ARTICLE OPEN

# Data driven models of the Madden-Julian Oscillation: understanding its evolution and ENSO modulation

Nicolás Díaz[1], Marcelo Barreiro[1] and Nicolás Rubido[2]✉

Understanding the physical mechanisms of the Madden-Julian Oscillation (MJO) and its evolution is a major concern within the climate community. Its main importance relies on its ability to act as a source of predictability within the intra-seasonal time-scale in tropical and extratropical regions, therefore filling the gap between weather and climate forecasts. However, most atmospheric general circulation models fail to correctly represent MJO's evolution, and their prediction skills are still far from MJO's theoretical predictability. In this work we infer low dimensional models of the MJO from data by applying a recently developed machine learning technique, the Sparse Identification of Non-linear Dynamics (SINDy). We use the daily-mean outgoing longwave radiation MJO index (OMI) as input data to infer bi-dimensional climatological models of the MJO, and analyse the inferred models during El Niño and La Niña years. This approach allows us to diagnose the MJO's behaviour in OMI's phase space. Our results show that MJO can be most frequently represented by a harmonic oscillator, which represents the MJO's eastward propagation and characteristic period. Upon this basic oscillatory behaviour, we find that small non-linear corrections play a fundamental role in representing MJO's non-uniform speed of propagation, explaining its acceleration over the Pacific Ocean region. Particularly, we find that MJO's evolution is most frequently non-linear [linear] during El Niño [La Niña] years. Overall, our work shows that SINDy can robustly model MJO's evolution as a linear oscillator with small non-linear corrections, contributing to understand the MJO's dynamics and dependency on El Niño-Southern Oscillation.

*npj Climate and Atmospheric Science* (2023)6:203 ; https://doi.org/10.1038/s41612-023-00527-8

## INTRODUCTION

The Madden-Julian Oscillation (MJO) is an atmospheric planetary-scale phenomenon that exists over the tropical region[1] but has a global impact. It influences remote regions of the planet by means of a tropical-tropical teleconnection—related to equatorially traped Kelvin and Rossby waves—and a tropical-extratropical teleconnection due to forcing of planetary Rossby waves. For example, its effects include changes in precipitation[2–4] and temperature[5,6], setting up conditions for marine heat-waves[7,8], modulating cyclo-genesis[9,10], and interacting with monsoon circulations systems[11,12], to name a few. By extension, the MJO is a key source of predictability within the intra-seasonal (IS) time-scale, which is a challenging scale to predict atmospheric behaviour since it falls in between the synoptic ($\lesssim$10 days) and longer climate time-scales ($\gtrsim$90 days)[13,14].

The MJO is composed of two anomalous convective centres: a wet centre, where anomalous precipitations occur, and a dry centre, where precipitations are inhibited. These two centres are coupled through large scale zonal baroclinic cells, with winds converging [diverging] over the wet [dry] centre at lower levels and diverging [converging] aloft[1]. This whole structure, which usually initiates over the warm waters of the Indian Ocean, slowly propagates eastward (particularly observed during the austral summer) and presents a characteristic periodicity within the IS time-scale. However, many aspects of the MJO need more research or show large variability – such as its eastward speed, which has been reported with values between 5.0 m s$^{-1}$ and 10.0 m s$^{-1}$ [15,16]. In particular, there are open questions regarding its interaction with the Maritime Continent[17] and its acceleration beyond the Maritime Continent, as well as the impact on MJO's variability that El Niño Southern Oscillation (ENSO)[18–21] or the quasi-biennial oscillation[22] can have. Consequently, advances in understanding MJO's dynamics and and its physical mechanisms are needed. A comprehensive summary and comparison of MJO's main current theories can be found in ref. [23].

Here, we infer minimal models of the MJO from observed data and the use of machine learning techniques. This allows us to diagnose MJO's behaviour according to the inferred models, revealing its main kinematic characteristics, the relevance of non-linear dynamics, and MJO's dependency on ENSO.

Our data-driven modelling approach is based on applying the Sparse Identification of Non-linear Dynamics (SINDy) method[24] to a random sample of the daily-mean Outgoing longwave radiation MJO Index (OMI)[25] for the 1979 to 2021 December to March (DJFM) months–or a selection of years according to La Niña or El Niño years (see Fig. 1 for a schematic representation of our methodology). The resultant models describe the MJO's behaviour from bi-dimensional equations of motion (the OMI index is composed of two principal component time-series), which we find have oscillatory characteristics with periods within the expected range and a dependence to the ENSO. Other works have modelled the MJO by constructing minimal physical models[26–28] or applied machine learning algorithms to obtain reliable forecasts within the IS time-scale[29,30]. We take both of these aspects into our approach in order to diagnose MJO's behaviour from the inferred minimal models. Hence, our results contribute to understanding the MJO's physical behaviours and could contribute to operational forecasting.

