulting optimal value of the genetic algorithm is at a boundary of the region, it is an indicative that the region is too narrow in that parameter and that the limit value should be changed. The estimated optimal values with the hybrid optimization algorithm for $F = 7$ and $F = 18$ are also shown in Table 2.

The Jensen-Shannon divergence sensitivity to each of the parameter values for the case $F = 7$, varying one parameter value and fixing the other two to the optimal values which resulted from the hybrid optimization algorithm, is shown in Fig. 6. Parameters exhibit strong sensitivity in a small region close to the optimal values. These sensitivity experiments are produced after the optimization, with independent integrations that are not related to the optimization method. For some parameter values, the Lorenz '96 model presents numerical instabilities. A uniform time series is assigned for these cases and so a delta PDF results, which in turn gives a large Jensen-Shannon divergence.

The sensitivity of the Jensen-Shannon divergence to each of the parameters for the case of $F = 18$ is shown in Fig. 7, while the other two parameters are fixed at the optimal values. The parameter a_0 exhibits a reasonable sensitivity around the optimal value. On the other hand, a_1 and a_2 show several peaks so that they are more difficult to be precisely estimated, however, the genetic algorithm is clearly able to find the global minimum even in the presence of these local minima (Fig. 7c).

As the information quantifiers give useful information on the optimal parameter values of the deterministic parameterization, we now turn our attention to stochastic parameterizations for the imperfect case. We include the first-order autoregressive process (16) in the parameterization (15), and search with the hybrid optimization algorithm for the optimal parameter values including the optimal standard deviation σ, (a_0, a_1, a_2, σ), and again we only explore for two fixed autoregressive parameters $\phi = 0$ and $\phi = 0.984$. The resulting optimal parameter values of the hybrid optimization algorithm are shown in Table 2.

The Jensen-Shannon divergence as a function of the standard deviation is depicted in Fig. 8 for the optimal deterministic parameter values (shown in Table 2). For an external forcing of $F = 7$ a smooth function is found with a clear minimum (see Fig. 8a). The minimum is found at $\sigma = 0.32$ for $\phi = 0$. Similar values of the Jensen-Shannon divergence are found at $\sigma = 0.15$ for $\phi = 0.984$. Both sets of values are suitable for representing the stochastic process that mimics the effects of Lorenz'96 small-scale variables. Note that the Jensen-Shannon divergence for the optimal σ value is smaller than the one for $\sigma = 0$ so that the stochastic parameterization improves the representation of small-scale variables. This is also valid when both deterministic and stochastic parameterizations have their own optimal parameters.

For $F = 18$ Jensen-Shannon divergence has smaller values than $F = 7$. This means that the parameterization is able to represent better the effects of the small-scale variables for this case due to the chaotic dynamics. The divergence depicts a noisy dependence, but a constrained optimal range of the standard deviation is still clearly identified from Fig. 8b. The minimum is at $\sigma = 4.67$ for the $\phi = 0$ experiment. Similar Jensen-Shannon divergence values are also found for the $\phi = 0.984$ experiment with a minimum at $\sigma = 2.12$.

To evaluate the information quantifiers as a method for model selection. We conducted an experiment in which we assume that the model has different parameterizations, changing the order of the polynomial function in the deterministic parameterization and for some experiments adding the stochastic process (16). A total of eight optimization experiments with different parameterizations were conducted for an observed time series taken from the two-scale Lorenz '96 system with $F = 18$. For each parameterization, the set of optimal parameters estimated by the hybrid optimization algorithm are stated in Table 3. The square root of Jensen-Shannon divergence for the optimal parameters is also shown in the Table. The best parameterization is the one that gives the minimal Jensen-Shannon divergence from the observed PDF. The quadratic polynomial parameterization is the best deterministic one. Interestingly, the stochastic parameterizations present a significantly better performance with this information measure. The higher-order polynomial terms are very sensitive to small changes in the variables and parameters and for some parameter values they produce numerical instabilities in the Lorenz '96 model (Pulido et al. 2016). Indeed, the optimization experiment with the fourth-order polynomial stochastic parameterization did not converge towards optimal parameter values because of these ubiquitous numerical instabilities (to overcome this, careful manual changes in the parameter limit values would be required).

The forcing given by the parameterizations with optimal parameters for the $F = 18$ experiments, including the quadratic deterministic, and the quadratic stochastic parameterizations with $\phi = 0$ and with $\phi = 0.984$ are shown in Fig. 9 (Panels (a), (b) and (c) respectively). The forcing given by the small-scale variables in the two-scale Lorenz '96 is also shown in the Figure (gray dots). We emphasize, this "true" forcing is only shown as the purpose of evaluation of the optimization experiments, but the time series of a single large-scale state variable is the only source of information used in the optimization experiments. The simple polynomial parameterizations with fixed standard deviation represent rather well the complex forcing dependencies given by the small-scale variable. However, they are obviously unable to represent the dependence of the standard deviation with the value of the state variable particularly at the tail (large X values) and with the $dX/dt > 0$ and $dX/dt < 0$ branches of the forcing, see Crommelin and Vanden-Eijnden (2008); Pulido et al. (2016).

