*Q. J. R. Meteorol. Soc.* (2003), **129**, pp. 239–275 doi: 10.1256/qj.01.174

# Inferring instantaneous, multivariate and nonlinear sensitivities for the analysis of feedback processes in a dynamical system: Lorenz model case-study

By FILIPE AIRES[1,2*] and WILLIAM B. ROSSOW[3]

[1]*Columbia University, NASA Goddard Institute for Space Studies, USA*
[2]*Laboratoire de Météorologie Dynamique du CNRS, France*
[3]*NASA Goddard Institute for Space Studies, USA*

(Received 17 October 2001; revised 17 May 2002)

SUMMARY

As an alternative to classical linear feedback analysis, we present a nonlinear approach for the determination of the sensitivities of a dynamical system from observations of its variations. The new methodology consists of statistical estimates of all the pair-wise relationships among the system state variables based on a neural-network modelling of the system dynamics (its time evolution). The model can then be used to estimate the instantaneous, multivariate, nonlinear sensitivities. Classical feedback analysis is re-examined in terms of these sensitivities, which are shown to be more fundamental in the analysis of feedback processes than estimates of feedback factors and to provide a more appropriate representation of the system's behaviour. The method is described and tested on synthetic observations of the time variations of the Lorenz low-order atmospheric model where the correct sensitivities can be evaluated analytically.

KEYWORDS: Climate sensitivities Feedback analysis Neural network

## 1. INTRODUCTION

Feedback processes are present in dynamical systems that, like the climate, involve nonlinear relationships among many variables integrated over time. A feedback process (or loop) involves at least two coupled variables where an initial perturbation of one causes a change of the other that causes further perturbation of the first. A formalism from electrical circuit theory (Bode 1945) has long been used to study feedback processes in climate (cf. Peixoto and Oort 1992; Curry and Webster 1999), especially in models used to predict climate change (e.g. Hansen *et al.* 1984; Schlesinger 1985). This approach combines the sensitivities of the system (first derivatives of one variable by another) to characterize changes in the equilibrium state of the system produced by an external forcing. Such an approach is valid in a theoretical model where the instantaneous sensitivities can be evaluated directly from the continuous equations of the system; however, its application is more questionable in situations where the underlying equations are unknown (or only partially known) and all we have are observations of the system behaviour in the form of discrete measurements of the system state at different times. This is the situation faced in the study of the real climate, but also encountered in the study of climate models which are numerically discrete approximations of the real system.

The sensitivities of climate models have been estimated in several ways. One approach is to introduce a perturbation of one variable at a time and then evaluate the changes of all the other variables. The problems associated with this approach are numerous. First, since the initial perturbation is limited to one variable only, the interdependence of the sensitivities is completely neglected. Even if many variables are perturbed together, it is difficult to be sure that the estimate of the multivariate sensitivities is complete. Second, in most cases the sensitivities are estimated from the finite differences between two (usually equilibrium) states of the system, either as differences

* Corresponding author: Department of Applied Physics and Applied Mathematics, Columbia University, NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY 10025, USA. e-mail: faires@giss.nasa.gov

with geographical location or time, depending on the strategy adopted (Slingo *et al.* 2000). Usually, this procedure also involves averaging the variables over large space and time domains, which suppresses all the possible non-local relationships. Third, in taking such differences over finite (large) space or time intervals, the individual relationships of pairs of variables are already 'contaminated' by actions of the feedback processes. For example, one feedback process can reduce the importance of another feedback process, leading to an underestimate of the latter by comparing differences over a finite time interval. This point is equivalent to saying that, as the differencing interval increases, the higher-order relationships (higher-order derivatives at least) become important. In this sense, the usual method for estimating the system sensitivities ignores these higher-order terms, which is equivalent to linearizing the system behaviour. As we will see, this simplistic approach can be highly misleading because the space-averaged and/or time-averaged sensitivities may not represent the system dynamics correctly. However, this classical feedback analysis can still be useful if one is interested in the equilibrium (or transient) response of one variable to a pertubation of one other, especially when one is comparing two nonlinear integrations.

The analyses described above can only be performed on models, usually not on observations of a real system. In the study of the climate, we cannot conduct controlled perturbation experiments or observe changes of the equilibrium state (although paleo-climate comparisons are assumed to represent changes of equilibrium); we have only the transient variations of the system, some induced by temporary perturbations (e.g. volcanic eruptions), some unforced variations, to work with. Hence the classical feedback analysis approach is not very helpful in understanding climate feedback processes from observations nor in verifying how well our climate models represent them because the hypotheses underlying the classical feedback analysis are too crude (as we will discuss in more detail): a mono-variable conception of forcing and response, a linear model, constant and mutually independent sensitivities. Thus, it seems questionable that a few (constant) feedback factors can be used to explain the time integral of the nonlinear, multivariate climate processes well enough to predict accurately the climate response to an external forcing as a change in the equilibrium state. Even if this were possible, such an approach would not describe the transient adjustment period between the beginning of the changed forcing and the attainment of new equilibrium, including its duration.

To avoid these problems we propose an alternative approach that takes the sensitivities themselves as the fundamental quantities defining the dynamical behaviour of a nonlinear system, rather than their combination into constant feedback factors. Our goal is not to develop a statistical method for climate prediction; the use of the sensitivities for this purpose is questionable because of the chaotic behaviour of the system. Instead, we believe that accurate estimates of the instantaneous, multivariate and state-dependent sensitivities can provide a more appropriate and better understanding of nonlinear climate feedback processes and a much better way to compare climate models with each other and with observations. If a model does not possess the same time/space-localized, multivariate and nonlinear sensitivities as those inferred from observations, the feedback processes of the model must be wrong. Stability analyses of the dynamical processes involved can also be performed using these sensitivities and, for prediction purposes, the temporal propagation of errors onto the state of the system can also be analysed (Smith 1997).

To characterize the full nonlinearity of the feedback processes, we will show that it is necessary to estimate the system sensitivities at space- and time-scales sufficiently small that they can be treated as constant and that the higher-order relationships can be neglected. This approach provides a more natural way to understand the physical

processes, their causal relationships, how they evolve during a transient climate change, and how they integrate over time to determine the change in the equilibrium climate state. Moreover, this approach suggests some ways to evaluate the completeness of the description of the climate system, either observations of it or a model of it, in terms of the list of variables, their resolutions and coverage needed to describe the system dynamics accurately. Understanding of the feedback mechanisms is a prerequisite to predicting them with useful skill, so carrying out an analysis along these lines can lead to improved numerical models that are a much more credible way to perform climate predictions.

In section 2, we first refine the terminology required to perform a feedback analysis with an emphasis on the discrete formulation of dynamical systems, which is better adapted to prediction, to the description of the cause-and-effect relations underlying the feedback processes, and to the direct analysis of observations. Then we develop a more general analysis of feedbacks, showing that the sensitivities are sufficient to define the interdependencies of one variable on another that cause the feedback loops in a nonlinear dynamical system. Finally, we compare this analysis with the classical linear feedback analysis to illustrate the limitations of feedback factors as descriptors of nonlinear dynamics. In section 3 we describe a general, multivariate, nonlinear statistical method to estimate the sensitivities of a dynamical system. As a test of the validity of this analysis approach, we need a dynamical system with known analytic equations that are simple enough to illustrate the workings of our analysis method and that allow for a direct evaluation of the sensitivities for comparison with the results of our analysis. We also need a system that has more than one variable and is sufficiently nonlinear in character to be a real challenge to the analysis, not just a trivial exercise. In section 4 we present a discrete version of the Lorenz low-order dynamical model, which meets all our requirements, and, though very simple, is also thought to possess some characteristics similar to the real climate. We also show the analytic expressions for this model that appear in the general feedback analysis and apply the classical feedback analysis to show the limitations of feedback factors for describing the system dynamics. Then in section 5, we apply our analysis method to the time-evolving output (observations) of the Lorenz model and evaluate the accuracy of the sensitivities determined by our statistical analysis method compared with the exact analytic expressions. This comparison shows how the instantaneous, multivariate, nonlinear sensitivities can accurately represent the system dynamics. Section 6 has some concluding remarks.

## 2. Feedbacks in a dynamical system

There are two general ways of formulating a dynamical system: the continuous and the discrete approaches. We prefer the discrete formulation because it is simpler to describe the cause-and-effect relationships between variables. Furthermore, the discrete approach is more practical for prediction when no theoretical physical evolution model is available. We adopt the discrete formalism in the following, but will refer sometimes to the continuous case. The goal of this section is to show how time integration of dynamical relationships leads to feedback processes and to highlight the role played by the sensitivities.

### (*a*) *Dynamical systems*

The object of this study is the analysis of a physical dynamical system by observing the time variations of the quantities defining the state of the system. A dynamical system is often described by a set of ordinary differential equations (ODEs) which come from

the physics of the problem. For practical considerations or because these ODEs are not known, the dynamical system is often discretized in the form:

$$\mathbf{X}(t + \Delta t) = \mathcal{A}(\mathbf{P}(t)) + \boldsymbol{\varepsilon}(t), \tag{1}$$

where $\mathbf{X}(t)$ is the $p$-dimensional vector of observable variables (defining the state of the system) at time $t$, $\mathbf{P}(t)$ is the $d$-dimensional vector of variables defining the system behaviour (predictors), which can include $\mathbf{X}(t)$, $\boldsymbol{\varepsilon}(t)$ is noise (observational or model errors), and $\mathcal{A}$ is a mapping, possibly nonlinear. This kind of model is often used in atmospheric and oceanic sciences to perform, for example, climatological predictions.

Determination of the good predictors, $\mathbf{P}(t)$, is the crucial issue for the quality of the model. This determination uses all the a priori physical knowledge about the system. If $\mathbf{P}(t) = \mathbf{X}(t)$, the system is said to be autoregressive. Sometimes, the prediction of $\mathbf{X}(t + \Delta t)$ requires the knowledge of

$$\mathbf{X}(t), \mathbf{X}(t - \Delta t), \ldots, \mathbf{X}(t - q\Delta t)$$

because the system dynamics has inertia that requires the knowledge of previous steps. Then, the system is said to be autoregressive with memory $q$, denoted as an AR($q$) model. However, defining a new state variable

$$\mathbf{X}'(t) = (\mathbf{X}(t), \mathbf{X}(t - \Delta t), \ldots, \mathbf{X}(t - q\Delta t)),$$

one can rewrite this AR($q$) system as an AR(1) model with memory 1. Since we are interested in using this model to retrieve sensitivities from two consecutive time steps, we will use, in the following, only a memoryless model (i.e. $q = 1$), which is consistent with the physics in general-circulation model numerical simulations.

If the dynamical system Eq. (1) is linearized near $\mathbf{P}(t_0)$, we obtain

$$\mathbf{X}(t_0 + \Delta t) - \mathbf{X}(t_0) = \mathbf{G}(\mathbf{P}(t_0))\Delta\mathbf{P}(t_0) + \boldsymbol{\varepsilon}(t_0), \tag{2}$$

where

$$\mathbf{G}(\mathbf{P}(t_0)) = \frac{\partial \mathbf{X}(t_0 + \Delta t)}{\partial \mathbf{P}(t_0)} \tag{3}$$

is the Jacobian or sensitivity matrix of the mapping $\mathcal{A}$ at state $\mathbf{P}(t_0)$.

The uncertainty $\boldsymbol{\varepsilon}(t)$ is often neglected, so the discretized system of Eq. (2) is entirely defined by the sensitivities $\mathbf{G}(\mathbf{P}(t))$ of the dynamical system and by an initial state $\mathbf{P}(t_0)$.

### (*b*) *General feedback analysis: time integration of sensitivities*

For simplicity of notation, we suppose, as is true in most cases, that the system is autoregressive, i.e. the predictors are equal to the state variables (the response), $\mathbf{P}(t) = \mathbf{X}(t)$.

In the simplest system, where the local mapping $\mathcal{A}$ of Eq. (1) is linear, we have $\mathbf{X}(t + \Delta t) = \mathbf{A}\mathbf{X}(t)$, where the matrix $\mathbf{A}$ is diagonal and independent of time. The variables of the system are independent and evolve independently as

$$X_i(t_0 + k\Delta t) = (A_{ii})^k X_i(t_0).$$

The absolute value $|A_{ii}|$ implies decreasing $X_i$ if $|A_{ii}| < 1$ or increasing $X_i$ if $|A_{ii}| > 1$.

However, if the matrix $\mathbf{A}$ is non-diagonal, i.e. some of the variables of the system are dependent on other variables, an initial perturbation of one variable will propagate into all the other variables that are directly or indirectly dependent on this initially perturbed

variable. After $k$ time steps, the state of the system is given by

$$\mathbf{X}(t_0 + k\Delta t) = \mathbf{A}^k \mathbf{X}(t_0).$$

So, the responses $\mathbf{X}(t)$, at any time $t$, are still a linear combination of the predictors at time $t_0$, but the impact of an initial perturbation has been mixed up into each linked variable because of the feedback loops. For the system to be stable, it is required that the eigenvalues of the matrix $\mathbf{A}^k$ be less than one, otherwise the system is unstable.

If the mapping $\mathcal{A}$ of Eq. (1) is nonlinear, the Jacobians are dependent on the state $\mathbf{X}(t)$. So, even if we linearize the mapping $\mathcal{A}$ using its Jacobians, $\mathbf{G}(\mathbf{X}(t))$, after $k$ time steps, the state of the system is given by

$$\mathbf{X}(t_0 + k\Delta t) = \left\{ \prod_{l=1}^{k} \mathbf{G}(\mathbf{X}(t_0 + l\Delta t)) \right\} \mathbf{X}(t_0),$$

which is more complex ($\prod$ is the product symbol). This Jacobian-product-based propagator is necessary also for linear but non-autonomous (unforced) mappings.

To define a feedback process in the discrete formulation of a dynamical system, we need at least two time steps to describe the feedback loops involved. If an initial perturbation $\Delta\mathbf{X}(t_0)$ is introduced into the system at time $t_0$, the response of the system at time $t_0 + \Delta t$ is approximated to first order by

$$\Delta\mathbf{X}(t_0 + \Delta t) \simeq \mathbf{G}(t_0, t_0 + \Delta t)\Delta\mathbf{X}(t_0), \tag{4}$$

where $\mathbf{G}(t_0, t_0 + \Delta t)$, the gain of the system from $t_0$ to $t_0 + \Delta t$, is the Jacobian matrix $\mathbf{G}(\mathbf{X}(t))$ of the mapping $\mathcal{A}$ between $[t_0, t_0 + \Delta t]$: matrix $\mathbf{G}(t_0, t_0 + \Delta t)$ has elements

$$\frac{\partial \mathcal{A}_i(t_0 + \Delta t)}{\partial X_j(t_0)} = \frac{\partial X_i(t_0 + \Delta t)}{\partial X_j(t_0)}$$

at coordinates $(i, j)$. An initial perturbation $\Delta X_j(t_0)$, on variable $X_j$ at time $t_0$, is then propagated at time $t_0 + \Delta t$ to each variable $X_i$ that is linked to $X_j$ via off-diagonal terms in $\mathbf{G}(t_0, t_0 + \Delta t)$. But the resulting perturbations $\Delta X_i(t_0 + \Delta t)$ are just the direct impact of the initial perturbation, so there is no feedback during $[t_0, t_0 + \Delta t]$.