## RESULTS

Our results come from OMI daily data from 1979 to 2021 (42 years) for the DJFM months. We infer models for the MJO's behaviour

[1]Universidad de la República, Instituto de Física de Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay. [2]University of Aberdeen, King's College, Institute for Complex Systems and Mathematical Biology, AB24 3UE Aberdeenx, UK. ✉email: nicolas.rubidoobrer@abdn.ac.uk

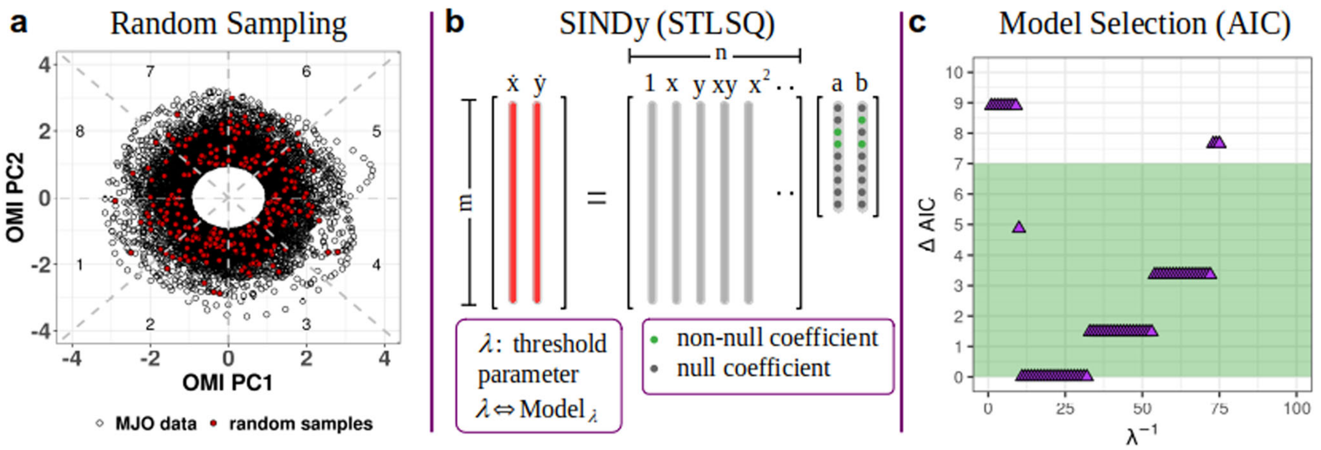


**Fig. 1 Schematic representation of the methodology.** **a** OMI data (unfilled black circles) from the 1979 to 2021 December to March months, where $m = 2^8$ points are randomly selected (filled red circles). **b** Sparse Identification of Nonlinear Dynamics (SINDy) method applied to $m$ randomly sampled OMI data ($x$ and $y$ components), which are used to construct the matrix with $n = 18$ polynomial predictors of the form $x^i y^j$ (with $i, j = 0, 1, 2$) and the velocity field $(\dot{x}, \dot{y})$. SINDy solves an optimisation problem that promotes sparsity in the unknown coefficients by using a threshold parameter $\lambda$ in a sequentially thresholded least-squares algorithm (STLSQ). **c** We obtain 100 models for the $m$ sampled data by tuning $\lambda^{-1}$ from 0 to 100 (with a step of 1) and then discard those with an Akaike Information Criterion (AIC) greater than 7 (values outside the shaded green area).

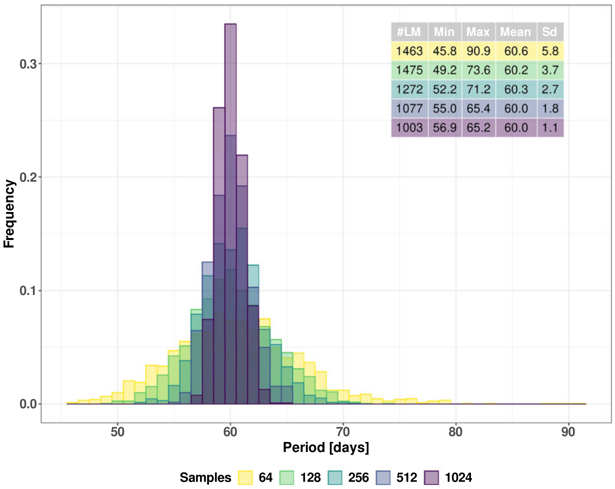
from these OMI data by following the methodology represented in Fig. 1. We start by taking $m$ random samples without replacement (panel a), then we fit these samples to a polynomial set of dynamical systems using SINDy (panel b), which we lastly filter according to the Akaike Information Criterion (AIC) to obtain the minimal models (panel c)—see SINDy's details in 'Implementing SINDy on MJO' and AIC's details in 'Model Selection Through Akaike Information Criterion'. We name the resultant data-driven models as the climatological models. Moreover, in order to have results with sufficient statistical power, we repeat this process for $N = 1000$ independent realisation of the randomly sampled data resulting in more than $N$ (possibly different) climatological models. For example, we obtain 1902 climatological models when considering $m = 2^8$ samples per realisation.