As an independent measure of the climatology of the model with optimal parameters, we use the classical histogram PDF. They were computed from the whole integration with the different optimal parameter values. Figure 10 shows the histogram PDF for the nature integration for $F = 18$ and the ones with the optimal parameters for the quadratic deterministic parameterization (dashed line) and for the stochastic parameterizations using $\phi = 0$ (dotted line) and $\phi = 0.984$ (gray line). A very good agreement between the true histogram PDF and the model PDF is achieved. The stochastic parameterizations give a slightly better agreement to the true histogram PDF.

5. Conclusions

Ordinal symbolic analysis only depends on the repetition of patterns within a time series. If it is combined with information measures, they represent a useful framework to evaluate models, in particular unresolved processes of multi-scale models. Since ordinal symbolic analysis does not depend directly on the state, the quantities can be used for long time intervals (time series) even in the presence of model error. The ordinal symbolic analysis is used in this work for long time series and it accounts for the model fidelity with strong sensitivity to the parameters of the subgrid parameterization which represents the small-scale processes.

Although stochastic parameterizations appear to give improvements in the atmospheric numerical models, the tuning of stochastic parameters represents a challenge. On-line parameter estimation techniques as Kalman filtering present difficulties estimating these stochastic parameters even for small and intermediate systems. DelSole and Yang (2010) show that it is not possible to constrain stochastic parameters with ensemble-based Kalmar filter augmenting the model state with the stochastic parameters. Ruiz et al. (2013b) show that a separate adaptive inflation treatment is required for the parameter covariance to avoid its collapse. Pulido et al. (2016) show that the time variability given by Kalman filtering parameter estimates is not useful to constrain stochastic parameters in a subgrid parameterization. In this work, we show that information measures from ordinal symbolic analysis are useful for tuning stochastic parameters with promising results.

This work evaluates the sensitivity of the parameters to the information measures, which is useful for model selection. Furthermore, for parameter optimization a hybrid optimization technique using a genetic and newUOA algorithms was implemented in this work for low dimensional models. For some cases, the information measures based on the ordinal symbolic analysis do not give smooth dependencies with the parameters. This may be a problem for traditional gradient descent optimization methods. For parameter estimation in high-dimensional models more sophisticated optimization techniques suitable for noisy cost functions, like simulated annealing, are required to minimize the Jensen-Shannon divergence for the probability distributions of observations and an imperfect model. The evaluation of optimization techniques in high-dimensional models with information measures will be examined in a follow-up work.

The proposed parameter estimation method offers an alternative framework to complex data assimilation systems which couple model state to observations. On the other hand, in the proposed method the model time series is generated independently of the observed state of the system. The model state is assumed to be in its own model attractor (which is not necessarily the one from nature). Only partial observation of the system is needed, indeed the observed time series may be a single relevant variable or a small set of variables. The information measures could be applied to a set of *free* integrations from different climate models or a set of *free* integrations from a single climate model with different parameterizations or parameters, to evaluate from an observed time series, which climate model or parameterization give the most accurate results—the closest PDF to the observed PDF.

This work evaluates the information measures with the Lorenz'96 system, which is a small model with 8 – 256 variables. Two major points need to be evaluated with more realistic models, the impact of a higher-dimensional state space on the information measures, and the length of the time series needed to compute the probability distributions. The length of the time series used in this work would represent about 70 years in the atmospheric time scale. It depends on two factors, the required time resolution and the length of the pattern used for the ordinal symbolic analysis. The time resolution used in this work is related to the time-scale of the resolved large-scale processes, and indeed the used time series corresponds to a large-scale variable. The length of the sequence is taken to be six in this work, as used in other applications Sippel et al. (2016); Serinaldi et al. (2014). However, Tirabassi and Massoller (2016) used three for monthly climate time series (which are of limited length) with meaningful results. The way to combine the information measures of different variables for high-dimensional problems needs to be explored.

The information measures can deal with weak observational noise (Rosso et al. 2007), however as expected Shannon entropy gives a maximum if the time series is stochastic without correlations— completely dominated by white noise. For the cases with strong observational noise, the signal may not be useful for analyzing fast processes, but averaging the time series and applying ordinal symbolic analysis in longer time steps may give useful information for slower physical processes.

Acknowledgments. We acknowledge CONICET Grant PIP 112-20120100414CO which funded a two-week visit of one of us (MP) to the Universidade Federal de Alagoas.

References

Abarbanel H. D. I., 1996: *Analisis of observed chaotic data.* Springer-verlag, New York.

Aksoy, A., 2015: Parameter Estimation. In: G. R. North (editor-in-chief), J. Pyle and F. Zhang (editors). *Encyclopedia of Atmospheric Sciences*, 2nd edition, **4**, 181–186.

Arnold H.M., Moroz I.M., Palmer T.N., 2013: Stochastic parametrizations and model uncertainty in the Lorenz 96 system. *Phil Trans R Soc A,* **371**, 20110479. http://dx.doi.org/10.1098/rsta.2011.0479

Bandt, C. and Pompe, B., 2002: Permutation entropy: a natural complexity measure for time series. *Phys. Rev. Lett.*, **88**, 174102.

Bollt E., T. Stanford, Y.-C. Lai, and K. Zyczkowski, 2000: Validity of threshold-crossing analysis of symbolic dynamics from chaotic time series, *Phys. Rev. Lett.* **85**, 3524.

Brissaud J. B., 2005: The meanings of entropy. *Entropy*,**7**, 68-96.

Charbonneau P. 2002. An introduction to genetic algorithms for numerical optimization. Technical Note TN-450+IA. NCAR: Boulder, MA.