At time $t = t_0 + 2\Delta t$, the impact on the system is given to first order by

$$\Delta\mathbf{X}(t_0 + 2\Delta t) \simeq \mathbf{G}(t_0 + \Delta t, t_0 + 2\Delta t)\Delta\mathbf{X}(t_0 + \Delta t), \tag{5}$$

$$\simeq \mathbf{G}(t_0 + \Delta t, t_0 + 2\Delta t)\mathbf{G}(t_0, t_0 + \Delta t)\Delta\mathbf{X}(t_0), \tag{6}$$

$$\simeq \mathbf{G}(t_0, t_0 + 2\Delta t)\Delta\mathbf{X}(t_0). \tag{7}$$

The previously propagated perturbations $\Delta X_i(t_0 + \Delta t)$ resulting from $\Delta X_j(t_0)$ can then perturb $\Delta X_j(t_0 + 2\Delta t)$, completing a feedback loop. The initial perturbation $\Delta X_j(t_0)$, can be amplified or damped to $\Delta X_j(t_0 + 2\Delta t)$. We see in this simple example that feedback processes result from the time integration of the variable dependencies of the system. The term $\mathbf{G}(t_0, t_0 + 2\Delta t)$, representing the evolution of the system in two time steps, includes these feedback loops.

### (c) *Forcing/response*

We introduce in this section the concept of external forcing to formalize the perturbations of the variables of the system we have discussed in the previous example. It is important to note that the feedback processes are present and active in a dynamical

system, even when no external forcing is applied and the system is in equilibrium (forcing and feedback are often confused).

An external forcing perturbs some internal variables of the system (i.e. variables that define the state of the system). The external forcing has an impact on the internal variables, but the reverse is not true: the forcing is independent of the internal variables. There are many ways an external forcing could operate. The simplest case is the introduction of an impulse perturbation at time $t_0$: $\mathbf{E}(t) = \mathbf{E}_0\delta(t_0 - t)$, a time-localized volcanic eruption for example. In this case, the initial perturbation will be propagated in time through the internal variables, following their interdependencies. This is the example discussed in the previous section.

The external forcing can also begin at time $t_0$ and remain constant in time:

$$\mathbf{E}(t) = \mathbf{E}_0,\ t \in [t_0,\ t_0 + \Delta t,\ \ldots\ ].$$

In this case, the relations Eqs. (5)–(6) become more complex:

$$\begin{aligned}\Delta\mathbf{X}(t_0 + 2\Delta t) &\simeq \mathbf{E}(t_0 + 2\Delta t) + \frac{\partial\mathbf{X}(t_0 + 2\Delta t)}{\partial\mathbf{X}(t_0 + \Delta t)}\mathbf{E}(t_0 + \Delta t) \\ &\quad + \frac{\partial\mathbf{X}(t_0 + 2\Delta t)}{\partial\mathbf{X}(t_0 + \Delta t)}\frac{\partial\mathbf{X}(t_0 + \Delta t)}{\partial\mathbf{X}(t_0)}\mathbf{E}(t_0), \qquad (8)\end{aligned}$$

$$\simeq \mathbf{E}_0 + \mathbf{G}(t_0 + \Delta t,\ t_0 + 2\Delta t)\mathbf{E}_0 + \mathbf{G}(t_0,\ t_0 + 2\Delta t)\mathbf{E}_0. \qquad (9)$$

If the gains of the system are constant, $\mathbf{G}$ (i.e. a linear dynamical system), then at time $t + k\Delta t$

$$\Delta\mathbf{X}(t_0 + k\Delta t) = (\mathbf{I} + \mathbf{G} + \mathbf{G}^2 + \cdots + \mathbf{G}^k)\mathbf{E}_0 = \frac{\mathbf{I} - \mathbf{G}^{k+1}}{\mathbf{I} - \mathbf{G}}\mathbf{E}_0, \qquad (10)$$

where $\mathbf{I}$ is the identity matrix. If the eigenvalues of matrix $\mathbf{G}$ have an absolute value lower than 1 (otherwise the system is unstable), the effect of the external forcing is stabilized, the dynamical system eventually reaches a stabilized state:

$$\Delta\mathbf{X}(t_0 + k\Delta t) \simeq \frac{\mathbf{I}}{\mathbf{I} - \mathbf{G}}\mathbf{E}_0, \quad \text{for } k \to +\infty \qquad (11)$$

$$\simeq \mathbf{E}_0 + \frac{\mathbf{G}}{\mathbf{I} - \mathbf{G}}\mathbf{E}_0. \qquad (12)$$

For example, a mono-variable system with $G = 1/2$, $E_0 = 1$ and $X(t_0) = 0$ stabilizes at

$$\lim_{k \to +\infty} X(t_0 + k\Delta t) = 2;$$

the forcing has changed the equilibrium state of the system. Figure 1 shows the values of the stabilized solutions of this simple system for different values of $G$. If $-1 < G < 1$ then the system stabilizes. If $G$ is close to 0, the system stabilizes near $E_0$. The closer the gain of the system $G$ to $1^-$ (i.e. lower but close to 1), the higher is the value at which it stabilizes. If the absolute value of $G$ is bigger than 1, the system is unstable.

### (*d*) *The classical analysis of a parallel feedback configuration*

The previous examples are very general since each variable of the system can be dependent on any other variables. But in some cases knowledge of cause-and-effect relationships provides a priori information about the ordering and the structure of the

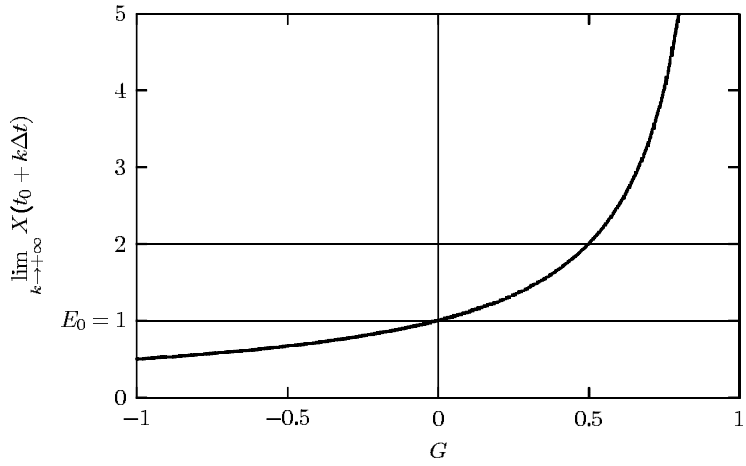



Figure 1. The stabilized values $\lim_{k\to+\infty} X(t_0 + k\Delta t)$ of a mono-variable linear system for different values of the gain of the system, $G$, with external forcing $E_0 = 1$. See text for explanation.

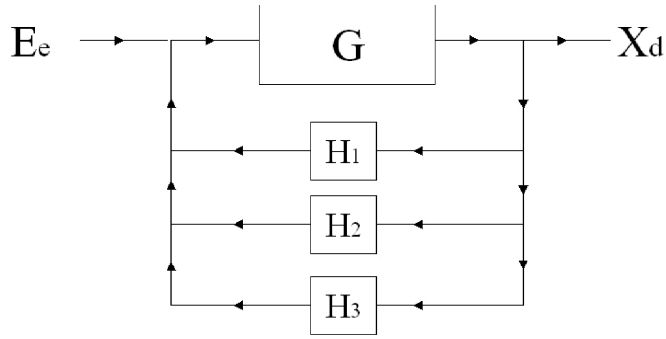



Figure 2. Feedback-loops system in parallel: $E_e$ is the external forcing in variable $X_e$, $G$ is the linear gain of the system, $X_d$ is the 'diagnose' variable where the impact of the forcing is observed, and the coefficients $H_i$ represent the feedbacks.

dependencies. It is then possible, and recommended, to use this information. This kind of a priori information is used, for example, in the cause-and-effect analysis technique (Andronova and Schlesinger 1991; Andronova and Schlesinger 1992).

In the classical feedback analysis (e.g. Hansen 1984; Schlesinger 1985), it is supposed that the external forcings $E_{X_e}$ of the system act on only one variable $X_e$ of the system, that all the other internal variables $\{X_i = X_i(X_d)\}$ are all dependent only on one particular internal variable $X_d$ (i.e. the diagnosed variable), that the impact of the external forcings is observed on this particular internal variable $X_d$ (i.e. $X_d$ is a function of $X_e$), and that the feedbacks act in parallel (Fig. 2). These assumptions are very strong cause-and-effect constraints: the diagnosed variable $X_d$ is supposed to be more important than the other internal variables $\{X_i\}$; by hypothesis the $\{X_i\}$ are dependent only on $X_d$, and they are not directly dependent on the external forcing or each other. In this case, the feedback processes are assumed to act in parallel, i.e. they do not interact with each other.

Since the external forcing $E_{X_e}$ acts on only one variable, $X_e$, of the system, the general multivariate expression in Eq. (8) becomes a scalar relation:

$$\Delta X_e(t_0+2\Delta t) \simeq E_{X_e}(t_0+2\Delta t) + \sum_i \frac{\partial X_e(t_0+2\Delta t)}{\partial X_i(t_0+\Delta t)} \Delta X_i(t_0+\Delta t) + \sum_i \sum_j \frac{\partial X_e(t_0+2\Delta t)}{\partial X_i(t_0+\Delta t)} \frac{\partial X_i(t_0+\Delta t)}{\partial X_j(t_0)} \Delta X_j(t_0). \tag{13}$$

We measure the effect of the constant external forcing $E_{X_e}$ on $X_d$, the diagnosed variable (Fig. 2). We then analyse the system $\Delta X_e \longrightarrow \Delta X_d$. So the perturbations $\Delta X_i(t_0+\Delta t)$ and $\Delta X_j(t_0)$ are considered only for the diagnosed variable $X_d$. In other words, the impact of the perturbations on variables other than $X_d$ are not taken into account in this classical analysis (see Fig. 2). Thus, Eq. (13) becomes

$$\Delta X_e(t_0+2\Delta t) \simeq E_{X_e}(t_0+2\Delta t) + \frac{\partial X_e(t_0+2\Delta t)}{\partial X_d(t_0+\Delta t)} \Delta X_d(t_0+\Delta t) + \sum_i \frac{\partial X_e(t_0+2\Delta t)}{\partial X_i(t_0+\Delta t)} \frac{\partial X_i(t_0+\Delta t)}{\partial X_d(t_0)} \Delta X_d(t_0). \tag{14}$$

Because of the hierarchical cause-and-effect dependencies adopted, i.e.

$$X_e(t_0) \to X_d(t_0+\Delta t) \to X_i(t_0+2\Delta t)$$

(see Fig. 2), some of the partial derivatives in Eq. (14) are zero:

$$\frac{\partial X_e(t_0+2\Delta t)}{\partial X_d(t_0+\Delta t)} = \frac{\partial X_e(t_0+2\Delta t)}{\partial X_e(t_0+\Delta t)} = 0, \tag{15}$$

and Eq. (14) simplifies to

$$\Delta X_e(t_0+2\Delta t) \simeq \underbrace{E_{X_e}(t_0+2\Delta t)}_{\text{external forcing}} + \underbrace{\sum_{i\neq d, i\neq e} \frac{\partial X_e(t_0+2\Delta t)}{\partial X_i(t_0+\Delta t)} \frac{\partial X_i(t_0+\Delta t)}{\partial X_d(t_0)} \Delta X_d(t_0)}_{\text{feedback terms}} \tag{16}$$

$$\simeq E_{X_e}(t_0+2\Delta t) + \sum_{i\neq d, i\neq e} H_i(t_0, t_0+2\Delta t) \Delta X_d(t_0), \tag{17}$$

where the terms $H_i(t_0, t_0+2\Delta t)$ are the products of first derivatives describing the cause-and-effect relations. Expression (17) can be multiplied by the gain $G(t_0+2\Delta t, t_0+3\Delta t)$ of the system $\Delta X_e(t_0+2\Delta t) \longrightarrow \Delta X_d(t_0+3\Delta t)$:

$$\left.\begin{aligned} \Delta X_d(t_0+3\Delta t) &\simeq G(t_0+2\Delta t, t_0+3\Delta t) \Delta X_e(t_0+2\Delta t) \\ &\simeq G(t_0+2\Delta t, t_0+3\Delta t) E_{X_e}(t_0+2\Delta t) \\ &\quad + G(t_0+2\Delta t, t_0+3\Delta t) \sum_{i\neq d, i\neq e} H_i(t_0, t_0+2\Delta t) \Delta X_d(t_0). \end{aligned}\right\} \tag{18}$$

If the system is in equilibrium, or if $\Delta t$, the time discretization, is sufficiently small, i.e. $\Delta X_d(t_0+3\Delta t) \simeq \Delta X_d(t_0)$,

$$\left(1 - G(t_0+2\Delta t, t_0+3\Delta t) \sum_{i\neq d, i\neq e} H_i(t_0, t_0+2\Delta t)\right) \Delta X_d(t_0+3\Delta t) \simeq G(t_0+2\Delta t, t_0+3\Delta t) E_{X_e}(t_0+2\Delta t). \tag{19}$$

So,

$$\Delta X_{\mathrm{d}}(t_0 + 3\Delta t) \simeq \frac{G(t_0 + 2\Delta t,\, t_0 + 3\Delta t)}{1 - G(t_0 + 2\Delta t,\, t_0 + 3\Delta t) \sum_{i \neq d, i \neq e} H_i(t_0,\, t_0 + 2\Delta t)} E_{X_{\mathrm{e}}}(t_0 + 2\Delta t) \quad (20)$$

$$\simeq \frac{G(t_0 + 2\Delta t,\, t_0 + 3\Delta t)}{1 - \sum_{i \neq d, i \neq e} f_i(t_0,\, t_0 + 3\Delta t)} E_{X_{\mathrm{e}}}(t_0 + 2\Delta t), \quad (21)$$

where the terms

$$f_i(t_0,\, t_0 + 3\Delta t) = G(t_0 + 2\Delta t,\, t_0 + 3\Delta t) H_i(t_0,\, t_0 + 2\Delta t)$$

are called the feedback factors. The gain with feedbacks is then defined by

$$G_f(t_0 + 2\Delta t,\, t_0 + 3\Delta t) = \frac{G(t_0 + 2\Delta t,\, t_0 + 3\Delta t)}{1 - \sum_{i \neq d, i \neq e} f_i(t_0,\, t_0 + 3\Delta t)}. \quad (22)$$

The feedback $f_i$ factors are dependent on *both* the variable $X_{\mathrm{e}}$ perturbed by the external forcing and the diagnosed variable $X_{\mathrm{d}}$ chosen in the beginning of the analysis. These feedback factors are time-dependent, but this expression is traditionally (Peixoto and Oort 1992; Curry and Webster 1999) given without time reference. This means that it is supposed that the system is in equilibrium or that the quantities are examined locally in time.

The classical way to find this expression is much more simple and the hypotheses that are made are not directly understood. The system is first formulated as a 'mono-variable' forced dynamical system $\Delta X_{\mathrm{d}}(t_0) \to \Delta X_{\mathrm{d}}(t_0 + \Delta t)$. The total gain of this system is defined as $GH$ which represents the feedback loops plus the linear gain, where

$$G = \frac{\partial X_{\mathrm{d}}(t_0 + \Delta t)}{\partial X_{\mathrm{e}}(t_0)}$$

is the gain without feedbacks of the system $\Delta X_{\mathrm{e}} \to \Delta X_{\mathrm{d}}$, and

$$H = \sum_{i \neq d, i \neq e} \frac{\partial X_{\mathrm{e}}(t_0 + 2\Delta t)}{\partial X_i(t_0 + \Delta t)} \frac{\partial X_i(t_0 + \Delta t)}{\partial X_{\mathrm{d}}(t_0)}$$

represents the feedbacks. The forcing of the variable $X_{\mathrm{d}}$ is given by $GE_{X_{\mathrm{e}}}$. In the limit of decreasing time steps, we could use Eq. (11) to obtain

$$\Delta X_{\mathrm{d}} = \frac{G}{1 - GH} E_{X_{\mathrm{e}}}. \quad (23)$$

This expression converges to the continuous case as $\Delta t \to 0$. In the original field where this formalism was developed, i.e. the analysis of electrical circuits (Bode 1945), the relation Eq. (14) is instantaneous since the electricity propagates (almost) instantly. In this continuous case, the time reference in Eq. (14) can be suppressed. The same remark holds if the system is in equilibrium, i.e. if the previous perturbations are constant in time. Thus, this analysis has to be done locally in time or at equilibrium.