In 'Climatological data-driven models of the MJO', we group the 1902 models according to their structure, analyse the trajectories generated by the largest groups of climatological models, orbit's characteristic periods and speed, and model's linear and non-linear components. We also study the number and properties of the models inferred with different $m$ samples. In 'Analysis of models according to ENSO's modulation', we analyse the effects of ENSO on the MJO by restricting the data to the corresponding 14 El Niño years and 15 La Niña years that had been identified according to the NOAA's Oceanic Niño Index. Similarly to 'Climatological data-driven models of the MJO', we analyse the resultant ENSO modulated MJO models.

## Climatological data-driven models of the MJO

Here, we present the main characteristics of the subset of climatological models that are linear when considering $m = 2^6, 2^7, 2^8, 2^9$ or $2^{10}$ samples. We find that irrespective of $m$, the largest fraction of climatological models are linear oscillators with a definite period (shown bellow). Therefore, we can determine analytically the period for each oscillatory model (from their trace and determinant[31]), finding that it falls within the IS time-scale—as observed for the MJO. The distribution of periods according to $m$ are shown in Fig. 2.

We note that, regardless of the number of samples $m$, the distributions of periods in Fig. 2 overlap around a mean value of nearly 60 days (close to the median value), becoming narrower for increasing $m$. This can be quantified by the standard deviation ($sd$) or the distribution range. For example, $sd = 5.8$ days if $m = 2^6$ samples are used, progressively decreasing up to $sd = 1.1$ days when $m = 2^{10}$ samples are used. Similarly, the range of the distribution of inferred periods goes from nearly 45 days (91–46) when $m = 2^6$ to 8 days (65–57) when $m = 2^{10}$. These results show that the inferred linear climatological models become more accurate in defining the period with increasing number of samples; but as $m$ grows, we get diminishing returns. Consequently, in what follows we restrict our analysis to $m = 2^8$ random samples per realisation to keep some variability.



| #LM | Min | Max | Mean | Sd |
|---|---|---|---|---|
| 1463 | 45.8 | 90.9 | 60.6 | 5.8 |
| 1475 | 49.2 | 73.6 | 60.2 | 3.7 |
| 1272 | 52.2 | 71.2 | 60.3 | 2.7 |
| 1077 | 55.0 | 65.4 | 60.0 | 1.8 |
| 1003 | 56.9 | 65.2 | 60.0 | 1.1 |

**Fig. 2 Distribution of periods I.** Distribution of periods for the MJO index obtained from sets of linear oscillatory models. Colour coded distributions correspond to taking different number of random samples ($m$) from all the available data to create the models. The top-right table shows a summary of the statistics for each distribution where #LM is the number of linear models, Min (Max) is the minimum (maximum) period, and Mean and Sd are the mean and standard deviation values).

Now, we analyse the set of 1902 climatological models inferred from the 1000 realisations of $m = 2^8$ random data samples. The statistics of the coefficients appearing in the models' equations of motion are summarised in Fig. 3.

On the top panel (a) of Fig. 3, we note that there are two coefficients that occur with frequency = 1, meaning that these