Christensen, H.M., I. M. Moroz and T. N. Palmer, 2015: Stochastic and perturbed parameter representations of model uncertainty in convection parameterization. *J. Atmos. Sci.*, **72**, 2525–2544.

Crommelin, D. and Vanden-Eijnden, E., 2008. Subgrid-scale parameterization with conditional Markov chains. *J. Atmos. Sci.*, **65**, 2661-2675.

DelSole, T., and X. Yang., 2010: State and parameter estimation in stochastic dynamical models. *Physica D*, **239**, 1781–1788.

Eckmann J. -P., Ruelle D., 1985: Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, **57**, 617–656.

Gray, R.M., 1990: *Entropy and Information Theory.* Springer, Berlin-Heidelberg, Germany.

Grosse I., Bernaola-Galván P., Carpena P., Román-Roldán R., Oliver J. and Stanley H.E., 2002: Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E*, **65** 041905.

Kantz H., Kurths J., Meyer-Kress G., 1998: *Nonlinear Analysis of Physiological Data*, Springer, Berlin

Kinchin I., 1957: *Mathematical foundations of the information theory*. Dover Publ. New York.

Klinker, E. and P. Sardeshmukh, 1992: The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atmos. Sci.*, **49**, 608-627.

Lang M., P. J. Van Leeuwen, and P. Browne, 2016: A systematic method of parameterisation estimation using data assimilation. *Tellus*, **68**, 29012.

Lamberti, P. W., Martín, M. T., Plastino, A., and O. A. Rosso, 2004: Intensive entropic non-triviality measure. *Physica A*, **334**, 119–131.

Leung, L.-Y., and G. R. North, 1990: Information theory and climate prediction. *J. Climate*, **3**, 5–14.

Lin J 1991: Divergence measures based on the Shannon Entropy *IEEE Transactions on Information Theory*, **37**, 145-151.

López-Ruiz R., Mancini H.L., Calbet X., 1995: A statistical measure of complexity. *Phys. Lett. A*, **209**, 321–326.

Lorenz, E. N., 1996: Predictability— A problem partly solved. In *Proc. Seminar on Predictability*, Shinfield Park, Reading, United Kingdom, European Centre for Medium-Range Weather Forecasting, 1-18.

Lott, F., L. Guez, and P. Maury, 2012: A stochastic parameterization of non-orographic gravity waves: Formalism and impact on the equatorial stratosphere. *Geophys. Res. Lett.* **39**, L06807, doi:10.1029/2012GL051001.

Majda, A.J. and Gershgorin, B., 2011: Improving model fidelity and sensitivity for complex systems through empirical information theory. *Proceedings of the National Academy of Sciences*, **108**, 10044–10049.

Martín M.T., Plastino A., Rosso O.A., 2006: Generalized statistical complexity measures: Geometrical and analytical properties. *Physica A*, **369**, 439–462.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Q. J. Roy. Meteor. Soc.*, **127**, 279–304.

Pesin Ya. B., 1977: Lyapunov characteristic exponents and smooth ergodic theory. *Usp. Mat. Nauk.*, **32**, 55.

Piani, C., W. A. Norton, and D. A. Stainforth, 2004: Equatorial stratospheric response to variations in deterministic and stochastic gravity wave parameterizations. *J. Geophys. Res.*, **109**, 1984–2012.

Powell, M.J., 2006. The NEWUOA software for unconstrained optimization without derivatives. In *Large-scale nonlinear optimization*. 255-297. Springer.

Pulido M, Polavarapu S, Shepherd TG, Thuburn J. 2012. Estimation of optimal gravity wave parameters for climate models using data assimilation. *Quart. J. R. Meteorol. Soc.*, **138**, 298-309.

Pulido, M, 2014: A simple technique to infer the missing gravity wave drag in the middle atmosphere using a general circulation model: Potential vorticity budget. *J. Atmos. Sci.*, **71**, 683-696.

Pulido M., G. Scheffler, J. Ruiz, M. Lucini and P. Tandeo, 2016: Estimation of the functional form of subgrid-scale schemes using ensemble-based data assimilation: a simple model experiment. In press *Q. J. Roy. Meteorol. Soc.* DOI: 10.1002/qj.2879.

Rodwell, M.J., and T. N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Q. J. Roy. Meteor. Soc.*, **33**, 129-146.

Rosso, O. A., Larrondo, H. A., Martín, M. T., Plastino, A., and M. A. Fuentes, 2007: Distinguishing noise from chaos. *Phys. Rev. Lett.*, **99**, 154102.

Rosso O. A. and Masoller C., 2009: Detecting and quantifying stochastic and coherence resonances via information-theory complexity measurements. *Phys. Rev. E*, **79**, 040106(R).

Rosso O. A. and Masoller C., 2009: Detecting and quantifying temporal correlations in stochastic resonance via information theory measures. *Eur. Phys. J. B*, **69**, 37–43.

Rosso O.A., Carpi L.C., Saco P.M., Gómez Ravetti M., Plastino A., Larrondo H., 2012: Causality and the Entropy-Complexity Plane: Robustness and Missing Ordinal Patterns. *Physica A*, **391**, 42–55.

Rosso O.A., Carpi L.C., Saco P.M., Gómez Ravetti M., Larrondo H., Plastino A., 2012: The Amigó paradigm of forbidden/missing patterns: a detailed analysis *Eur. Phys. J. B* **85**, 419–430.