The gain of the system, $G_f$, is very sensitive to the estimation of the feedback factors $f_i$. Furthermore, it is very important to estimate all these factors simultaneously since the effect of one particular feedback is sensitive to the presence or absence of other feedbacks.

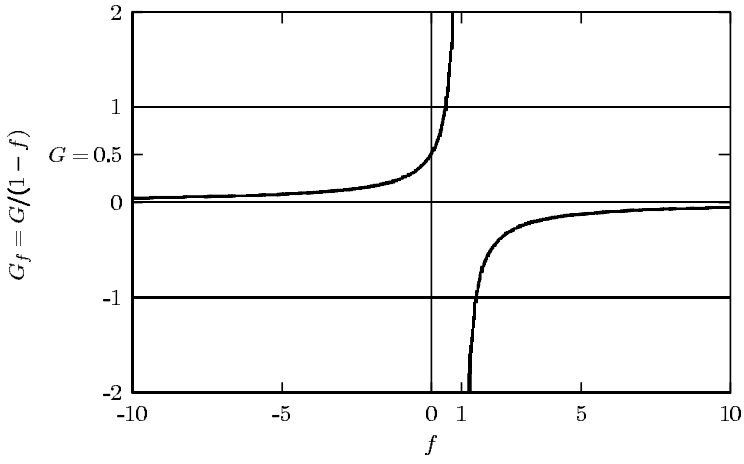



Figure 3. Analysis of the gain $G_f$ of the system as a function of the unique feedback factor $f$; the gain without feedback $G = 0.5$.

Figure 3 shows the gain of a simple system, $G_f$, as a function of the feedback factor $f$, supposing that the gain of the system without feedback is $G = 0.5$. If $f < 0$, the gain with feedback is reduced: $0 < G_f < G$. If $f = 0$, the gain with feedback is unchanged: $G_f = G$. If $0 < f < 1$, the gain of the system with feedback is increased: $G_f > G$, and $\lim_{f \to 1} = +\infty$ (the system becomes unstable). If $f > 1$, $G_f$ is negative, so the system oscillates, and it is unstable if $G_f < -1$. We see in Fig. 3 how the effect of a feedback factor on the system can be highly nonlinear. So the significance of a feedback factor is strongly dependent on the availability of the feedback factors of *all* variables: an isolated feedback factor cannot characterize any relevant behaviour of the whole system.

### (*e*) *The classical example*

The following example has been extensively used in the climate literature. We suppose that the global-mean net radiative flux (solar minus terrestrial) at the top of the atmosphere (TOA) is in equilibrium ($\Delta F_{\mathrm{TOA}} = 0$). The question is: if an external forcing is introduced into the system, how will the system react? The global-mean surface temperature $T_{\mathrm{s}}$ is taken as the diagnosed variable since a lot of other internal variables of the system are (assumed to be) dependent on this variable. Then, we can analyse the feedback process loops acting on $T_{\mathrm{s}}$ using the above formalism and assuming that they all act in parallel.

A forcing $E_{X_{\mathrm{ext}}}$ is introduced onto an external variable, $X_{\mathrm{ext}}$ (i.e. the solar insolation, volcanic eruptions, etc.). We analyse the system:

$$F_{\mathrm{TOA}}(t + \Delta t) = F(X_{\mathrm{ext}}(t),\, X_i(t),\, T_{\mathrm{s}}(t)). \tag{24}$$

The terms $X_i$ are the internal variables of the system (i.e. that depend on the surface temperature, $X_i = X_i(T_{\mathrm{s}})$) like the albedo, the water vapour, the lapse rate, the clouds, etc.

We suppose here that it is possible to express the external forcing $E_{X_{\mathrm{ext}}}$, in terms of perturbations of the net radiative flux, $E_{\mathrm{TOA}}$. The forcing introduces perturbations onto the variables of the system; the link between these perturbations can be expressed by

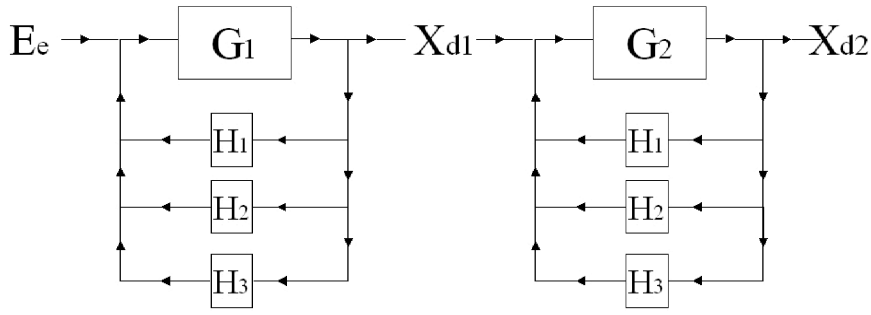



Figure 4. Feedback-loops system in series: $E_e$ is the external forcing in variable $X_e$, $G_1$ and $G_2$ are the linear gains of the systems $X_e \rightarrow X_{d1}$ and $X_{d1} \rightarrow X_{d2}$, $X_{d2}$ is the 'diagnose' variable where the impact of the forcing is observed, and the coefficients $H_i$ represent the feedbacks.

(see Eq. (16)):

$$\Delta F_{\mathrm{TOA}}(t_0 + 2\Delta t) = \underbrace{E_{\mathrm{TOA}}(t_0 + 2\Delta t)}_{\text{external forcing}} + \underbrace{\sum_i \frac{\partial F_{\mathrm{TOA}}(t_0 + 2\Delta t)}{\partial X_i(t_0 + \Delta t)} \frac{\partial X_i(t_0 + \Delta t)}{\partial T_s(t_0)}}_{\text{feedback loops}} \Delta T_s(t_0). \tag{25}$$

If the equilibrium state is reached, or if the sensitivities are instantaneous, the reference to time can be suppressed:

$$\Delta F_{\mathrm{TOA}} = E_{\mathrm{TOA}} + \left(\sum_i H_i\right) \Delta T_s, \tag{26}$$

where the terms $H_i$ are the products of first derivatives describing the cause-and-effect relations in Eq. (25). By multiplying this expression by the gain of the system without feedbacks, $G = \partial T_s / \partial F_{\mathrm{TOA}}$, we finally obtain the following familiar expression:

$$\Delta T_s = \frac{G}{1 - G \sum_i H_i} E_{\mathrm{TOA}} = \frac{G}{1 - \sum_i f_i} E_{\mathrm{TOA}}. \tag{27}$$

### (*f*) *The classical analysis in a series feedback configuration*

It is supposed again that the external forcing $E_{X_e}$ of the system acts on only one variable $X_e$ of the system. There are two diagnosed variables: $X_{d1}$ and $X_{d2}$, $X_{d2}$ being dependent on $X_{d1}$. Some of the internal variables $\{X_{i1} = X_{i1}(X_{d1})\}$ are dependent on $X_{d1}$, and some others $\{X_{i2} = X_{i2}(X_{d2})\}$ are dependent on $X_{d2}$. The impact of the external forcing is observed on the diagnosed variable $X_{d2}$ (Fig. 4). This internal structure describes a dynamical system $X_e \rightarrow X_{d1} \rightarrow X_{d2}$, with feedbacks in series. In this case, the gain of the subsystems $X_e \rightarrow X_{d1}$ and $X_{d1} \rightarrow X_{d2}$ would be computed as in section 2(d). Then the global gain of the system would be $G_f = G_{f_2} G_{f_1}$.

### (*g*) *Comments on the classical feedback analysis*

We have seen in the two previous subsections that where particular cause-and-effect relations in the system are known (or assumed), the time reference is required in the discrete case, but can be suppressed in two situations:

(i) In an *equilibrium state*: the perturbations are constant $\partial X/\partial t = 0$ (not to be confused with zero forcing), so they are the same at each time step. The feedback analysis is then only a characterization of the equilibrium state. There is no estimation of the time required to reach the equilibrium and we cannot predict the transient behaviour of the system. Furthermore, we do not know a priori the sensitivities in the equilibrium state, so we are required to assume (without evidence) that the sensitivities are constant and that we have a good estimate of them. Finally, the system does not have to be in static equilibrium; they may be unforced variations.

(ii) When the sensitivities are *instantaneous*: the relations between the perturbations of each variable of the system are then valid without a time reference. But in this case, instantaneous estimates of the sensitivities are required and the feedback factors have to be computed at each time because of their state dependence. To our knowledge, this approach has not yet been investigated since no technique has been available to estimate these instantaneous, multivariate and nonlinear sensitivities.

The classical approach to feedback analysis from the electrical circuits theory (Bode 1945) was first used on simple energy-balance models of the climate where instantaneous sensitivities are available because they can be evaluated directly from the simple equations of these models. Even if the estimation of sensitivities was crude, the applicability of the technique was justified when the cause-and-effect relationships were supposed to be known. In more recent studies, and particularly in the analysis of observations, this approach to the estimation of sensitivities is highly questionable. In particular, the use of this characterization of the equilibrium state to predict the system response to an external forcing is inappropriate since the sensitivities used to produce the equilibrium state are unknown. Some of the limitations of actual studies are as follows:

(i) *Model used.* The hierarchical model of cause-and-effect relations, described by greatly simplified relations among the sensitivities, is usually much too simple. For example, the fact that the forcing/gain/response system has to be mono-variable is a very strong simplification: such assumptions result in the suppression/neglect of some perturbations and some first derivatives in the system. Moreover, it is usually assumed that the feedbacks operate only in parallel, which is not general.

(ii) *Estimation of sensitivities.* The sensitivities are often estimated by finite differences between two (usually equilibrium) states of the system. First, this approach measures only the coincidence of the changes in two quantities, but it does not mean that there is a cause-and-effect relationship between these variables. The relationships might also be indirect (via the ordering of the dependencies). Second, this approach measures the changes in two quantities and the sensitivity is then estimated assuming that the other variables do not interact. This is a strong limitation since there are a lot of cross-linkages in the variables of the climate system. Third, the finite difference for the estimation of the sensitivities can be highly misleading if the sensitivities of the system are not constant in time.

(iii) *Forcing process.* The forcing model is often not expressed: is it localized in time, constant, growing in time, cyclic, etc.? The way that the external forcing evolves in time is also important for the study of the transient response.

(iv) *Better description.* Previous approaches to feedback analysis are often only a characterization of the equilibrium state of the system after the introduction of an external forcing. The transient period between the beginning of the forcing and the equilibrium state is not described, the time to reach the equilibrium is not estimated.

This is a real drawback for the understanding of these phenomena. Furthermore, the gain of the system with feedback factors is highly dependent on the precision of the sensitivity estimates.

In conclusion, the actual application of the classical feedback analysis is limited by some very strong assumptions like linearity (i.e. sensitivities constant in time), static equilibrium, mono-variable cause-and-effect relationships, etc., and so does not seem at all appropriate for application to the climate system. Moreover, since the resulting expressions for the feedback factors are products of the instantaneous sensitivities, it would seem more straightforward to evaluate these sensitivities instead. To avoid the classical limitations, the general feedback formulation should be used to evaluate the nonlinear, multivariate and instantaneous sensitivities in both numerical models and observations. We have developed a method to estimate these sensitivities from the time evolution of the system state which we describe in the next section.

## 3. A NONLINEAR REGRESSION SCHEME FOR ESTIMATION OF SENSITIVITIES

To estimate the sensitivities of the dynamical model in Eq. (1), we use a multivariate nonlinear regression fit to the statistics produced by observing the behaviour of the system over a time period long enough to provide a good sample of the different states of the system. Any multivariate nonlinear regression technique, such as spline interpolation or ARMAX* models, etc., could be used. For this purpose, we use a neural-network (NN) technique because of its ability to process large-dimension datasets (which will be helpful for further experiments on numerical models) and its capacity to integrate a priori information about the problem (Aires 1999). This technique has been used extensively to estimate physical relationships such as inverse radiative-transfer models (see, for example, Aires *et al.* 2002b,c,d).

### (*a*) *The neural-network model*

The Multi-Layer Perceptron (MLP) network is a mapping model composed of parallel processors called 'neurons'. These processors are organized in distinct layers: the first layer (number 0) represents the input $\mathbf{P} = (p_i;\ 0 \leq i \leq m_0)$ with $m_0$ the number of neurons in layer 0. The last layer (number $L$) represents the output mapping $\mathbf{X} = (x_k;\ 0 \leq k \leq m_L)$. The intermediate layers ($0 < m < L$) are called the 'hidden layers'. These layers are connected via neuronal links (Fig. 5): two neurons, $i$ and $j$, between two consecutive layers have synaptic connections associated with a synaptic weight $w_{ij}$.

Each neuron $j$ executes two simple operations: first, it makes a weighted sum of its inputs from the previous layer $z_i$; this signal is called the activity of the neuron:

$$a_j = \sum_{i \in \text{Inputs}(j)} w_{ij} z_i . \tag{28}$$

Then, it transfers this signal to its output through a so-called 'transfer function', often a sigmoid function such as $\sigma(a) = \tanh(a)$. The output $z_j$ of neuron $j$ in the hidden layer is then given by

$$z_j = \sigma\left( \sum_{l \in \text{Inputs}(j)} w_{lj} z_l \right).$$

Generally, for regression problems, the neurons in the output (last) layer have a transfer function of identity. For example, in a one-hidden-layer MLP (Fig. 5), the $k$th output $x_k$

* AutoRegressive Moving-Average with eXternal process.

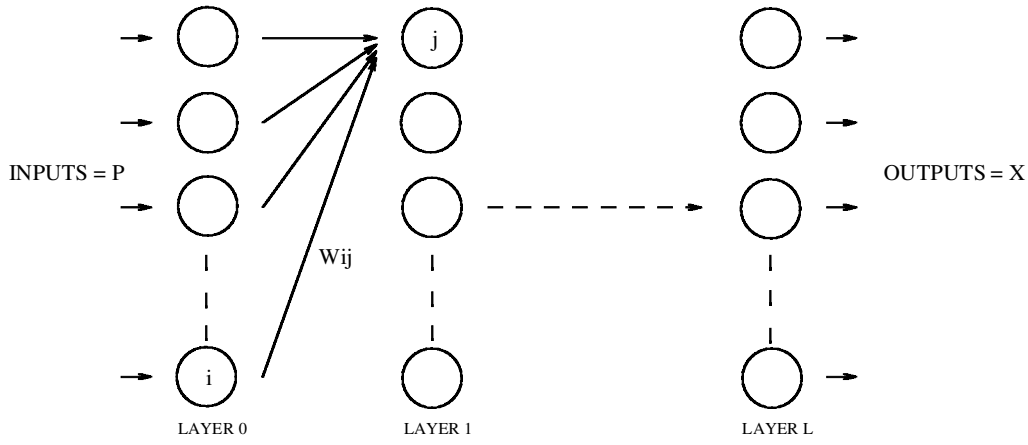



Figure 5. Architecture of a multi-layer perceptron neural network with $L$ layers, with synaptic weights $w_{ij}$, inputs $\mathbf{P}$ and outputs $\mathbf{X}$.

of the network is defined as

$$x_k(y) = \sum_{j \in S_1} w_{jk}\sigma(a_j) = \sum_{j \in S_1} w_{jk}\sigma\left(\sum_{i \in S_0} w_{ij}\, p_i\right), \tag{29}$$

where $\sigma$ is the sigmoid function, $a_j$ is the activity of neuron $j$, and $S_i$ is the $i$th layer of the network (with $i = 0$ for the input layer). We have deliberately omitted the usual bias term in this formula for clarity, but include it in the actual network.