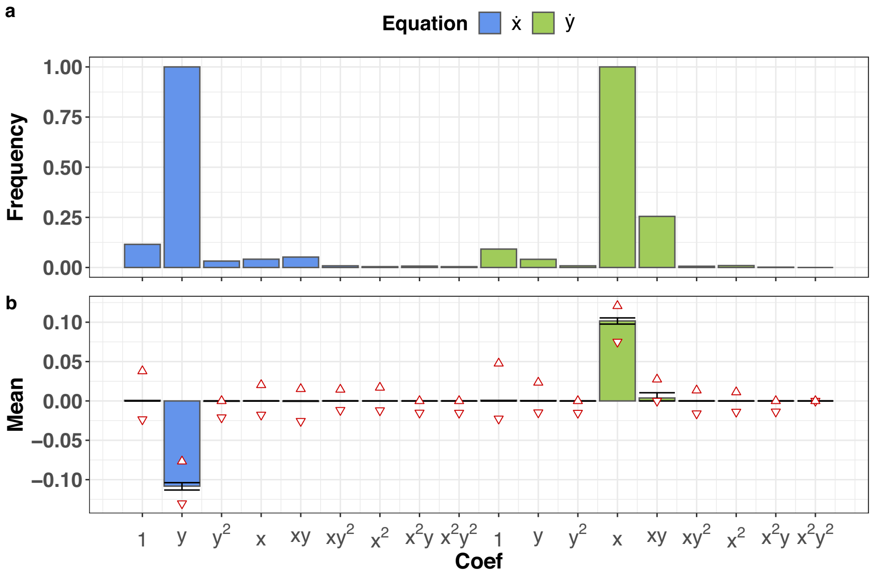


**Fig. 3 Statistics of coefficients.** **a** Frequency of coefficients found over the total set of selected models (linear and non-linear) for $\dot{x}$ (blue colour) and $\dot{y}$ (green colour). The *x*-axis shows the polynomial terms over which we perform SINDy's linear regression. **b** Statistics of the coefficients found over the total set of selected models for each polynomial term. The blue and green colour correspond to the mean coefficient values for the $\dot{x}$ and $\dot{y}$ equations, respectively, the black error bars to the interquartile range, and the red triangles to the minimum and maximum coefficients values.

models have a non-null coefficient after the hard thresholding procedure of SINDy is applied. These coefficients correspond to the *y* term in the equation for $\dot{x}$ —coefficient $a_{01}$—and the *x* term in the equation for $\dot{y}$—coefficient $b_{10}$. As a result, all the inferred climatological models share a common structure given by

$$\begin{cases} \dot{x} \ = \ a_{01}\, y \\ \dot{y} \ = \ b_{10}\, x. \end{cases} \tag{1}$$

The mean value (and interquartile range) of all the possible coefficients in the equations of motion appear in the bottom panel (b) of Fig. 3, showing that the coefficients in eq. (1) have definite signs ($a_{01} < 0$ and $b_{01} > 0$) with nearly the same mean absolute value (~0.10).

Equation (1) can be written as $\ddot{x} = -\omega_0^2 x$, which corresponds to the canonical form of the harmonic oscillator with $\omega_0 = \sqrt{|a_{01} b_{10}|}$ the angular frequency. This model is conservative – the energy $E = \dot{x}(t)^2/2 + \omega_0^2 x(t)^2/2$ is constant for all *t*—and holds circular orbits of radius 2*E* if $|a_{01}| = |b_{10}|$ or ellipses in any other case (note that if energy is written using the *x* and *y* variables of eq. (1), it takes constant values whenever $a_{01}^2 y(t)^2/2E + |a_{01} b_{10}| x(t)^2/2E = 1$).

From the bottom panel (b) in Fig. 3, we note that the next more frequently appearing coefficient corresponds to the $\dot{y}$ equation—coefficient $b_{11}$—and has a positive sign. Next are the coefficients that are a constant term in the equations for $\dot{x}$ and $\dot{y}$— $a_{00}$ and $b_{00}$, respectively. However, these constants have different signs across the inferred models from the different realisations. This sign variability could potentially determine whether a fixed point is linearly stable or not, becoming a likely control parameter for the model's bifurcations; such as what happens in a saddle-node bifurcation[31]. The remaining coefficients show frequencies below 0.05 and with characteristic mean values that are at least three orders of magnitude smaller than the leading linear terms.

Figure 3 gives a statistical view of the coefficients across the total set of 1902 inferred dynamical models (linear and non-linear) and their values, but it leaves the structure of any given model undefined, i.e., which are the specific non-null coefficients in each model. By taking this into account, we group the 1902 climatological models into 76 classes, where any two classes differ by containing models that have at least one non-null coefficient, which is present in one class but not in the other.

Figure 4 shows the structure of the models with non-null coefficients, where the structures define different classes. The 11 classes in Fig. 4 appear more than 1% of the time (i.e., fr > 0.01), containing a total of 1712 models (namely, 90% of the 1902 inferred models) sharing a basic structure of coefficients corresponding to the harmonic oscillator, which is consistent with the results of Fig. 3. The differences between these classes appear because of coefficients that skew the harmonic oscillator dynamics. Specifically, ~66% of the inferred models in Fig. 4 are linear oscillatory models – *C.M*1, *C.M*3, *C.M*4, *C.M*7, and *C.M*8—while the remaining ~24% are non-linear oscillatory models—*C.M*2, *C.M*5, *C.M*6, *C.M*9, *C.M*10, and *C.M*11. Consequently, these results show that the deterministic component of the MJO – represented by OMI's two principal components – can be frequently modelled (i.e., at least 90% of the time) by a linear oscillator with small non-linear corrections.