Ruiz J., M. Pulido and T. Miyoshi, 2013: Estimating parameters with ensemble-based data assimilation. A review. *J. Meteorol. Soc. Japan*. **91**, 79–99.

Ruiz J., M. Pulido and T. Miyoshi, 2013: Estimating parameters with ensemble-based data assimilation. Parameter covariance treatment. *J. Meteorol. Soc. Japan*. **91**, 453-469.

Ruiz J. and M. Pulido, 2015: Parameter estimation using ensemble based data assimilation in the presence of model error. *Mon. Wea. Rev.*, **143**, 1568–1582.

Serinaldi F., L. Zunino, O. A. Rosso, 2014: Complexity-entropy analysis of daily stream flow time series in the continental United States. *Stoch. Environ. Res. Risk Assess.*, **28**, 16851708

Shannon C.E., 1948: A mathematical theory of communication. *Bell Syst. Technol. J.*, **27**, 379–423, 623–656.

Shannon C.E. and Weaver W., 1949: The Mathematical Theory of Communication, University of Illinois Press, Champaign, IL, USA.

Schuster, H.G. and Just, W., 2006. *Deterministic chaos: an introduction*. John Wiley and Sons.

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.* **131**, 3079–3102.

Sippel S., Lange H., Mahecha M.D., Hauhs M., Bodesheim P., Kaminski T., Gans F. and Rosso O. A., 2016: Diagnosing the Dynamics of Observed and Simulated Ecosystem Gross Primary Productivity with Time Causal Information Theory Quantifiers. *PLoS ONE*, **11**, e0164960. doi:10.1371/journal.pone.0164960

Stainforth D.A., T. Aina, C. Christensen, M. Collins, N. Faull, D. J. Frame, J. A. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Piani, D. Sexton, L. A. Smith, R. A. Spicer, A. J. Thorpe and M. R. Allen, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature*, **433**, 403–406. doi:10.1038/nature03301

Tirabassi and Massoller C. 2016: Unravelling the community structure of the climate system by using lags and symbolic time-series analysis. *Scientific Reports* , **6** 29804. doi:10.1038/srep29804

Wilks D. S., 2005: Effects of stochastic parameterizations in the Lorenz 96 system. *Q. J. R. Meteorol. Soc.*, **131**, 389–407. doi: 10.1256/qj.04.03

Zanin M., Zunino L., Rosso O. A. and Papo D., 2012: Permutation entropy and its main biomedical and econophysics applications: A review. *Entropy*, **14**, 1553–1577.

LIST OF TABLES

Table 1. Values of the parameters (a_i, i degree of the polynomial term and standard deviation, σ) for the quadratic stochastic parameterization in the perfect-model experiment. The true values correspond to the values used to generate the observations. The optimal values obtained with the hybrid optimization algorithm for $\phi^t = 0$ and $\phi^t = 0.984$ experiments. 37

Table 2. Values of the parameters (a_i, i degree of the polynomial term and σ). The maximum and minimum values used to constrain the optimization and the optimal values obtained with the hybrid optimization algorithm corresponding to the deterministic (Det), $\phi = 0$ and $\phi = 0.984$ experiments. 38

Table 3. Estimated values of the parameters (a_i, i degree of the polynomial term, and stochastic parameter σ) for the deterministic and stochastic parameterizations with $\phi = 0$ in the imperfect-model experiment. 39

	a_0	a_1	a_2	σ
True Values	17.0	-1.20	0.035	1.0
$\phi_T = 0$	17.0	-1.17	0.031	0.82
$\phi_T = 0.984$	17.0	-1.19	0.034	0.88

TABLE 1. Values of the parameters (a_i, i degree of the polynomial term and standard deviation, σ) for the quadratic stochastic parameterization in the perfect-model experiment. The true values correspond to the values used to generate the observations. The optimal values obtained with the hybrid optimization algorithm for $\phi^t = 0$ and $\phi^t = 0.984$ experiments.

Coef	$F = 7$					$F = 18$				
	Min	Max	Det	$\phi = 0$	$\phi = 0.984$	Min	Max	Det	$\phi = 0$	$\phi = 0.984$
a_0	2.0	8.0	5.79	5.78	6.97	14.0	19.0	17.7	18.5	17.1
a_1	-3.5	0.0	-2.79	-1.76	-2.18	-3.0	0.0	-1.19	-1.28	-1.26
a_2	0.0	0.8	0.50	0.22	0.25	0.0	0.5	0.038	0.039	0.049
σ	0.0	2.0		0.32	0.15	0.0	5.0		4.67	2.13

TABLE 2. Values of the parameters (a_i, i degree of the polynomial term and σ). The maximum and minimum values used to constrain the optimization and the optimal values obtained with the hybrid optimization algorithm corresponding to the deterministic (Det), $\phi = 0$ and $\phi = 0.984$ experiments.

	a_0	a_1	a_2	a_3	a_4	σ	$\sqrt{D_{JS}}$
Linear	18.36	-0.981					0.3950E-01
Quadratic	17.7	-1.19	0.038				0.3224E-01
Cubic	18.6	-1.50	0.062	0.0002			0.3434E-01
Quartic	18.2	-1.35	0.094	-0.0046	0.00007		0.3309E-01
Linear	19.1	-1.00				3.83	0.3120E-01
Quadratic	18.5	-1.28	0.039			4.67	0.2910E-01
Cubic	17.1	-1.15	0.073	-0.0033		1.49	0.3050E-01

TABLE 3. Estimated values of the parameters (a_i, i degree of the polynomial term, and stochastic parameter σ) for the deterministic and stochastic parameterizations with $\phi = 0$ in the imperfect-model experiment.