The key to our analysis is that any continuous function can be represented by a one-hidden-layer MLP with this kind of sigmoid transfer function (Hornik *et al.* 1989; Cybenko 1989). Hence the process of training the MLP to fit the observed multivariate, nonlinear relationship statistics is equivalent to deriving a multivariate, nonlinear function that behaves as the dynamical system in question. The key advantage of the NN approach over some other methods is that the Jacobians (i.e. sensitivities) can be evaluated directly from the MLP (see later).

### (*b*) *The learning algorithm*

Given a neural architecture (number of layers, neurons and connections, type of transfer functions), all the information of the network is contained in the set of synaptic weights $w_{ij}$. The learning algorithm is an optimization technique that estimates the network parameters $W = \{w_{ij}\}$ by minimizing a loss function, $C(W)$, required to fit the desired function defined by observations as closely as possible. The criterion usually used to adjust $W$ is the mean-square error in network outputs:

$$C(W) = \tfrac{1}{2} \sum_{k=1}^{m_L} \iint (x_k(\mathbf{P};\, W) - t_k)^2 H(t_k/\mathbf{P}) H(\mathbf{P})\, \mathrm{d}t_k\, \mathrm{d}\mathbf{P}, \tag{30}$$

where $t_k$ is the $k$th desired output component, $x_k$ the $k$th neural output component, $H(t_k/\mathbf{P})$ the probability function of output $t_k$ given the input $\mathbf{P}$, and $H(\mathbf{P})$ the probability density function of input data $\mathbf{P}$. If specific a priori information about the probability distribution functions is available, quality criteria other than least-squares can be used. For example, criteria involving higher-order statistics have been defined (Aires *et al.*

2000). Practically, $C(W)$ is approximated by the classical least-square criterion:

$$\overline{C}(W) = \frac{1}{2E} \sum_{e=1}^{E} (x_k(\mathbf{P};\, W) - t_k)^2. \tag{31}$$

The error back-propagation algorithm (Rumelhart *et al.* 1986) is used to minimize $\overline{C}(W)$. It is a stochastic steepest descent (i.e. Newtonian minimization procedure) very well adapted to the MLP hierarchical architecture because the computational cost is only linearly related to the number of parameters.

### (*c*) *The neural Jacobians*

The important feature of the NN for our purpose is that the adjoint model of the neuronal model is directly available (Aires 1999; Aires *et al.* 1999, 2001). The computation of this adjoint model (or neural Jacobians) is accurate and very fast. Since the NN is nonlinear, these Jacobians depend on the situation $x$. For example, the neural Jacobians in the previous example of Eq. (29) (a MLP network with one hidden layer) are

$$\frac{\partial x_k}{\partial p_i} = \sum_{j \in S_1} w_{jk} \frac{\partial \sigma}{\partial a} \left( \sum_{i \in S_0} w_{ij}\, p_i \right) w_{ij}, \tag{32}$$

where $\partial \sigma / \partial a$ is the derivative of the transfer function $\sigma$. For a more complex MLP network with many hidden layers, there still exists a back-propagation algorithm for efficiently computing these neural Jacobians.

The neural Jacobians concept is a very powerful tool because it allows for the direct statistical evaluation of the multivariate and nonlinear sensitivities of the dynamical system under study.

## 4. Analysis of the discrete Lorenz model

To test the definitions and the technique previously presented, we apply it to a simple nonlinear, multivariate, chaotic, non-stationary and forced dynamical model for which the sensitivities are known analytically. We choose here a discrete form of the low-order Lorenz model (Lorenz 1984). This model is very general since it is not a mono-variable structure, as described in sections 2(d) and (f), and it exhibits very complex behaviour. Nonetheless, we can define the time relationships directly from the equations of the model to test our ability to infer these relationships from the observed behaviour (model output). We have discretized the Lorenz continuous model to make it easier to describe the cause-and-effect relations of the feedback processes.

### (*a*) *Continuous Lorenz model*

The low-order model used in this study was developed by Lorenz (Lorenz 1984; Lorenz 1990) to analyse the chaos and stability assumptions about the atmospheric circulation. This simple model is able to represent the Hadley circulation and is used to determine the stability or the instability of this circulation (stationary or migratory disturbance). This model is defined by three ODEs:

$$\left.\begin{aligned} \mathrm{d}X(t)/\mathrm{d}t &= -Y^2(t) - Z^2(t) - aX(t) + aF_1, \\ \mathrm{d}Y(t)/\mathrm{d}t &= X(t)Y(t) - bX(t)Z(t) - Y(t) + F_2, \\ \mathrm{d}Z(t)/\mathrm{d}t &= bX(t)Y(t) + X(t)Z(t) - Z(t), \end{aligned}\right\} \tag{33}$$

where

- $t$ is the time (equivalent in units to about 1 day);
- $X$ is the intensity of the symmetric, globe-encircling, westerly wind current and also the poleward temperature gradient (assumed to be in equilibrium with it);
- $Y$ is the cosine phase of a series of superposed large-scale eddies, which transport heat poleward at a rate proportional to the square of their amplitudes;
- $Z$ is the sine phase of a series of superposed large-scale eddies, which transport heat poleward at a rate proportional to the square of their amplitudes;
- $F_1$ is a zonally symmetric thermal forcing on $X$;
- $F_2$ is a zonally asymmetric thermal forcing on $Y$.

The two forcings $F_1$ and $F_2$ are the values to which $X$ and $Y$ would be driven if the westerly current and the eddies were not coupled.

The discretization of these ODEs is a very delicate process, but the Runge–Kutta fourth-order technique can be used for this purpose. Figure 6 shows the integration of Eq. (33) from $t_0 = 0$ to $T = t_0 + N\Delta t$, using $a = 0.25$, $b = 4$, $F_1 = 8$, $F_2 = 1$ and $\Delta t = 0.08$. The initial state of the system at time $t = 0$ is taken as $X(0) = 1.31$, $Y(0) = 1.48$ and $Z(0) = 0.34$. Lorenz has shown that this system with these parameter values exhibits chaotic behaviour.

### (*b*) *Discretization of the dynamical system*

We are not interested in a perfect simulation of the Lorenz model; rather, we are interested in a representation of this system in a form like

$$\begin{Bmatrix} X(t+\Delta t) \\ Y(t+\Delta t) \\ Z(t+\Delta t) \end{Bmatrix} = \mathcal{A} \begin{Bmatrix} X(t) \\ Y(t) \\ Z(t) \end{Bmatrix} \tag{34}$$

as a test of our analysis technique. By discretizing Eqs. (33) with the forward Euler step (Runge–Kutta fourth-order technique), we obtain

$$\left.\begin{aligned} X(t+\Delta t) &= \Delta t(-Y(t)^2 - Z(t)^2 + aF_1) + (1 - a\Delta t)X(t), \\ Y(t+\Delta t) &= \Delta t(-bX(t)Z(t) + F_2), +(1 - \Delta t + \Delta t X(t))Y(t), \\ Z(t+\Delta t) &= \Delta t\, bX(t)Y(t) + (1 + \Delta t X(t) - \Delta t)Z(t), \end{aligned}\right\} \tag{35}$$

where $\Delta t$ is the discrete time step. Discretization schemes other than the forward Euler scheme can be used (improved Euler, Crank–Nicholson, backward Euler step, etc.). Using a different finite-difference scheme would change the results we obtain by introducing higher-order terms. This is normal since different simulation schemes result in different systems, which result in different sensitivities/feedbacks. For our purpose, we take Eqs. (35) to be the exact model.

The size of $\Delta t$ needs to be sufficiently small so that the linearization of the system during a single time step is accurate, i.e. so that the hypothesis that the Jacobians of the system are constant during the time interval is true. The time discretization is also directly related to the regularity of the Jacobians of the system: high complexity requires small time steps to ensure a good description of the evolution of the Jacobians. We take $\Delta t = 0.08$ (when re-scaled, equivalent to two hours); this time step leads to a good simulation of the Lorenz system.

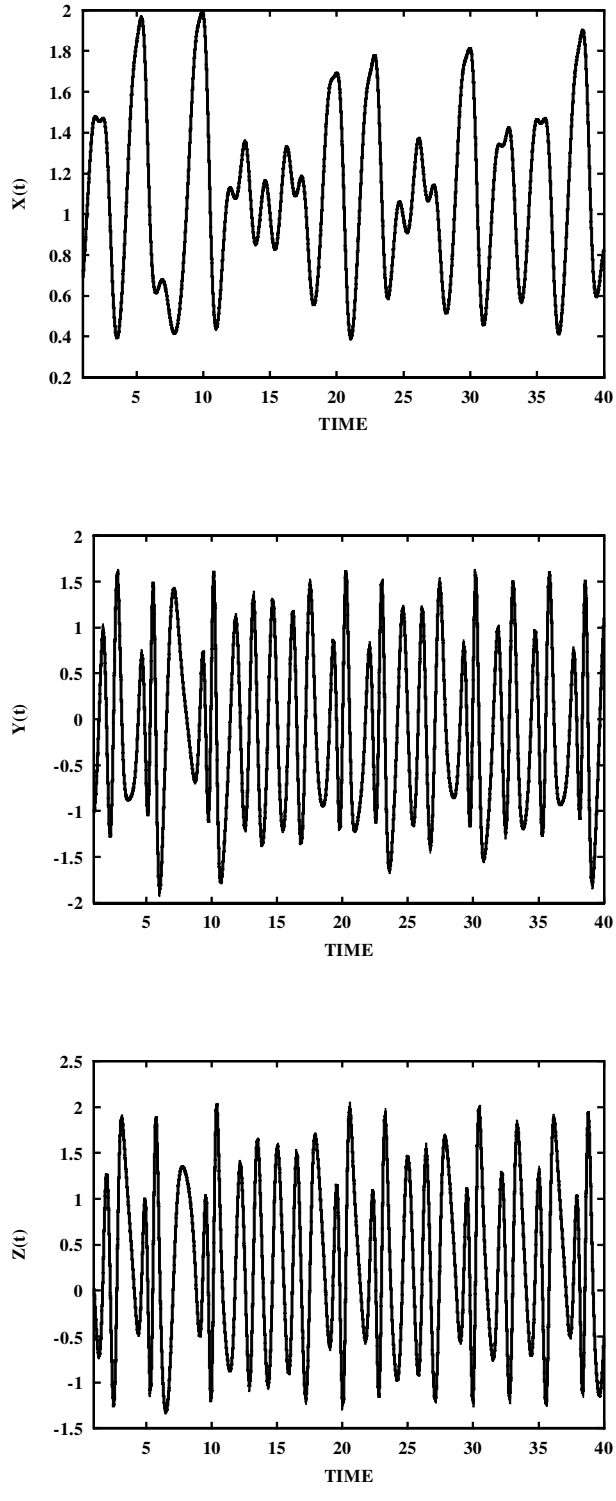



Figure 6. Time evolution of the state variables $X$, $Y$ and $Z$ of the Lorenz model, with parameters $a = 0.25$, $b = 4$, $F_1 = 8$, $F_2 = 1$ and $\Delta t = 0.08$, simulated by fourth-order Runge–Kutta. See text for explanation.

### (c) *Sensitivities of the dynamical system*

The Jacobian matrix of the discrete system is

$$\mathbf{G}\begin{Bmatrix} X(t+\Delta t) \\ Y(t+\Delta t) \\ Z(t+\Delta t) \end{Bmatrix} = \begin{pmatrix} \dfrac{\partial X(t+\Delta t)}{\partial X(t)} & \dfrac{\partial X(t+\Delta t)}{\partial Y(t)} & \dfrac{\partial X(t+\Delta t)}{\partial Z(t)} \\ \dfrac{\partial Y(t+\Delta t)}{\partial X(t)} & \dfrac{\partial Y(t+\Delta t)}{\partial Y(t)} & \dfrac{\partial Y(t+\Delta t)}{\partial Z(t)} \\ \dfrac{\partial Z(t+\Delta t)}{\partial X(t)} & \dfrac{\partial Z(t+\Delta t)}{\partial Y(t)} & \dfrac{\partial Z(t+\Delta t)}{\partial Z(t)} \end{pmatrix} \tag{36}$$

$$= \begin{pmatrix} 1 - a\Delta t & -2\Delta t Y(t) & -2\Delta t Z(t) \\ -\Delta t\, bZ(t) + \Delta t Y(t) & 1 - \Delta t + \Delta t X(t) & -b\Delta t X(t) \\ \Delta t\, bY(t) + \Delta t Z(t) & \Delta t\, bX(t) & 1 + \Delta t X(t) - \Delta t \end{pmatrix}. \tag{37}$$

These Jacobians are dependent on the state of the system (in this sense, we say that the sensitivities are nonlinear), so they are also dependent on time. Thus, the hypothesis of constant Jacobians, as in classical feedback analysis, cannot be used to understand this system. For our analysis we assume only that Jacobians are constant over a single time step, but the elements in Eq. (36) still vary with time.

Local sensitivities are a linearization of the nonlinear dynamical system but we estimate these sensitivities over the whole state space. We estimate the global sensitivity $G(\cdot)$, which is state dependent. The sensitivity $G(a_1 x1 + a_2 x2)$ is not equal to $a_1 G(X_1) + a_2 G(X_2)$, which means that the sensitivity operator $G$ is nonlinear.

### (d) *Theoretical feedback analysis*

The two external forcing, $aF_1$ on $X$ and $F_2$ on $Y$, are continuous and constant in Eq. (33). In the discrete formalization, this is represented by

$$\left.\begin{aligned} \Delta X(t_0 + k\Delta t) &= \Delta t\; aF_1, \\ \Delta Y(t_0 + k\Delta t) &= \Delta t F_2, \end{aligned}\right\} \quad \text{for } k = 1, \ldots, N. \tag{38}$$

If the beginning state of the simulation is chosen as ($X(t_0) = 0$, $Y(t_0) = 0$, $Z(t_0) = 0$), then the state of the system at the next time step is given by

$$\left.\begin{aligned} X(t_0 + \Delta t) &= \Delta t\; aF_1, \\ Y(t_0 + \Delta t) &= \Delta t F_2, \\ Z(t_0 + \Delta t) &= 0, \end{aligned}\right\} \tag{39}$$

and, for the next time step,

$$\left.\begin{aligned} X(t_0 + 2\Delta t) &= 2\Delta t\; aF_1 - 2\Delta t^3 {F_2}^2 - \Delta t^2 a^2 F_1, \\ Y(t_0 + 2\Delta t) &= 2\Delta t F_2 - \Delta t^2 F_2 + 2\Delta t^3 a F_1 F_2, \\ Z(t_0 + 2\Delta t) &= \Delta t^3 ab F_1 F_2 \end{aligned}\right\} \tag{40}$$

and so on. We analyse the impacts of the external forcings, $aF_1$ and $F_2$, on the diagnosed variable, chosen here to be $X$ for illustration.

The perturbation at time $t_0 + \Delta t$, $\Delta X(t_0 + \Delta t) = \Delta t\, aF_1$, is straightforward. At time $t_0 + 2\Delta t$, without feedbacks, the forcing would simply be added:

$$\Delta X(t_0 + 2\Delta t) = 2\Delta t\, aF_1. \tag{41}$$

With feedbacks, the true perturbation is given by

$$\begin{aligned}\Delta X(t_0 + 2\Delta t) &= E_X(t_0 + 2\Delta t) + \frac{\partial X(t_0 + 2\Delta t)}{\partial X(t_0 + \Delta t)} E_X(t_0 + \Delta t) \\ &\quad + \frac{\partial X(t_0 + 2\,\Delta t)}{\partial Y(t_0 + \Delta t)} E_Y(t_0 + \Delta t) \\ &= 2\Delta t\, aF_1 - 2\Delta t^3 {F_2}^2 - \Delta t^2 a^2 F_1. \end{aligned} \tag{42}$$

Comparing Eqs. (41) and (42), we note the presence of two correction factors giving the contribution of the 'direct' feedback processes: these feedbacks are caused only by the integration of the variables over time and the fact that the effects change with time. This expression for the perturbation is in agreement with the first part of Eq. (40).