The first two classes in Fig. 4 contain 68% of the models. The largest class, *C.M*1 (first row in Fig. 4), contains 53% of all the models and corresponds to linear harmonic oscillators. The following class, *C.M*2 (second row in Fig. 4), contains 15% of the models, and it is a non-linear oscillator with the non-linear *xy* term in the $\dot{y}$ equation. These models are part of the Lienard family of non-linear systems [31], and the *xy* term can be interpreted as an 'energy-like' source/sink (depending on the coefficient's sign).

In order to have a representative climatological model for each class, we use the average value of the coefficients for the models in each class. For the classes *C.M*1 and *C.M*2, we get the following representative models:

$$C.M1: \begin{cases} \dot{x} = \langle a_{01} \rangle\, y, & \langle a_{01} \rangle \simeq -0.108\, days^{-1}, \\ \dot{y} = \langle b_{10} \rangle\, x, & \langle b_{10} \rangle \simeq 0.102\, days^{-1}, \end{cases} \tag{2}$$

where $T_{C.M1} = 2\pi/\omega_0 = 2\pi/\sqrt{|\langle a_{01} \rangle \langle b_{10} \rangle|} = 2\pi/0.105\, days^{-1} \sim 59.9\, days$ is the resultant harmonic oscillator period and

$$C.M2: \begin{cases} \dot{x} = \langle a_{01} \rangle\, y, & \langle a_{01} \rangle \simeq -0.108\, days^{-1}, \\ \dot{y} = \langle b_{10} \rangle\, x + \langle b_{11} \rangle\, xy, & \langle b_{10} \rangle \simeq 0.102\, days^{-1},\ \langle b_{11} \rangle \simeq 0.015\, days^{-1}, \end{cases} \tag{3}$$

for the Lienard system.

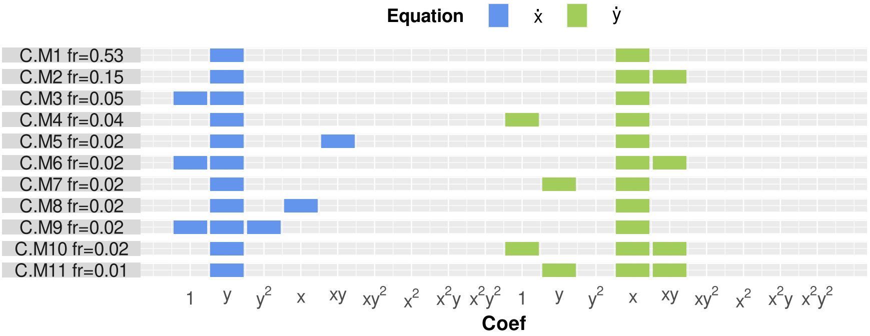


**Fig. 4 Model's structure I.** Coefficient structure of the most frequently ($fr > 0.01$) found classes of models. Models *C.M*1, *C.M*3, *C.M*4, *C.M*7 and *C.M*8 present only linear dependencies, while models *C.M*2, *C.M*5, *C.M*6, *C.M*9, *C. M*10 and *C.M*11 present at least one non-linear dependency. For each class (row), the non-null coefficients are indicated by blue ($\dot{x}$ equation) and green ($\dot{y}$ equation) colours.

We base our analysis of the *C.M*1 and *C.M*2 models on numerical simulations from a set of pre-defined initial conditions. Specifically, we obtain the trajectories for these representative models by integrating Eqs. (2) and (3) from a set of uniformly drawn initial conditions of $y(t=0)$ and fixed $x(t=0)=0$. Namely, the initial conditions correspond to starting the MJO's evolution at the end of phase 2. Figure 5 shows the resultant trajectories which are closed curves for both models—as seen from the black curves—where the speed along any given orbit is given in colour scale.

We note that the *C. M*1 linear model on the left panel (a) of Fig. 5 shows elliptic trajectories with low eccentricity and nearly-uniform speed – as expected from a harmonic oscillator with $|a_{01}|$~$|b_{10}|$. Also, the period of this model is independent of the initial condition, which explains the radial increase in speed over the set of trajectories. On the other hand, the *C. M*2 non-linear model on the right panel (b) of Fig. 5 shows egg-like shaped trajectories with non-uniform speed and a period that depends on the initial condition. In particular, we note that the slowest [fastest] velocities happen when the trajectories are going through phases 1–4 [5–8], increasing their magnitude as the orbit's amplitude is increased, which corresponds to intense MJO events. Consistently, both models have similar dynamics whenever the amplitude takes small values ($\lesssim 1$).

Figure 6 shows the total time that each trajectory spends over the different MJO phases according to *C.M*1 or *C.M*2 representative models. Given the models' symmetry with respect to the *y*-axis, we only evaluate these times for phases 3–4 and 5–6 (whose symmetrical opposites across the vertical axis are phases 1–2 and 7–8, respectively).