LIST OF FIGURES

Fig. 1. (a) Permutation entropy ($\mathscr{H}$), (b) permutation statistical complexity ($\mathscr{C}$), and (c) the causal entropy-complexity plane ($\mathscr{H} \times \mathscr{C}$) for two-scale Lorenz '96 integrations as a function of the forcing F with a resolution of $\delta F = 0.01$. Vertical dotted lines in (a) and (b) divide the four dynamical regimes found. The minimal and maximal complexity values, $\mathscr{C}_{min}$ and $\mathscr{C}_{max}$ as a function of permutation entropy are shown with black solid curves in panel (c). The transition points between the regions are not represented in panel (c) to improve visibility of the different regimes. 41

Fig. 2. Time series of a two-scale Lorenz '96 variable for (a) $F = 4$ (b) $F = 7$ (c) $F = 18$. 42

Fig. 3. (a) Permutation entropy ($\mathscr{H}$), (b) permutation statistical complexity ($\mathscr{C}$), and (c) the causal entropy-complexity plane ($\mathscr{H} \times \mathscr{C}$) for two-scale Lorenz '96 integrations with varying coupling constant h in steps of $\delta h = 0.1$ for $F = 4$ (continuous line, circle points), $F = 6$ (dotted curves, square points) and $F = 18$ (dashed curves, triangle points). 43

Fig. 4. Jensen-Shannon divergence as a function of a_0, a_1 and a_2 under the perfect model assumption. The true parameter values are shown with vertical dotted lines. The square root of Jensen-Shannon divergence is used to make visible small values close to the minimum. . . . 44

Fig. 5. Jensen-Shannon divergence as a function of the standard deviation, σ, and autoregressive parameter values $\phi^t = 0$ (continuous line), and $\phi^t = 0.984$ (dotted line) for the perfect model experiments. The other parameters are kept fixed at the optimal values. 45

Fig. 6. Jensen-Shannon divergence between the probability distribution from the model integration with a given set of parameters, and the one from the natural two-scale Lorenz '96 system evolution for an external forcing of $F = 7$. Two parameters are kept fixed at the estimated values (Table 2) and the third one is varied: a) a_0, b) a_1, and c) a_2 dependencies. 46

Fig. 7. Jensen-Shannon divergence between the probability distribution from the model integration with a given set of parameters, and the one from the natural two-scale Lorenz '96 system evolution for an external forcing of $F = 18$. One parameter value is varied and the other two are kept fixed at the optimal values (Table 2), (a) a_0 parameter is varied, (b) a_1 and (c) a_2. . . . 47

Fig. 8. Jensen-Shannon divergence as a function of σ the standard deviation of the AR1 process for the cases $\phi = 0$ and $\phi = 0.984$. (a) $F = 7$, (b) $F = 18$. 48

Fig. 9. Scatterplots of the forcing as a function of the state variable given by the two-scale Lorenz '96 model (gray dots) and the one given by the deterministic (a) and the stochastic parameterizations, $\phi = 0$ (b) and $\phi = 0.984$ (c), with optimal parameters (black dots) for the $F = 18$ case. 49

Fig. 10. Histogram PDFs for the nature integration with $F = 18$ (continuous line), for the imperfect model with the deterministic parameterization using optimal parameter values (dashed line) and for the imperfect model with the stochastic parameterization using $\phi = 0$ (dotted line) and $\phi = 0.984$ (gray line). 50

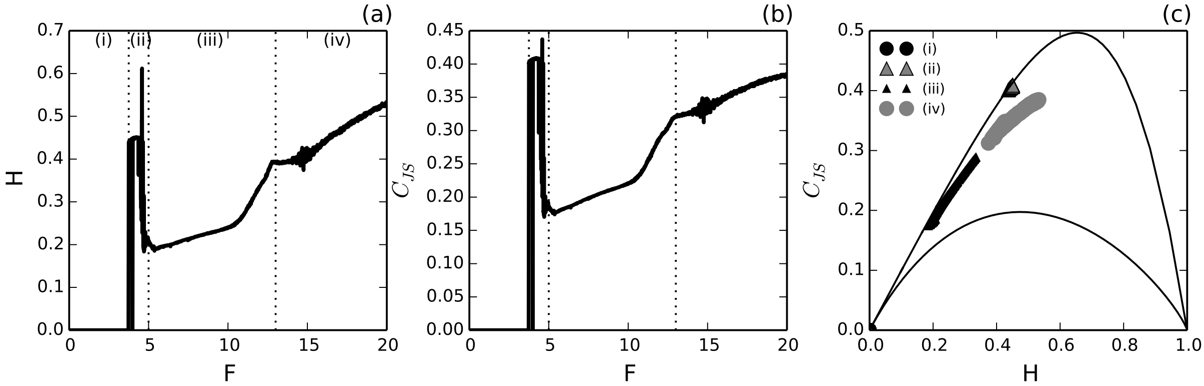

FIG. 1. (a) Permutation entropy ($\mathscr{H}$), (b) permutation statistical complexity ($\mathscr{C}$), and (c) the causal entropy-complexity plane ($\mathscr{H} \times \mathscr{C}$) for two-scale Lorenz '96 integrations as a function of the forcing F with a resolution of $\delta F = 0.01$. Vertical dotted lines in (a) and (b) divide the four dynamical regimes found. The minimal and maximal complexity values, $\mathscr{C}_{min}$ and $\mathscr{C}_{max}$ as a function of permutation entropy are shown with black solid curves in panel (c). The transition points between the regions are not represented in panel (c) to improve visibility of the different regimes.