For the description of the other ('indirect') feedbacks, three time steps are required. At time $t_0 + 3\Delta t$, the integration of the external forcings is even more complex:

$$\begin{aligned}\Delta X(t_0 + 3\Delta t) = {} & \text{external forcing} + \text{direct feedbacks} \\ & + \text{indirect feedbacks 1} + \text{indirect feedbacks 2},\end{aligned} \tag{43}$$

where

$$\text{external forcing} = E_X(t_0 + 3\Delta t),$$

$$\text{direct feedbacks} = \frac{\partial X(t_0 + 3\Delta t)}{\partial X(t_0 + 2\Delta t)} E_X(t_0 + 2\Delta t) + \frac{\partial X(t_0 + 3\Delta t)}{\partial Y(t_0 + 2\Delta t)} E_Y(t_0 + 2\Delta t),$$

$$\begin{aligned}\text{indirect feedbacks 1} = \Bigg\{ & \frac{\partial X(t_0 + 3\Delta t)}{\partial X(t_0 + 2\Delta t)} \frac{\partial X(t_0 + 2\Delta t)}{\partial X(t_0 + \Delta t)} \\ & + \frac{\partial X(t_0 + 3\Delta t)}{\partial Y(t_0 + 2\Delta t)} \frac{\partial Y(t_0 + 2\Delta t)}{\partial X(t_0 + \Delta t)} \\ & + \frac{\partial X(t_0 + 3\Delta t)}{\partial Z(t_0 + 2\Delta t)} \frac{\partial Z(t_0 + 2\Delta t)}{\partial X(t_0 + \Delta t)} \Bigg\} E_X(t_0 + \Delta t),\end{aligned}$$

$$\begin{aligned}\text{indirect feedbacks 2} = \Bigg\{ & \frac{\partial X(t_0 + 3\Delta t)}{\partial X(t_0 + 2\Delta t)} \frac{\partial X(t_0 + 2\Delta t)}{\partial Y(t_0 + \Delta t)} \\ & + \frac{\partial X(t_0 + 3\Delta t)}{\partial Y(t_0 + 2\Delta t)} \frac{\partial Y(t_0 + 2\Delta t)}{\partial Y(t_0 + \Delta t)} \\ & + \frac{\partial X(t_0 + 3\Delta t)}{\partial Z(t_0 + 2\Delta t)} \frac{\partial Z(t_0 + 2\Delta t)}{\partial Y(t_0 + \Delta t)} \Bigg\} E_Y(t_0 + \Delta t).\end{aligned}$$

We note in this expression some terms that do not appear in the classical analysis formalism. For example, the direct feedbacks terms (due to time integration of the variables) are suppressed in the classical analysis. Furthermore, we see that in this expression both forcings (on variable $X$ and on variable $Y$) are taken into account, which is not possible in the classical approach.

Integrating the system for one more time step would be highly complex, this is the reason why analysis of this kind of dynamical system is a difficult problem. To perform

prediction, the model needs to represent the sensitivities with a high degree of precision. Otherwise, an error at one time step is rapidly amplified in the next time steps.

The classical formalism for the feedback analysis is not well adapted to the analysis of the Lorenz model since there is no preferred variable on which the other two variables of the model depend solely. So we already see in this simple example how limited the assumptions used in the classical feedback analysis formalism are and how such an analysis could be very misleading. Again, it is clear that evaluation of the sensitivities is more straightforward than evaluation of feedback factors, which are products of sensitivities. However, for illustrative purposes we will use the classical formalism to calculate feedback factors just because they are more familiar. If we choose the variable $Y$ as the variable affected by the external forcing and $X$ as the diagnosed variable, the gain of the system $E_Y \rightarrow \Delta X$ is given by (see Eq. (23))

$$\Delta X = \frac{G}{1 - GH} E_Y = G_f E_Y \tag{44}$$

$$= \frac{G}{1 - \sum_i f_i^{XY}} E_Y, \tag{45}$$

where

- $G = \partial X / \partial Y$ is the gain without feedbacks of the system $E_Y \rightarrow \Delta X$;
- $H = \sum_i (\partial Y / \partial X_i)(\partial X_i / \partial X)$.

The three feedback factors for this mono-variable system are defined as

$$f_X^{YX} = \frac{\partial X}{\partial Y} \frac{\partial Y}{\partial X}, \tag{46}$$

$$f_Y^{YX} = \frac{\partial X}{\partial Y} \frac{\partial Y}{\partial Y} \frac{\partial Y}{\partial X}, \tag{47}$$

$$f_Z^{YX} = \frac{\partial X}{\partial Y} \frac{\partial Y}{\partial Z} \frac{\partial Z}{\partial X}. \tag{48}$$

Note that the sensitivities used in this relation still are dependent on time (and have to be estimated precisely), so that the feedback factors are also time dependent. As we will show, this fundamental property of complex, nonlinear dynamical systems reduces the value of the classical (linear) feedback analysis for understanding the system behaviour. Note again that the above quantities are not the true feedback factors for the Lorenz model since they are defined using invalid assumptions. In particular, the cause-and-effect structure of variable relationships are not as simple as described in the classical parallel feedback scheme (see section 2(d)).

## 5. Experimental results

### (*a*) *Construction of the dataset*

The quality of the dataset used to evaluate the sensitivities is a crucial issue. For example, using data from a system in equilibrium or from a system during a transient change may not give the same results in the analysis. Ideally, a good dataset would be one including all ranges of variability for all combinations of the variables of the system. The more situations that are included in the dataset, the larger will be the range of validity of the sensitivity estimates. This situation parallels that in climate analysis where the range of validity is limited by the range of climate states actually observed.

The discrete dynamical version of the Lorenz model stabilizes more rapidly onto a limit cycle than the continuous version. So, to create a dataset closer to the behaviour of the continuous system, we chose randomly 200 noisy states from the Runge–Kutta integration of the continuous Lorenz model. These 200 states of the system are used as initial states for 200 trajectories of 1000 time steps each from the discrete system in Eq. (35). The final dataset is then composed of $N = 200\,000$ couples $\{(\mathbf{I}^k, \mathbf{O}^k);\ k = 1, \ldots, N\}$, where

$$\mathbf{I}^k = (X(t_0 + k\Delta t),\ Y(t_0 + k\Delta t),\ Z(t_0 + k\Delta t))$$

is an $N \times 3$ matrix of the inputs of the system and

$$\mathbf{O}^k = (X(t_0 + (k+1)\Delta t),\ Y(t_0 + (k+1)\Delta t),\ Z(t_0 + (k+1)\Delta t))$$

is an $N \times 3$ matrix of the outputs. Each couple is linked by $\mathbf{O}^k = \mathcal{A}(\mathbf{I}^k)$.

The parameters for the Lorenz model are the same as previously: $a = 0.25$, $b = 4$, $F_1 = 8$, $F_2 = 1$ and $\Delta t = 0.08$, but we have introduced Gaussian-distributed noise $\mathcal{N}(0, 0.001)$ at each time step and in each variable during the simulation in order to be closer to an experiment with observation errors. Figure 7 shows the resulting noisy trajectories included in the dataset.

### (*b*) *Linear and nonlinear regressions*

If a priori information is available to define good predictors, the dynamical system can be described as a linear model. In the Lorenz case, the good predictors, $\mathbf{P}(t)$, of the general model Eqs. (1) can be determined directly from the equations of the model Eqs. (35):

$$\mathbf{P}(t) = (X(t),\ Y(t),\ Z(t),\ Y^2(t),\ Z^2(t),\ X(t)Y(t),\ X(t)Z(t),\ F_1,\ F_2). \tag{49}$$

In this configuration, the dynamical system of Eq. (35) becomes

$$\begin{Bmatrix} X(t+\Delta t) \\ Y(t+\Delta t) \\ Z(t+\Delta t) \end{Bmatrix} = \mathbf{A}(X(t),\ Y(t),\ Z(t),\ Y^2(t),\ Z^2(t),\ X(t)Y(t),\ X(t)Z(t),\ F_1,\ F_2), \tag{50}$$

where the constant matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{pmatrix} 1 - a\Delta t & 0 & 0 & -\Delta t & -\Delta t & 0 & 0 & a\Delta t & 0 \\ 0 & 1 - \Delta t & 0 & 0 & 0 & -\Delta t & -b\Delta t & 0 & \Delta t \\ 0 & 0 & 1 - \Delta t & 0 & 0 & b\Delta t & \Delta t & 0 & 0 \end{pmatrix}. \tag{51}$$

A linear regression in this case would give a good estimate of the elements of matrix $\mathbf{A}$. This is a very general idea: all nonlinear dynamical systems could be simplified, and even linearized, if all of the good predictors (all the terms in the equations) are known.

In practice, this a priori information is not available, so choosing good variables to predict system behaviour is a key question that has no general answer. Usually, then, the predictors are chosen as the state variables; model Eqs. (1) becomes

$$\begin{Bmatrix} X(t+\Delta t) \\ Y(t+\Delta t) \\ Z(t+\Delta t) \end{Bmatrix} = \mathcal{A} \begin{Bmatrix} X(t) \\ Y(t) \\ Z(t) \end{Bmatrix}. \tag{52}$$

Now, a linear regression analysis approximates the nonlinear function $\mathcal{A}$ by a linear model: $\mathcal{A}$ is replaced in Eq. (52) by a $3 \times 3$ matrix $\mathbf{A}$. This matrix is estimated by

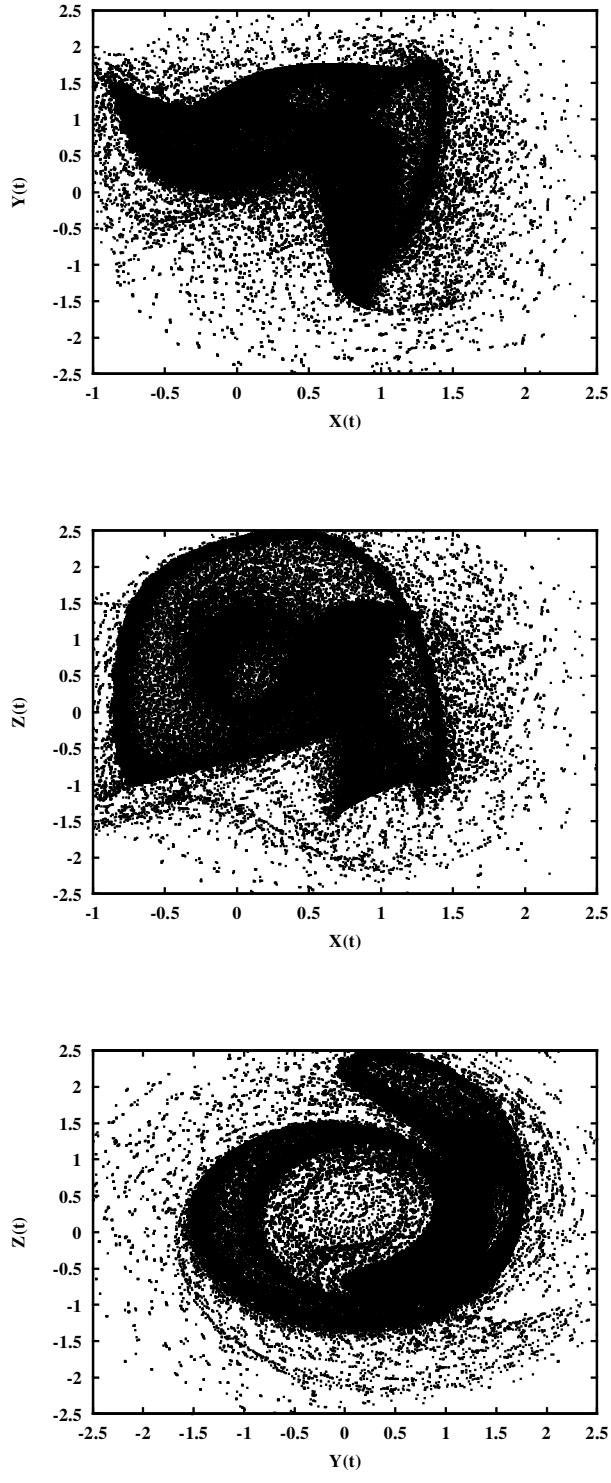



Figure 7. Noisy trajectories of the dataset from the discrete Lorenz dynamical system: $X$, $Y$ and $Z$ are the state variables of the system.

minimizing the least-squares criterion: if the square of errors criterion $C$ is imposed so that $\partial C/\partial A_{ij} = 0$, we obtain the linear expression $\mathbf{I}^{\mathrm{T}}\mathbf{I}\mathbf{A} = \mathbf{I}^{\mathrm{T}}\mathbf{O}$, which gives

$$\mathbf{A} = (\mathbf{I}^{\mathrm{T}}\mathbf{I})^{-1}\mathbf{I}^{\mathrm{T}}\mathbf{O}. \tag{53}$$

The use of this linear regression is already an improvement compared to classical approaches because it allows the simultaneous estimation of multivariate sensitivities.

For a nonlinear regression, we use an MLP network with one hidden layer. The architecture has three units in the input layer coding $\mathbf{I} = (X(t), Y(t), Z(t))$, 30 units in the hidden layer (this number was chosen by trial in the learning phase) and three units in the output layer coding the prediction,

$$\mathbf{O} = (X(t + \Delta t), Y(t + \Delta t), Z(t + \Delta t)).$$

The total number of parameters is $3 \times 30 + 30 \times 3 = 180$.

For the training of the NN (i.e. estimation of the parameters for the nonlinear regression), we have used 150 000 time-pairs randomly chosen from the dataset previously constructed and for the test data (i.e. to measure the ability of the model to generalize to unknown data), we have taken the remaining 50 000 points.

The quality of the training dataset is essential in an NN experiment. The dataset used should contain many samples of all the different states that occur. It is difficult to quantify the number of samples required to efficiently train the NN (no theoretical results being available for that); only an empirical experiment with various trials could answer to this question. In our Lorenz model experiment, the 150 000 samples dataset is sufficiently rich to represent all the situations. The selection of the number of hidden nodes is based on achieving a minimum fit error: significantly fewer or more nodes than 30 increases the fit error.

In Figure 8, the theoretical function $\mathcal{A}$ generated by integrating the nonlinear equations forward using the 4th Runge–Kutta scheme, and its two estimates (by linear and NN regressions) are illustrated. For display purposes, each plot presents one of the variables at time $t + \Delta t$, as a function of a variable at time $t$, supposing that the two other variables are equal to their mean values. It is clear that the NN regression is very precise (differences with the theoretical function are undetectable) and useful for representing nonlinear behaviour ($X(t + \Delta t)$ as a function of $Y(t)$, for example), where the linear regression is very poor. This figure shows how important the nonlinear aspect is: even the multivariate linear regression is not sufficient. The errors of the linear regression are nearly as large as the variability of the quantities, whereas the errors of the NN fit are usually about one to two orders of magnitude smaller than the variability of the quantities (except for the one sensitivity that is constant).