From the top panel (a) of Fig. 6, we observe that the linear model *C.M*1 (magenta colour) spends nearly 15 days over phases 3–4 (circles) and 5–6 (diamonds) irrespective of the initial condition. This models an MJO with quasi-uniform speed over different phases and a definite period of nearly 60 days, which can be seen in the bottom panel of Fig. 6 (in accordance with Figs. 2 and 5).

On the other hand, Fig. 6 shows that the non-linear representative model *C.M*2 (orange colour), has a slower propagation over the phases 3–4 (circles) with respect to the phases 5–6 (diamonds). In particular, both regions show a dependency on the initial condition, with larger time differences for trajectories starting further away from the origin. As a result, *C.M*2 models an MJO with a slower propagation over phases 1–4 compared to phases 5–8, with time differences that range from nearly 3 days for trajectories close to the origin to 7 days for trajectories far from it.

Comparing the *C.M*1 and *C.M*2 models, we note that the non-linear (orange) curves approach the linear (magenta) curves as the initial condition gets smaller. The reason is that the non-linear term gets vanishingly small with respect to the linear ones for trajectories close to the origin, where the dynamics of both models coincide. From the bottom panel (b), we also note that the non-linear period is always larger than the linear one, which shows that the non-linear term affects MJO's mean global speed of propagation by slowing it down.

Overall, the representative non-linear model shows an enhanced asymmetry in its propagation characteristics—showing a non-uniform behaviour across the different phases of the MJO—and a larger period—which approaches the linear one for trajectories close to the origin.

## Analysis of models according to ENSO's modulation

From the $N=1000$ realisations of $m=2^8$ randomly selected OMI data during the 14 El Niño years and during the 15 La Niña years, we get a total of 2894 and 2136 models, respectively. Similarly to the models drawn from all the available data, a large fraction of these ENSO-dependent models correspond to linear dynamical systems with a definite period. Specifically, we find 1002 linear models out of the 2894 for El Niño and 1834 linear models out of the 2136 for La Niña. Their distribution of characteristic periods for these two types of ENSO-dependent linear models are shown in Fig. 7.

We note that both distributions lie within the IS time-scale, with La Niña distribution presenting a larger mean characteristic period (~66 days) in comparison with El Niño (~58 days). Also, the La Niña distribution of periods is wider, with a standard deviation of $Sd\sim3.0$ days against a $Sd\sim2.2$ days for El Niño. We also note that the climatological distribution of periods shown in Fig. 2 is located between the El Niño and La Niña distributions. In particular, these results show that La Niña years are characterised by a slower MJO propagation with a larger variability around its mean period.

Similarly to 'Climatological data-driven models of the MJO', we group all the inferred models for the El Niño and La Niña years into classes, each class being characterised by a particular set of non-null coefficients. The most frequent ($fr > 0.01$) classes of models for El Niño (a) and La Niña (b) years are shown in Fig. 8, where we can see that 8 and 9 classes account for nearly 92% and 93% out of the total models, respectively.

As for the climatological models, we find that all ENSO modulated classes of models share a common structure: the harmonic oscillator. Moreover, the harmonic oscillator is still the most frequently inferred model in both ENSO phases (i.e., *EN.M*1 and *LN.M*1). We note that the remaining classes show models containing small corrections to the harmonic oscillator. However, the El Niño classes contain more non-linear models than linear ones—57% of non-linear models (*EN.M*2 to *EN.M*8) versus 35% of linear ones (*EN.M*1)—whilst the La Niña classes mostly contain linear models—with 83% of linear models (*LN.M*1, *LN.M*2, *LN.M*3, *LN.M*4, *LN.M*8 and *LN.M*9) against 10% of non-linear models (*LN.M*5 to *LN.M*7).