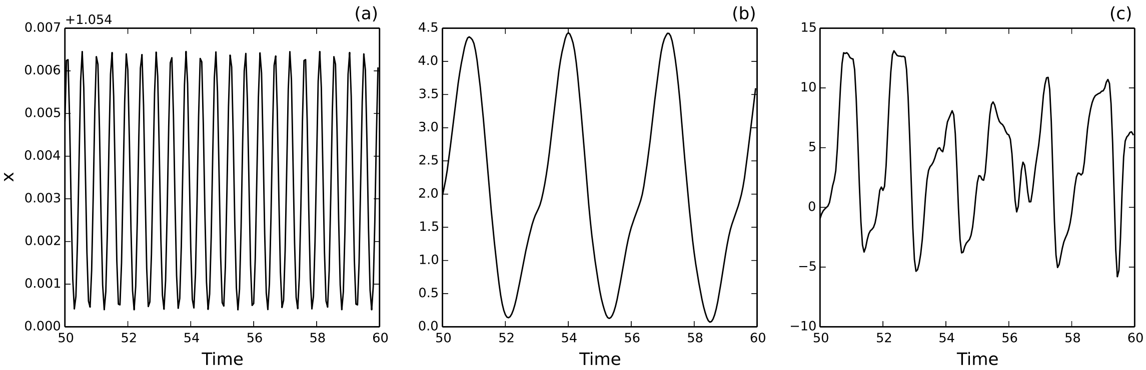

FIG. 2. Time series of a two-scale Lorenz '96 variable for (a) $F = 4$ (b) $F = 7$ (c) $F = 18$.

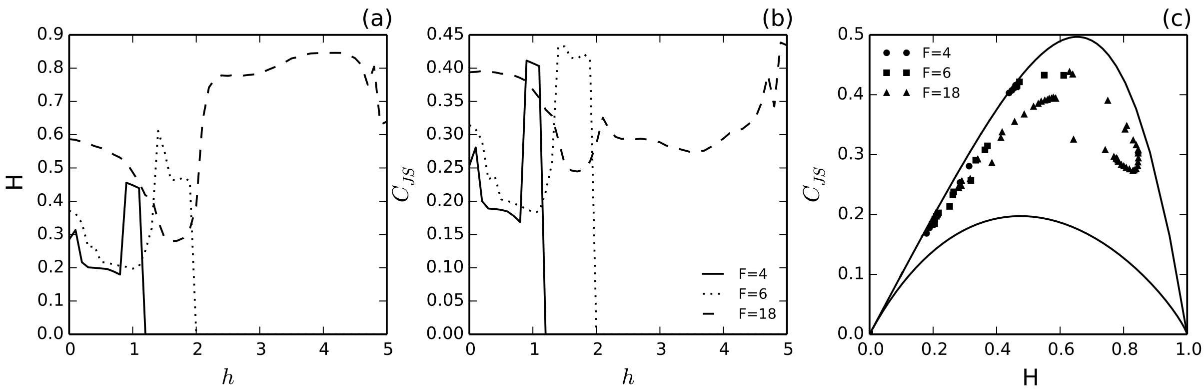

FIG. 3. (a) Permutation entropy ($\mathscr{H}$), (b) permutation statistical complexity ($\mathscr{C}$), and (c) the causal entropy-complexity plane ($\mathscr{H} \times \mathscr{C}$) for two-scale Lorenz '96 integrations with varying coupling constant h in steps of $\delta h = 0.1$ for $F = 4$ (continuous line, circle points), $F = 6$ (dotted curves, square points) and $F = 18$ (dashed curves, triangle points).

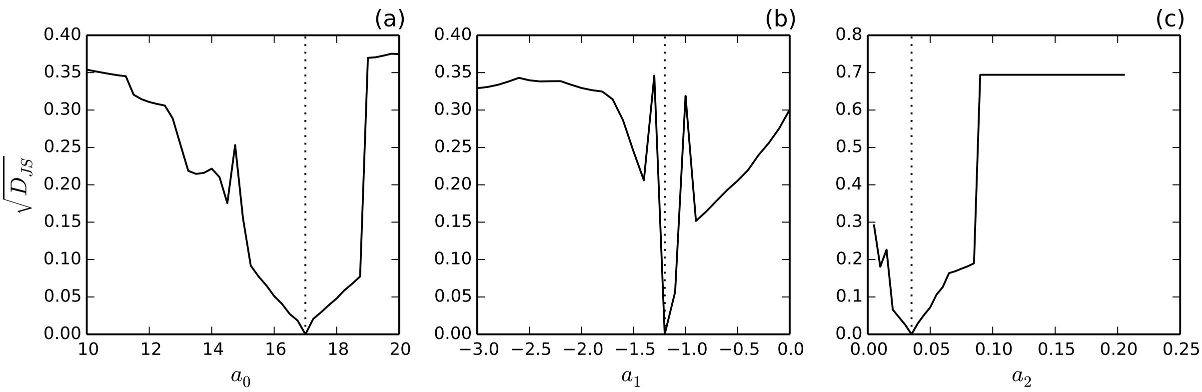

FIG. 4. Jensen-Shannon divergence as a function of a_0, a_1 and a_2 under the perfect model assumption. The true parameter values are shown with vertical dotted lines. The square root of Jensen-Shannon divergence is used to make visible small values close to the minimum.