A dilemma that we will face in applying this technique to a real case, numerical model or observations of the climate, is that we do not know the true answer as we do here for the Lorenz model. Hence, we must develop practical ways to assess the fidelity of the analysis results. One possibility is to conduct 'prediction' experiments where we pick many specific and different episodes in the observed record (preferably time periods not included in the original analysis), initialize the NN at the beginning state, and calculate forward for a short time interval. The goal of such experiments is diagnostic, to test quantitatively whether the derived sensitivities used in the NN can reproduce the observed system dynamics in cases not included in the original analysis. We are not proposing that an NN be used for climate forecasts (i.e. be used as a statistical forecast model) in place of a physical model of the climate (see Yuval 2000) for a previous study on this subject). Rather, we are interested in whether the derived sensitivities can be

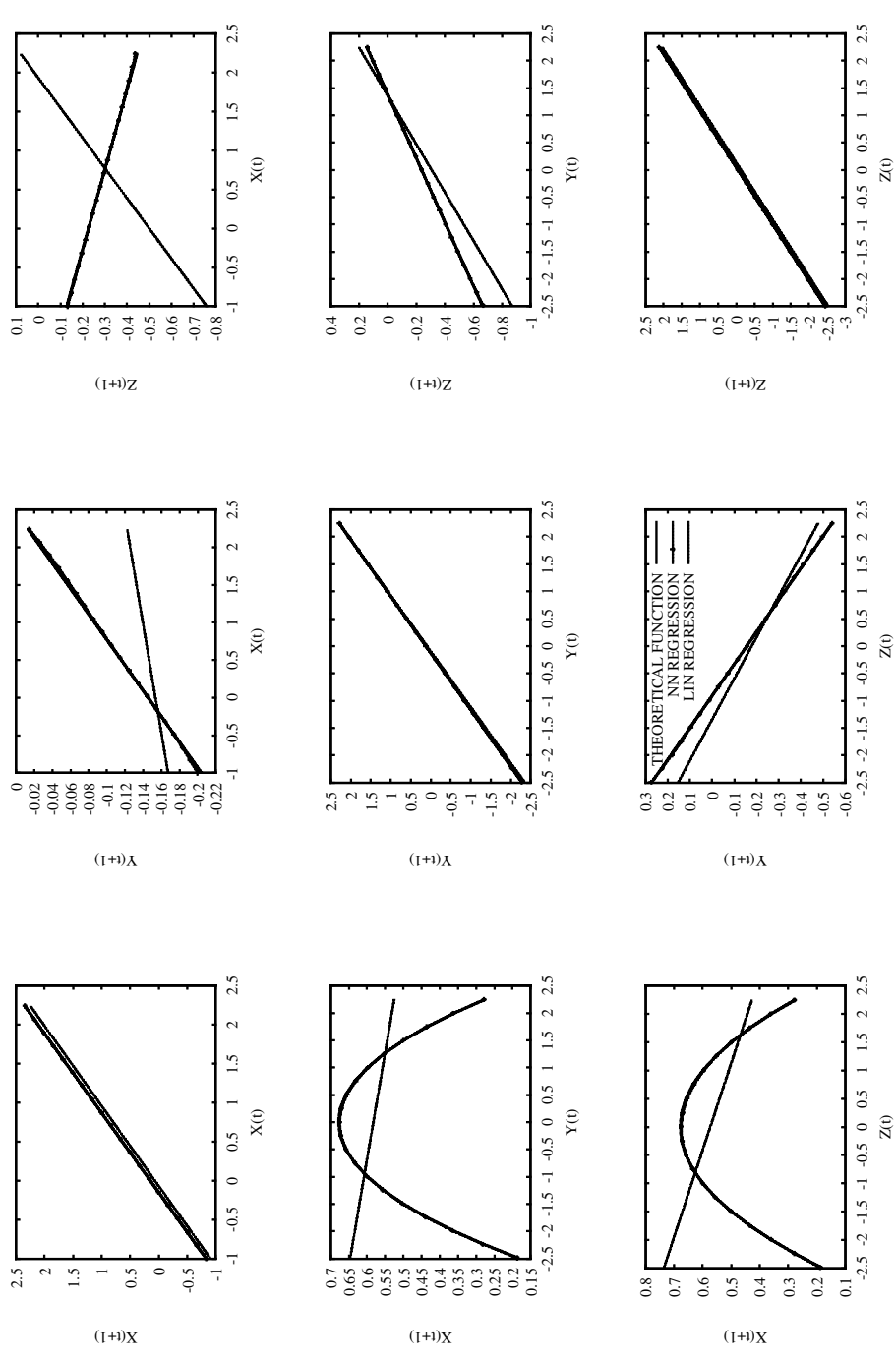



Figure 8. Representation of the theoretical Lorenz dynamical operator (continuous line), its neural-network estimate (dotted with crosses), and its linear-regression estimate (dashed). $X$, $Y$ and $Z$ are the state variables. See text for explanation.

used to understand the physical processes. At least, the sensitivities of a model can be compared with sensitivities inferred from observations.

We have tested this idea by making prediction runs with our NN representation of the Lorenz model: the calculation proceeds by calculating the state of the system at time step $t + \Delta t$, from the state and sensitivities of the system at time $t$; the sensitivities are then calculated at time $t + \Delta t$, and used in the next cycle. Figure 9 shows the evolution of the r.m.s. error of the predictions based on the linear and our nonlinear statistical models against the actual model started at the same state. As in section 4(b), each time step is 0.08 units, about 2 h in the scaling of the equations (i.e. 0.08 of the one day unit of the original Lorenz model). As expected, the nonlinear regression by the NN does much better than the linear regression, but the fact that the Lorenz system is chaotic (with the particular parameter values used) results in a relatively rapid increase of prediction error, even with a more accurate approximation of the system dynamics. Nonetheless, the nonlinear analysis extends the period of useful prediction accuracy by at least a factor of three. Figure 10 illustrates the time records from the prediction model and the actual model.

### (*c*) *Analysis of sensitivities*

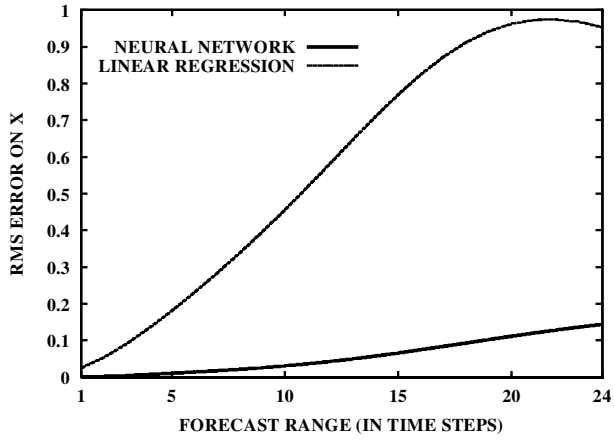
We illustrate the retrieval of the variable sensitivities in the form of histograms of their values as a function of $X(t)$. Similar figures (not shown) are obtained as functions of $Y(t)$ or $Z(t)$. The standard deviation of the theoretical sensitivities of the system (i.e. the linear sensitivities that come from the Jacobian of the system equations) are shown in Fig. 11, indicating that the all sensitivities of the system, except for $\partial X(t + \Delta t)/\partial X(t)$, are not constant. In this figure, the $y$-axis represents the standard deviation of the sensitivities and these quantities are represented for different states of the system (the sensitivities being dependent on the system) characterized by different states of the $X(t)$ variable. Since the system is not uniformly frequently in each $X(t)$ state, the number of points used to compute the standard deviation of the sensitivity is different for each $X(t)$ range. But our standard-deviation estimates are robust since the number of data values used is quite large (200 000 points). Note that the $y$-axes are all different since the standard-deviation ranges are different for the different sensitivities.

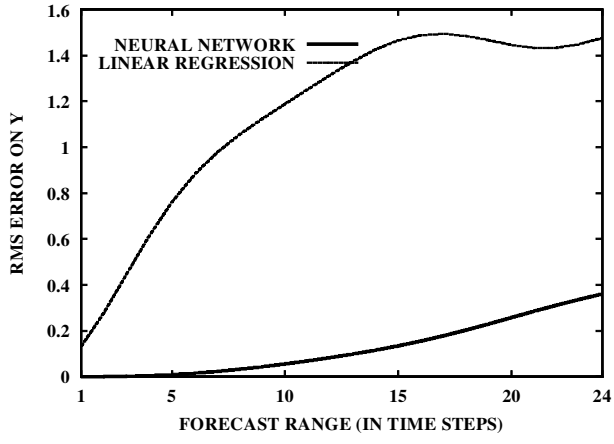
The classical approach for the estimation of sensitivities takes the finite differences of two variables between two (usually equilibrium) states of the system or two extreme events. For example, for the estimation of $\Delta X/\Delta Y$, two sets of extreme events of the variable $Y$ could be selected in the observations and the averages of the state differences $\langle \Delta X \rangle$ and $\langle \Delta Y \rangle$ estimated. Then, the following approximation would be used:

$$\frac{\partial Y}{\partial X}(t) \simeq \frac{\langle \Delta X \rangle}{\langle \Delta Y \rangle}. \tag{54}$$

We see how this approach can go wrong because it is so dependent on the selection of data: at best, it gives a crude estimate of the mean sensitivity for the selected dataset of extremes, which usually becomes worse as the time interval grows larger. The results of this approach for the Lorenz model would be very poor.

The particular sensitivity $\partial X(t + \Delta t)/\partial X(t)$ is the only one that is constant, i.e. does not depend on the state of the system: in Eq. (37), $\partial X(t + \Delta t)/\partial X(t) = 1 - a\Delta t$ (the values in Fig. 11 are not perfectly equal to zero due to numerical imprecision). The linear regression, for this particular sensitivity, is then a good estimate. So the results are good in this particular case, but for the eight other sensitivities, the results of the linear regression are insufficient. In a real-world case, we would not know which results are correct, if any.



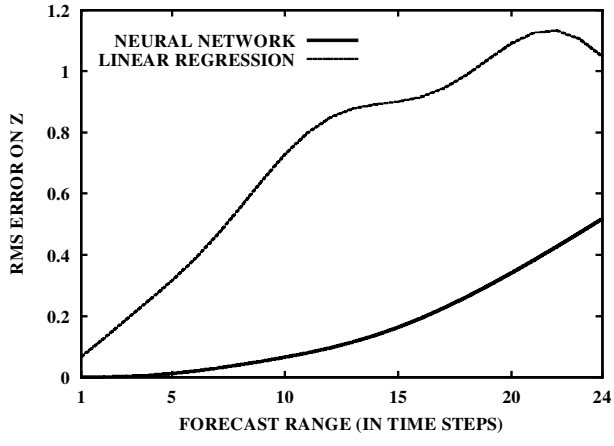




Figure 9. Prediction root-mean-square error of the state variables for the neural-network regression (continuous line) and the linear regression (dashed) as a function of the forecast range (in time steps $\Delta t = 0.08$).

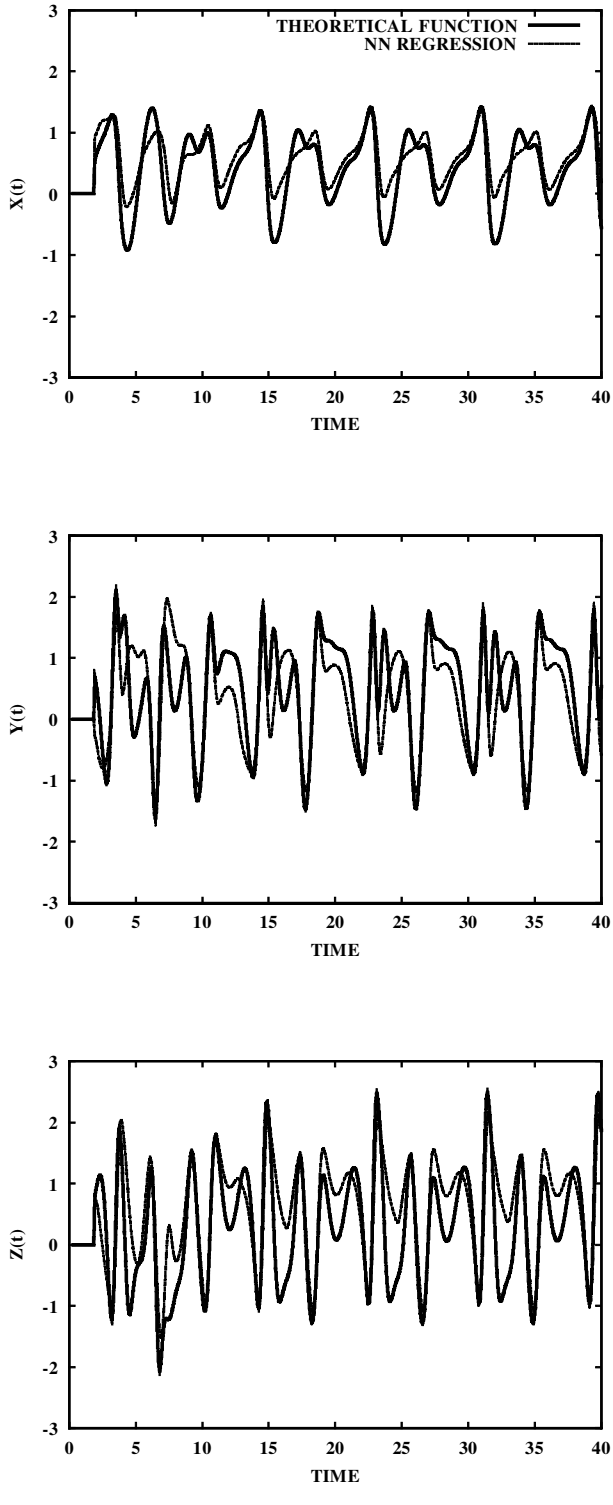



Figure 10. An example of prediction of $X$, $Y$ and $Z$ state variables with a forecast range of 24 time steps: theoretical function and neural-network regression.

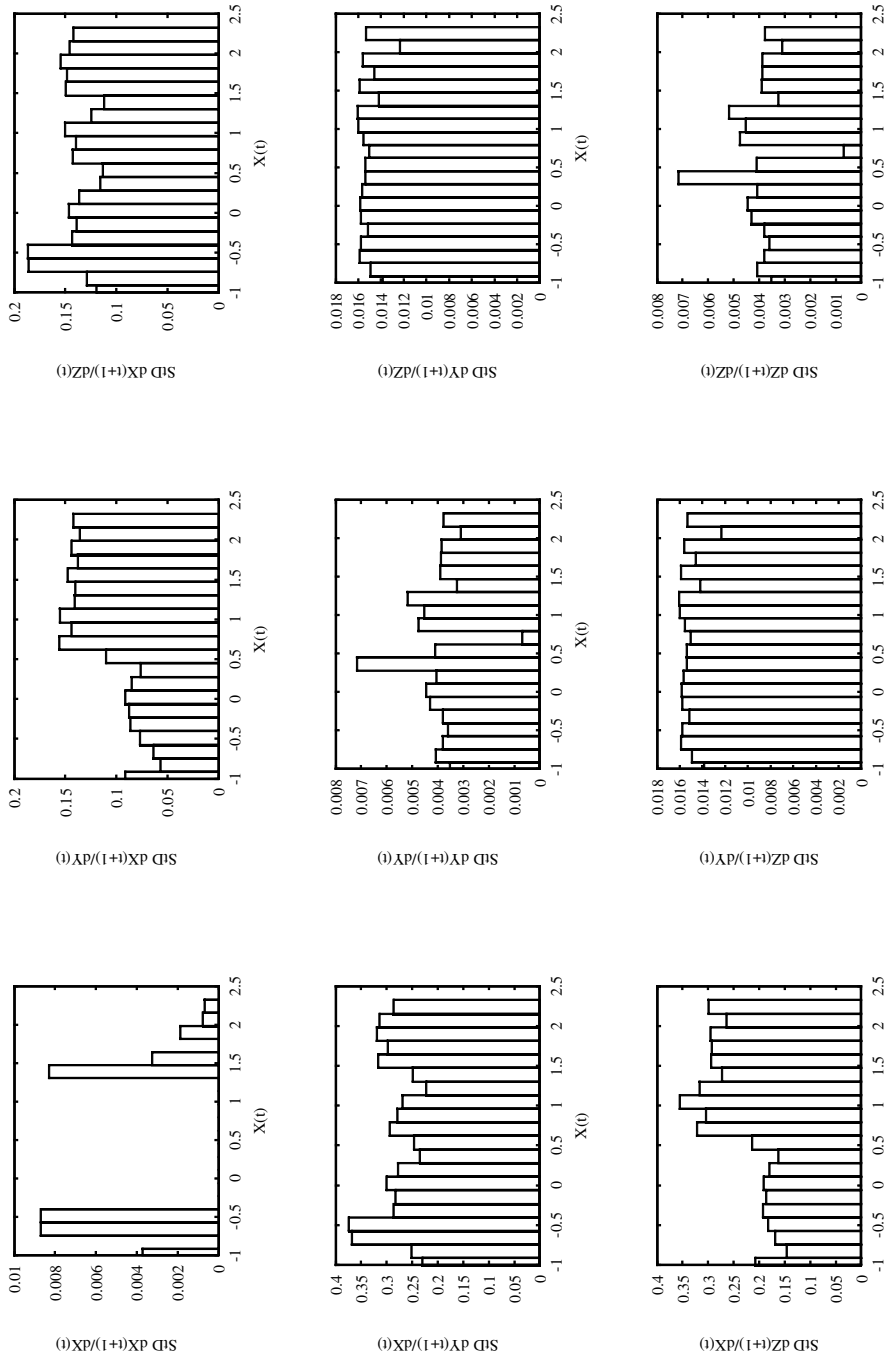



Figure 11. Standard deviation of the sensitivities of the discrete Lorenz dynamical system. $X$, $Y$ and $Z$ are the state variables.