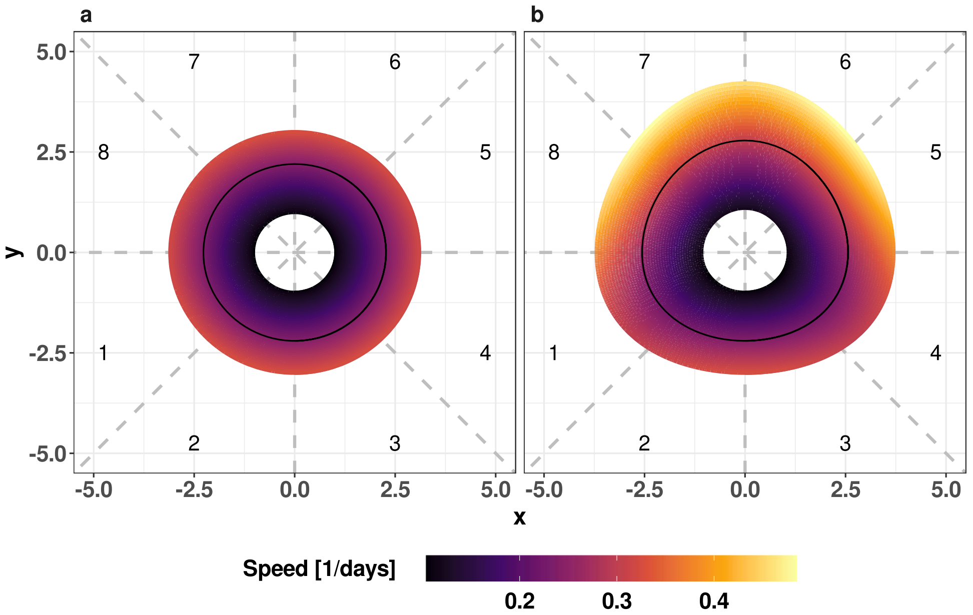


**Fig. 5 Model's trajectories I.** Trajectories for the representative linear model from class *C.M*1 (**a**) and non-linear model from class *C.M*2 (**b**). The set of trajectories are integrated for $t = 65$ *days* with a time increment of $\Delta t = 0.01$ and using initial conditions with $x(t = 0) = 0$ and a set of $y(t = 0)$ values from $-1$ to $-3$ with a step of $-0.05$. The colour-bar shows the local speed of each trajectory in $days^{-1}$ and dashed lines show MJO's phase boundaries. The black curves represent an arbitrary highlighted trajectory for each model.

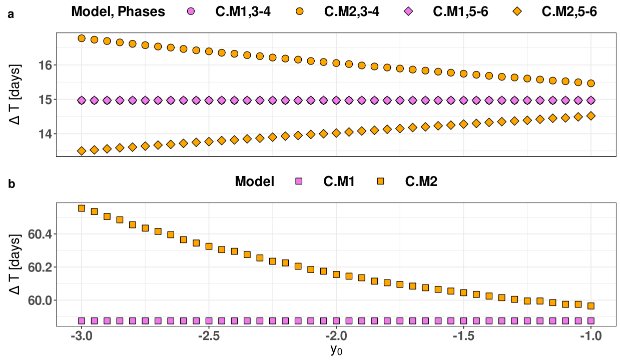


**Fig. 6 Propagation times I.** MJO's propagation time according to the *C.M*1 (magenta) and *C.M*2 (orange) average models. **a** The panel shows the dependence on initial condition for the propagation time across phases 3–4 (circles) and 5–6 (diamonds). **b** The panel shows the models' period dependency on the initial condition.

In order to get a representative model for a given class, we take the average value of the coefficients within that class. In particular, we construct the representative models for the most frequently appearing classes (aside from the harmonic oscillator). Namely, models *EN.M*2 and *EN.M*3 (with frequencies $fr = 0.34$ and $fr = 0.16$) for El Niño years and *LN.M*2 and *LN.M*3 (with frequencies $fr = 0.14$ and $fr = 0.12$) for La Niña years. For the El Niño, these representative models are defined by the following equations (which follow from Fig. 8).

$$EN.M2:\begin{cases}\dot{x}=\langle a_{01}\rangle\, y, & \langle a_{01}\rangle \simeq -0.110\,days^{-1}\\ \dot{y}=\langle b_{10}\rangle\, x+\langle b_{11}\rangle\, xy, & \langle b_{10}\rangle \simeq 0.111\,days^{-1},\ \langle b_{11}\rangle \simeq 0.029\,days^{-1},\end{cases} \tag{4}$$

$$EN.M3:\begin{cases}\dot{x}=\langle a_{01}\rangle\, y+\langle a_{11}\rangle\, xy, & \langle a_{01}\rangle \simeq -0.111\,days^{-1},\ \langle a_{11}\rangle \simeq 0.017\,days^{-1}\\ \dot{y}=\langle b_{10}\rangle\, x+\langle b_{11}\rangle\, xy, & \langle b_{10}\rangle \simeq 0.111\,days^{-1},\ \langle b_{11}\rangle \simeq 0.029\,days^{-1}.\end{cases} \tag{5}$$

A set of trajectories obtained from numerical simulations of these representative models of El Niño years are shown in the top panels (a and b) of Fig. 9. We note that the *EN.M*2 has the same structure as the non-linear climatology model (*C.M*2), but with a larger non-linear coefficient ($b_{11}$). This amplifies the non-linear effects, as can be seen by comparing the orbits generated by these models. Namely, the *EN.M*2 shows a more pronounced asymmetry with respect to the *x*-axis, with lager amplitudes—particularly over the phases 5 to 8—and a maximum speed which approximately doubles the one shown in *C.M*2. This pronounced asymmetry has a major role in the average time the MJO spends on each phase (see Fig. 10).