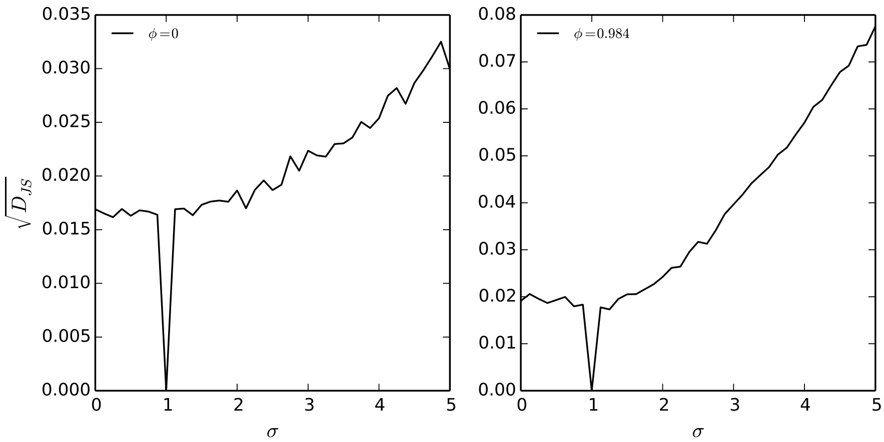

FIG. 5. Jensen-Shannon divergence as a function of the standard deviation, σ, and autoregressive parameter values $\phi^t = 0$ (continuous line), and $\phi^t = 0.984$ (dotted line) for the perfect model experiments. The other parameters are kept fixed at the optimal values.

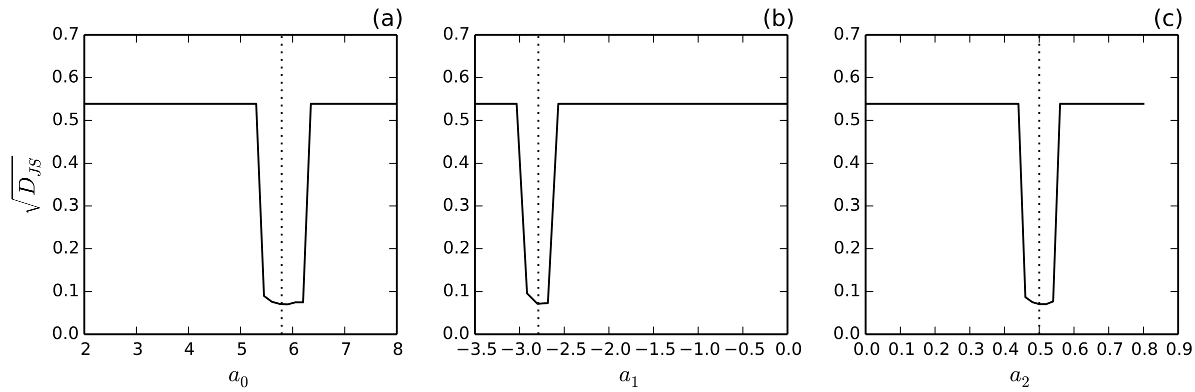

FIG. 6. Jensen-Shannon divergence between the probability distribution from the model integration with a given set of parameters, and the one from the natural two-scale Lorenz '96 system evolution for an external forcing of $F = 7$. Two parameters are kept fixed at the estimated values (Table 2) and the third one is varied: a) a_0, b) a_1, and c) a_2 dependencies.

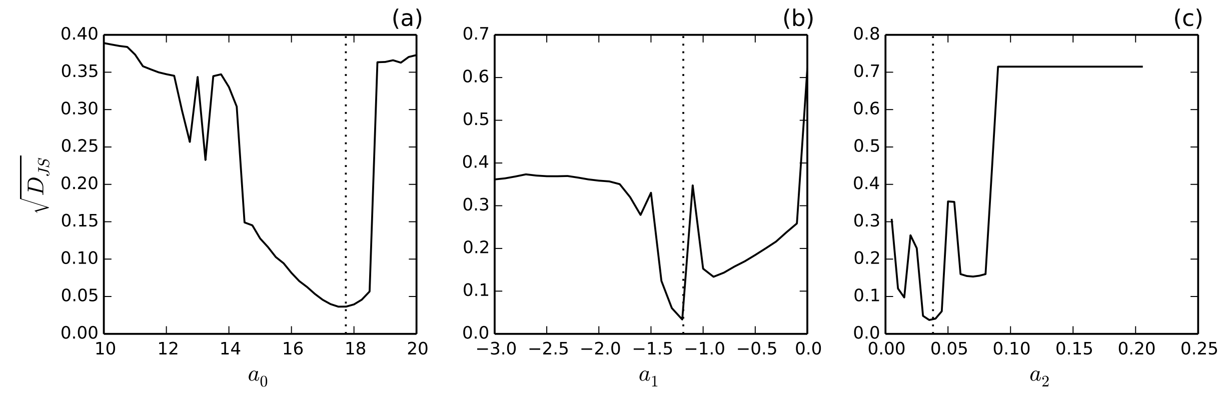

FIG. 7. Jensen-Shannon divergence between the probability distribution from the model integration with a given set of parameters, and the one from the natural two-scale Lorenz '96 system evolution for an external forcing of $F = 18$. One parameter value is varied and the other two are kept fixed at the optimal values (Table 2), (a) a_0 parameter is varied, (b) a_1 and (c) a_2.