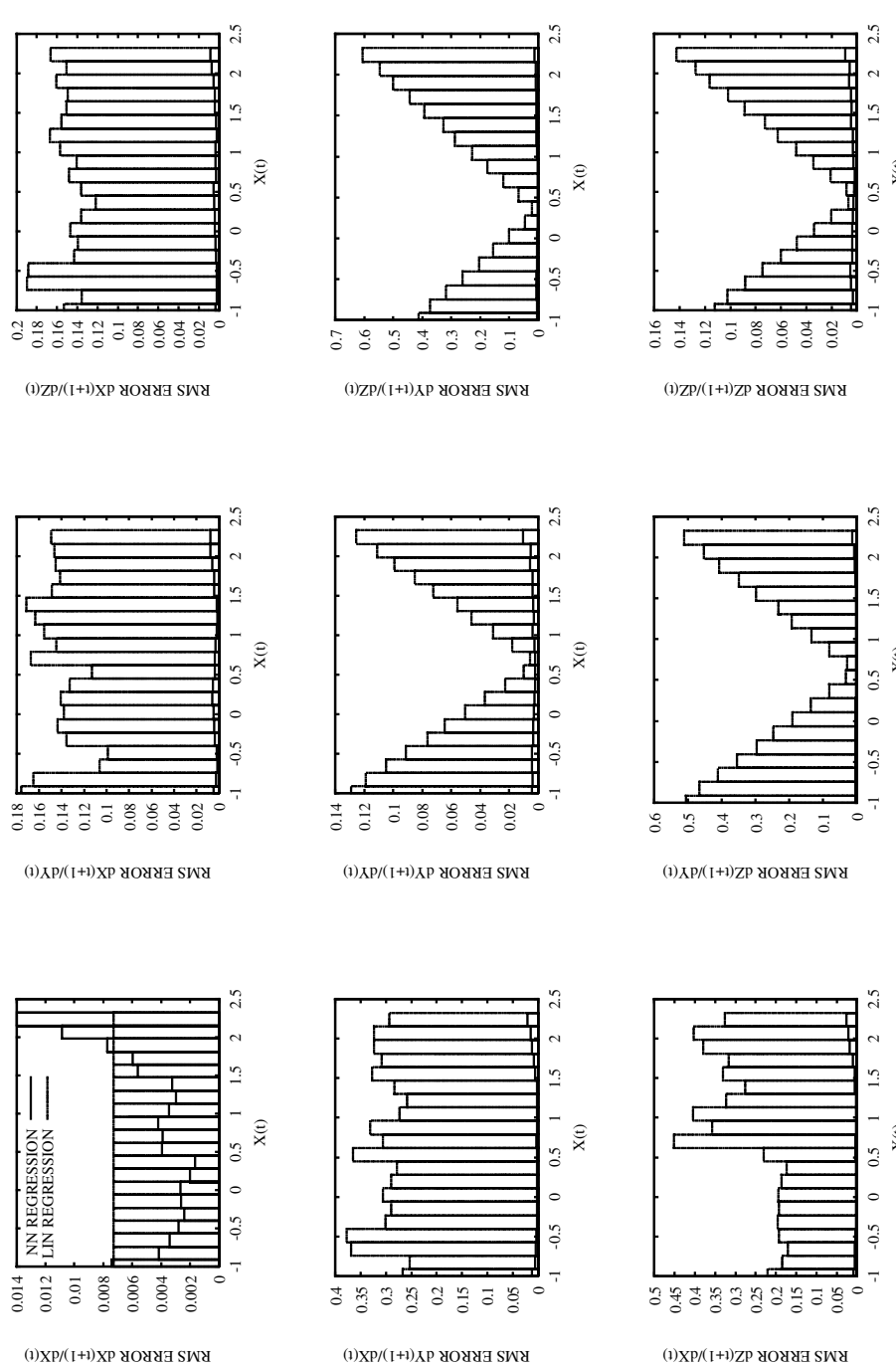



Figure 12. Root-mean-square error for the sensitivities estimates with respect to the 'theoretical' sensitivities: neural-network-based estimates (continuous lines) and linear-regression-based estimates (dashed). $X$, $Y$ and $Z$ are the state variables.

TABLE 1. STATISTICS ON TRUE AND RETRIEVED SENSITIVITIES

| Sensitivity | Statistics | Theoretical | Linear | Neural Network |
|---|---|---|---|---|
| $\partial X(t+\Delta t)/\partial X(t)$ | Mean | 0.980 | 0.973 | 0.981 |
| | Standard deviation | 0.000 | 0.000 | 0.003 |
| | R.m.s. error | | 0.007 | 0.003 |
| $\partial X(t+\Delta t)/\partial Y(t)$ | Mean | −0.077 | −0.025 | −0.076 |
| | Standard deviation | 0.133 | 0.000 | 0.132 |
| | R.m.s. error | | 0.144 | 0.004 |
| $\partial X(t+\Delta t)/\partial Z(t)$ | Mean | −0.057 | −0.064 | −0.057 |
| | Standard deviation | 0.146 | 0.000 | 0.145 |
| | R.m.s. error | | 0.147 | 0.004 |
| $\partial Y(t+\Delta t)/\partial X(t)$ | Mean | −0.077 | 0.014 | −0.077 |
| | Standard deviation | 0.297 | 0.000 | 0.297 |
| | R.m.s. error | | 0.310 | 0.003 |
| $\partial Y(t+\Delta t)/\partial Y(t)$ | Mean | 0.955 | 0.979 | 0.956 |
| | Standard deviation | 0.048 | 0.000 | 0.048 |
| | R.m.s. error | | 0.054 | 0.003 |
| $\partial Y(t+\Delta t)/\partial Z(t)$ | Mean | −0.141 | −0.133 | −0.141 |
| | Standard deviation | 0.192 | 0.000 | 0.193 |
| | R.m.s. error | | 0.193 | 0.003 |
| $\partial Z(t+\Delta t)/\partial X(t)$ | Mean | 0.184 | 0.259 | 0.184 |
| | Standard deviation | 0.281 | 0.000 | 0.281 |
| | R.m.s. error | | 0.291 | 0.004 |
| $\partial Z(t+\Delta t)/\partial Y(t)$ | Mean | 0.141 | 0.226 | 0.142 |
| | Standard deviation | 0.192 | 0.000 | 0.192 |
| | R.m.s. error | | 0.210 | 0.003 |
| $\partial Z(t+\Delta t)/\partial Z(t)$ | Mean | 0.955 | 0.962 | 0.955 |
| | Standard deviation | 0.048 | 0.000 | 0.048 |
| | R.m.s. error | | 0.049 | 0.003 |

Figure 12 represents the r.m.s. error in sensitivity with respect to the theoretical sensitivities for the neural network and linear estimates. The NN-based estimates of the sensitivities are a considerable improvement in comparison to the linear-regression-based ones (except for the constant sensitivity $\partial X(t+\Delta t)/\partial X(t)$ at extreme values of $X(t)$, but the differences are still negligible). Note that the magnitudes of the sensitivities are very different, yet our technique seems to be able to handle this situation. Furthermore, these results are good if we compare the r.m.s. errors with the natural variability described in Fig. 11. These results are summarized in Table 1. Since linear-regression-based sensitivities are constant by assumption, the r.m.s. errors of this representation are essentially equal to the standard deviations of the sensitivities. The improvement of the NN-based sensitivities is considerable with respect to the linear regression: standard-deviation errors are always (except for the constant sensitivity) smaller than the natural standard deviation of the theoretical sensitivities by one and sometimes two orders of magnitude. Given the large range of the sensitivity magnitudes, it is notable that the r.m.s. errors of the NN are uniformly distributed over the nine sensitivities, even if the variability of the sensitivities is quite different. Table 1 summarizes the improvement gained by use of the NN Jacobians to estimate the instantaneous, multivariate and non-linear sensitivities of the discrete Lorenz dynamical system.

Figure 13 shows an example of the evolution in time of the theoretical and NN estimates of sensitivities. This figure also highlights the more complex situation when feedback processes intersect: when the state of the system reaches some extreme value, the sensitivities change, even in their sign, taking the system back towards a middle range of values and finally stabilize the system on its attractor.

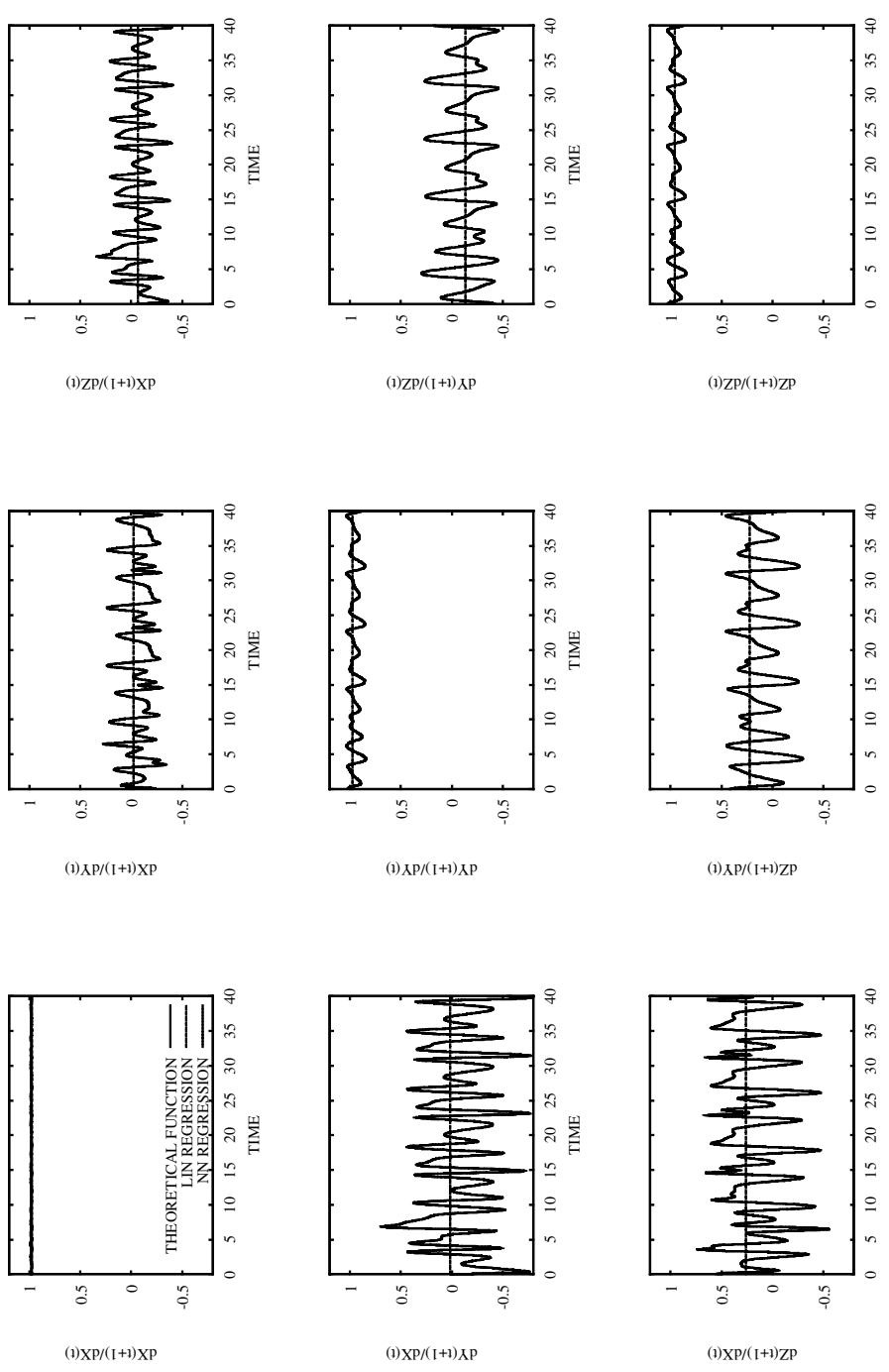
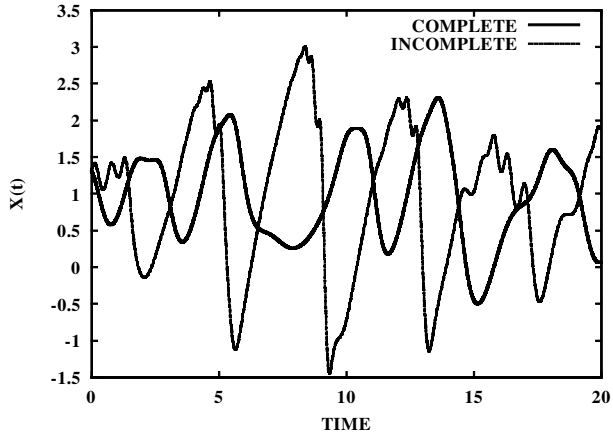
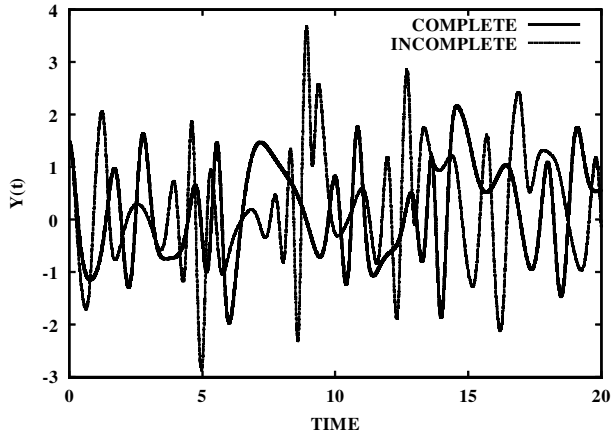



Figure 13. Jacobians evolution through time: theoretical Jacobians (continuous line), linear-regression-based estimates (dotted), and neural-network-based estimates (dashed). $X$, $Y$ and $Z$ are the state variables. The differences between theory and the neural network are not distinguishable.



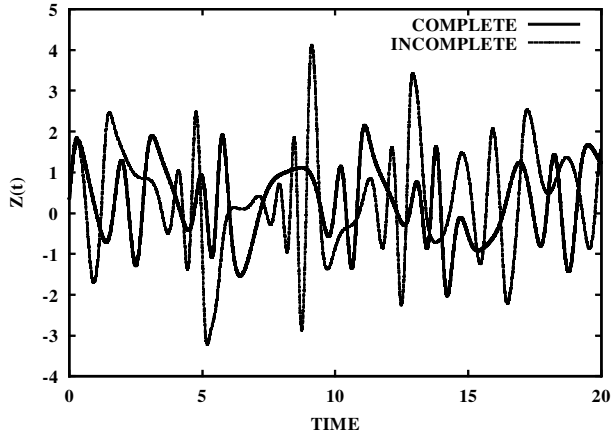




Figure 14. Discrete Lorenz model (continuous line) and discrete Lorenz model minus the sensitivity $\partial X(t + \Delta t)/\partial Z(t)$ (dashed), where $X$, $Y$ and $Z$ are the state variables.

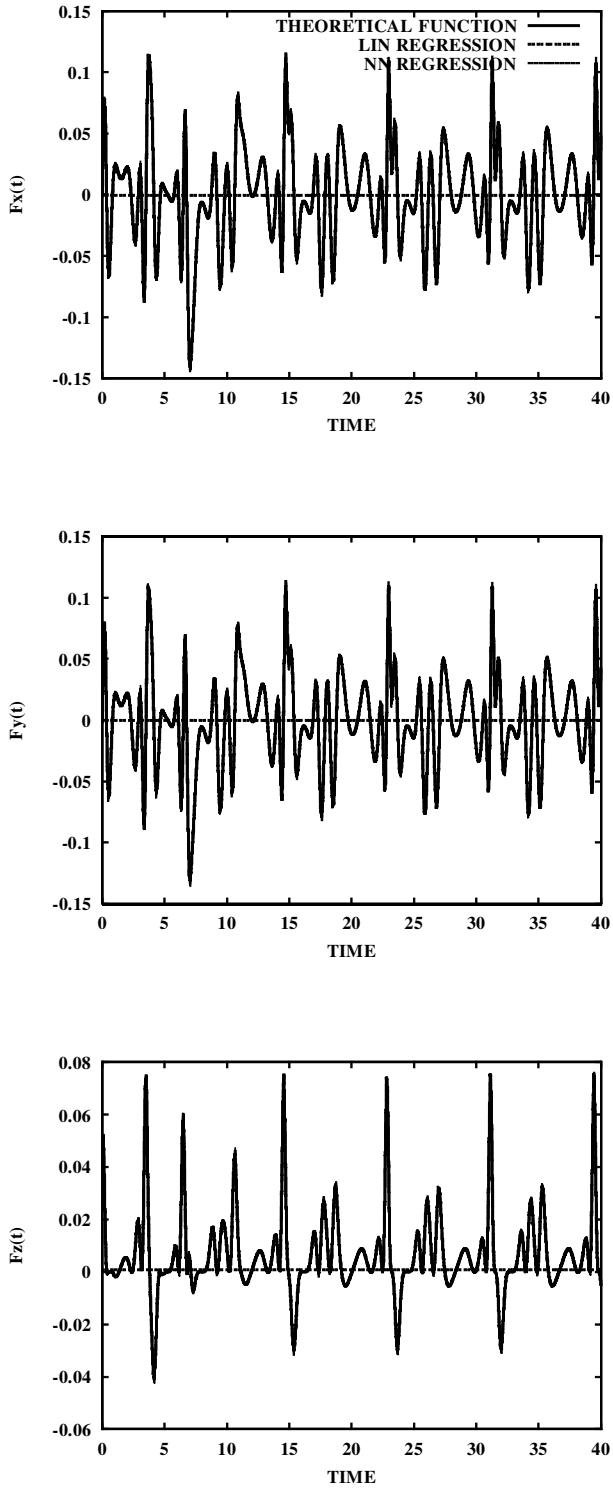



Figure 15. Feedback factors $F_x$, $F_y$ and $F_z$ (see Eqs. (46)–(48)): evolution through time using theoretical, linear-regression and neural-network sensitivities—the theoretical and neural-network feedback factors cannot be distinguished.

For example, using the theoretical sensitivities in Eq. (37), we can analyse the relation between the variables $X$ and $Y$. If $Y$ is large and positive, then the sensitivity $\partial X(t + \Delta t)/\partial Y(t) = -2\Delta t Y(t)$ becomes large and negative. So, if $Y$ continues to increase, the variable $X$ will decrease even more rapidly. But the auto-sensitivity $\partial Y(t + \Delta t)/\partial Y(t)$ (the most important sensitivity for the variable $Y$) is equal to $1 - \Delta t + \Delta t X(t)$, which will be less than 1 (damping effect) when $X$ is less than 1.

One consequence of this behaviour is that particular sensitivities, even when they are small on average, can still have a strong impact on the behaviour of the system. A linear regression analysis assuming that the sensitivities are constant in time, may provide some estimate of mean sensitivities from a dataset. For example, the sensitivity $\partial X(t + \Delta t)/\partial Z(t)$ is, on average, nearly zero. A linear analysis, in this case, might suggest neglecting this relationship in understanding the system. Figure 14 shows how wrong this approximation would be: this figure represents the discrete Lorenz model defined in Eq. (35) with and without this particular sensitivity set to zero. The two trajectories have quite distinct behaviour: the simulation without the variable sensitivity oscillates more strongly and with a different time-scale. The behaviour of the complete system is produced by oscillations of the particular sensitivity, depending on the state of the system, between a positive and a negative value, thereby stabilizing the system dynamics.

The sensitivities have a general tendency to exhibit similar shapes in their time records (Fig. 13), which means that they are closely linked with each other (i.e. for example, when one is high another is low). This type of nonlinear behaviour prevents a linear, even multivariate, regression analysis from extracting even approximate information about the system dynamics. Understanding of the system seems to require a more accurate representation of the time evolution of the multivariate sensitivities.

### (*d*) *Feedback analysis*

We have seen that the classical approach for the feedback analysis, which makes strong (and incorrect) hypotheses about the dynamical system, is not well adapted to the Lorenz model. However, the feedback factors can still be computed for the theoretical function, the linear regression model and the NN model according to Eq. (46)–(48). We suppose here that these expressions are applicable to show that these feedback factors evolve in time (Fig. 15), in violation of one of the assumptions used to obtain the expressions. The feedback factors Eqs. (46)–(48) are not simple and do not improve our understanding of the system since their physical interpretation is confused because they are products of the (simpler) sensitivities. The sensitivities themselves seem to be the more fundamental quantities. Furthermore, as we showed in section 2, without all the assumptions of this formalism (linearity, constant sensitivities, hierarchical cause-and-effect relationships, constant forcing, static equilibrium state, etc.), the whole formulation in terms of feedback factors falls apart.

## 6. Concluding remarks

What we have learned with this study of the Lorenz model is that the feedback processes are dependent on some important particular properties of the dynamical system under study. First, the feedback processes appear in a dynamical system when multivariate sensitivities are integrated over time. Second, if the system is nonlinear (i.e. the dynamical operator in Eq. (1) is nonlinear), the sensitivities are state-dependent and therefore not constant with time, which means that the feedback processes evolve

in time. Third, each feedback has a strong impact on the character and behaviour of the dynamical system, even those that may have a small time-averaged magnitude can have a stabilizing effect that changes drastically the characteristics of the system. Without such feedbacks the dynamical system would have more tendency to destabilize when an external forcing is introduced. The feedback processes can have a stabilization effect, so the system does not diverge too much from its initial equilibrium. But this new (statistical) equilibrium state could be different, with for example a higher frequency of extreme events. This is a theory that has been discussed recently by Palmer (1999).

We have shown that the classical technique (from the electrical circuit theory) to analyse feedbacks is, by its hypotheses, very limited in its validity when applied to highly nonlinear, multivariate systems like the climate. Furthermore, the results of this kind of classical analysis are no more than a 'schematic' measure of feedback processes at system equilibrium, which may be very misleading.

In comparison, the multivariate, instantaneous and nonlinear sensitivity concept is more generally applicable without these constraints and appears to be a good way of understanding the behaviour of a system with coupled feedback processes. This general technique allows the quantification of these processes both spatially and temporally. This dynamical information seems to be more useful than the classical feedback factor which provides only one number per variable. Furthermore, if a priori information about the cause-and-effect physical relationships is available (like in a reduced-form model), it is possible to introduce this additional knowledge into the NN model.

It is very important to note again that our goal in this work is not to perform statistical prediction of climate. Our analysis of local sensitivities tends to show that the statistical prediction of the climate based on too simple assumptions, in particular the linearity of the response to forcing, is not useful. Our goal is to improve numerical models of climate by understanding the processes better, then use the improved numerical models to do the climate predictions.

The dataset used in our analysis technique needs to satisfy some statistical requirements. First, the space and time sampling needs to be adequate to the description of the space and time variability of the sensitivities that originate the feedbacks, so that the assumption that the sensitivities are constant over one time step or one space interval is an accurate approximation. Using too coarse time sampling is equivalent to using time-averaged data, which mixes many physical processes and ruins the sensitivity estimates. Using space-averaged data is also dangerous; for example, a mean sensitivity equal to zero could be generated by two opposite regimes with non-zero sensitivity. In other words, even if we are studying the longer-term behaviour of the system, we must resolve the dynamics appropriately or the nonlinear integration will be incorrect. Some averaging is unavoidable, however, as processes occur on a range of spatial and temporal scales. The analysis ignores, due to limited data/model resolution, smaller scales making the assumption that they are not important in the feedback processes considered. A study of the space/time (state) variability of the sensitivities is then a prerequisite for the definition of the dataset sampling needed for the feedback analysis. Second, the dataset has to have good space and time coverage in order to provide many samples of as many climatological situations as possible. In other words, the dataset should contain 'all possible' combinations of the state variables. The more situations in the dataset, the better will be the 'laws' inferred by the analysis. These comments also mean that the dynamics of the system cannot be correctly deduced from datasets where individual quantities have been separately averaged over space and time because such treatment would alter the dynamical relationships. These points are a major argument to use detailed (e.g. hourly or daily), long-term datasets instead of generating new ones,

limited in time. Fortunately, the size of observational and model data available in climate studies is now becoming large enough to support such an NN training.

Some additional technical aspects need further investigation. In particular, the question of how to handle high-dimension data needs to be addressed. There are different possible approaches to reduce the dimension of data in order to apply this technique to climate data from observations or GCMs. First, if we are interested in a short time-scale (i.e. for local processes), it may be possible to perform the analysis on a limited number of locations. If the time-scales considered are larger, then they involve larger-scale spatial information. In that case, a way to reduce the dimension of these large-scale data would be to use an EOF or ICA (Aires *et al.* 2000, 2002a) procedures, it is possible to estimate our sensitivities on the principal components themselves. Another way to deal with large-scale structures is to use indices (like El Niño Southern Oscillation, Quasi-Biennial Oscillation, North Atlantic Oscillation, etc.) as proxies for climate processes.

Our technique has the advantage of being applicable to numerical-model output as well as observations, which means that the important work of inter-comparison of models with each other and models to data could be carried out in a more meaningful way, by comparing the sensitivities of the variables of the system and their state dependence. This diagnostic usage is particularly interesting because it concerns very intuitive and physical quantities. Comparisons of the sensitivity relationships could also be made with field experiment data to understand how physical processes produce these sensitivities. Thus, our analysis approach provides a framework for a whole new attack on these problems.

The statistical model estimating the sensitivities can also be used to study the change of the system to a new equilibrium state, including the time to reach equilibrium after a small perturbation. This simplified model could also be used to analyse the propagation of uncertainties when predictions are performed. In other words, the NN statistical model provides a better approximation of 'small perturbation' behaviour than attempts to linearize the system by dropping relationships or averaging in space and time. The next step for these ideas is to use this new technique for more complicated climate systems involving real observations or numerical model outputs.

## REFERENCES


Aires, F. 1999 'Problème inverses et réseaux de neurones: application à l'interféromètre haute résolution IASI et à l'analyse de séries temporelles'. PhD thesis, Université Paris IX-Dauphine

Aires, F., Schmitt, M., Scott, N. A. and Chédin, A. 1999 The weight smoothing regularization for Jacobian stabilization. *IEEE Trans. Neural Networks*, **10,** 1502–1510

Aires, F., Chédin, A. and Nadal, J.-P. 2000 Independent component analysis of multivariate time series. Application to the tropical SST variability. *J. Geophys. Res.*, **105,** 17437–17455

Aires, F., Prigent, C., Rossow, W. B. and Rothstein, M. 2001 A new neural network approach including first-guess for microwave retrieval of atmospheric water vapor, cloud liquid water path, surface temperature and emissivities over land from SSM/I observations. *J. Geophys. Res.*, **106,** 14887–14907

Aires, F., Rossow, W. B. and Chédin, A. 2002a Rotation of EOFs by the independent component analysis: Towards a solution of the mixing problem in the decomposition of geophysical time series. *J. Atmos Sci.*, **59,** 111–123

Aires, F., Rossow, W. B., Scott, N. and Chédin, A. 2002b Remote sensing from the IASI instrument. 2: Simultaneous retrieval of temperature, water vapor and ozone atmospheric profiles. *J. Geophys. Res.*, **107,** doi: 10.1029/2001JD001591

2002c Remote sensing from the IASI instrument. 1: Compression, de-noising, and first-guess retrieval algorithms. *J. Geophys. Res.*, **107,** doi: 10.1029/2001JD000955

Aires, F., Chédin, A., Scott, N. and Rossow, W. B. 2002d A regularized neural net approach for retrieval of atmospheric and surface temperatures with the IASI Instrument. *J. Appl. Meteorol.*, **41,** 144–159

Andronova, N. G. and Schlesinger, M. E. 1991 The application of cause-and-effect analysis to mathematical models of geophysical phenomena. 1: Formulation and sensitivity analysis. *J. Geophys. Res.*, **96,** 941–946

1992 The application of cause-and-effect analysis to mathematical models of geophysical phenomena. 2: Stability analysis. *J. Geophys. Res.*, **97,** 5911–5919

Bode, H. W. 1945 *Network analysis and feedback amplifier design*. Van Nostrand, New York

Curry, J. A. and Webster, P. J. 1999 *Thermodynamics of atmospheres and oceans*. Academic Press, London

Cybenko, G. 1989 Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, **2,** 303–314

Hansen, J., Lacis, A., Rind, D., Russel, G., Stone, P., Fung, I., Ruedy, R. and Lerner, J. 1984 'Climate sensitivity: Analysis for feedback mechanisms'. Pp. 120–163 in *Climate processes and climate sensitivity*. (Maurice Ewing Series, No. 5). American Geophysical Union, Washington DC

Hornik, K., Stinchcombe, M. and White, H. 1989 Multilayer feedforward networks are universal approximators. *Neural Networks*, **2,** 359–366

Lorenz, E. N. 1984 Irregularity: A fundamental property of the atmosphere. *Tellus*, **36A,** 98–110

1990 Can chaos and intransitivity lead to interannual variability? *Tellus*, **42A,** 378–389

Palmer, T. N. 1999 A nonlinear dynamical perspective on climate prediction. *J. Climate*, **12,** 575–591

Peixoto, J. and Oort, A. H. 1992 *Physics of Climate*. American Institute of Physics, New York

Rumelhart, D., Hinton, G. and Williams, R. 1986 'Learning representations by back-propagating error'. Pp. 318–362 in *Parallel distributed processing*, Vol. I. MIT Press

Schlesinger, M. E. 1985 'Feedback analysis of results from energy balance and radiative–convective models'. Pp. 280–318 in *The potential climatic effects of increasing carbon dioxide*. Eds. M. C. McCracken and F. M. Luther. DOE/ER-0237, US Department of Energy, DOE/ER-0237

Slingo, A., Pamment, J. A., Allan, R. P. and Wilson, P. S. 2000 Water vapour feedbacks in the ECMWF re-analyses and Hadley Centre climate model. *J. Climate*, **13,** 3080–3098

Smith, L. 1997 Uncertainty dynamics and predictability in chaotic systems. *Q. J. R. Meteorol. Soc.*, **123,** 1–34

Yuval 2000 Neural network training for prediction of climatological time series, regularized by minimization of the Generalized Cross Validation Function. *Mon. Weather Rev.*, **128,** 1456–1473