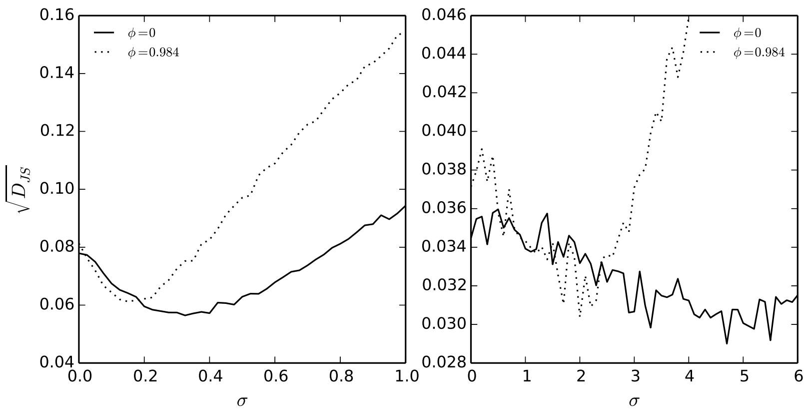

FIG. 8. Jensen-Shannon divergence as a function of σ the standard deviation of the AR1 process for the cases $\phi = 0$ and $\phi = 0.984$. (a) $F = 7$, (b) $F = 18$.

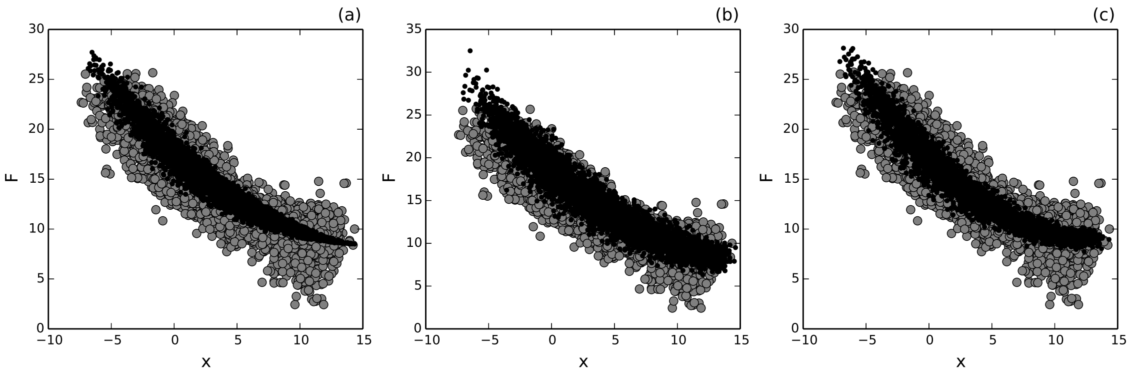

FIG. 9. Scatterplots of the forcing as a function of the state variable given by the two-scale Lorenz '96 model (gray dots) and the one given by the deterministic (a) and the stochastic parameterizations, $\phi = 0$ (b) and $\phi = 0.984$ (c), with optimal parameters (black dots) for the $F = 18$ case.

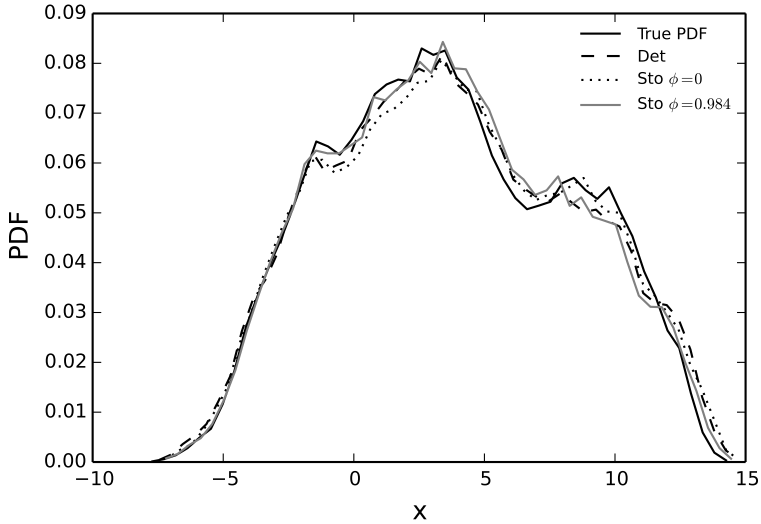

FIG. 10. Histogram PDFs for the nature integration with $F = 18$ (continuous line), for the imperfect model with the deterministic parameterization using optimal parameter values (dashed line) and for the imperfect model with the stochastic parameterization using $\phi = 0$ (dotted line) and $\phi = 0.984$ (gray line).