The *EN.M*3 shares the same structure as the *EN.M*2 with an extra (non-linear) polynomial term—the *xy* dependency on the $\dot{x}$ equation given by $a_{11}$. We highlight that the average coefficient $\langle a_{11}\rangle$ has the same order of magnitude as the non-linear correction $\langle b_{11}\rangle$ (although with a smaller value) and is the symmetrical *xy* dependency. Consequently, this model shows a more complex asymmetry of orbits, with the minimum (purple colours) and maximum (orange colours) local speeds found over the transition regions from phases 3–4 and 7–8, respectively.

On the other hand, the representative models during La Niña years, *LN.M*2 and *LN.M*3, are linear dynamical systems defined by

$$LN.M2:\begin{cases}\dot{x}=\langle a_{01}\rangle\, y, & \langle a_{01}\rangle \simeq -0.096\,days^{-1}\\ \dot{y}=\langle b_{10}\rangle\, x+\langle b_{01}\rangle\, y, & \langle b_{10}\rangle \simeq 0.098\,days^{-1},\ \langle b_{01}\rangle \simeq 0.017\,days^{-1},\end{cases} \tag{6}$$

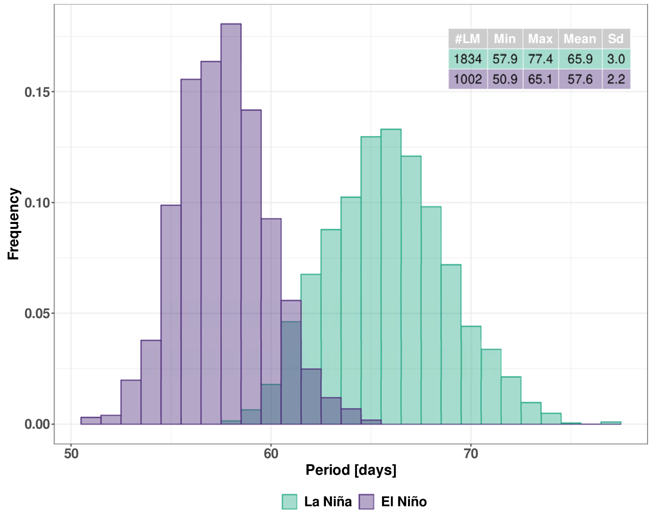
$$LN.M3:\begin{cases}\dot{x}=\langle a_{01}\rangle\, y+\langle a_{10}\rangle\, x, & \langle a_{01}\rangle \simeq -0.097\,days^{-1},\ \langle a_{10}\rangle \simeq -0.016\,days^{-1}\\ \dot{y}=\langle b_{10}\rangle\, x, & \langle b_{10}\rangle \simeq 0.095\,days^{-1}.\end{cases} \tag{7}$$

Equations (6) and (7) have the structure of dissipative harmonic oscillators, where the damping is controlled by the $\langle b_{01}\rangle$ and $\langle a_{10}\rangle$ coefficients, respectively. For the *LN.M*2 [*LN.M*3] representative model, $\langle b_{01}\rangle > 0[\langle a_{10}\rangle < 0]$, thus giving trajectories that increase [decrease] their amplitudes in time, as shown in the bottom panels (c and d) of Fig 9.

We now analyse the local behaviour of the MJO according to the set of selected models (i.e., *EN.M*1, *EN.M*2, *EN.M*3, *LN.M*1, *LN.M*2 and *LN.M*3). To do this, we subdivide the MJO cycle into 16 sub-phases, each of length $\pi/8$, starting from the *x*-axis (horizontal axis). We then calculate the time it takes for a given trajectory to cross each of these sub-phases. Figure 10 shows the resultant times for El Niño (a) and La Niña (b) models.

Firstly, we note that *EN.M*1 and *LN.M*1 (green solid lines)—being harmonic oscillators—present an almost constant behaviour—as we see for the climatology model *C.M*1. Also, the mean time for *EN.M*1 representative model is below 4 days in each 1/16-th of its cycle, while for the *LN.M*1 is above—consistent with the distributions shown in Fig. 7.

Secondly, we note that for the El Niño representative models *EN.M*2 (orange) and *EN.M*3 (purple), the times elapsed present two local minima (which are fast regions) at $\frac{\pi}{4}$ and $\frac{7\pi}{8}$, and two local maxima (which are slow regions) at $\frac{5\pi}{4}$ and $\frac{15\pi}{8}$. Moreover, the average time for phases between 0 and $\pi$ is smaller than the average time for the phases between $\pi$ and $2\pi$ (approximately 25 and 33 days, respectively). This shows that, according to these models, the MJO presents different speed of propagation over phases 1–4 and 5–8. Specifically, if we consider the longitudinal span of phases 5–8 and 1–4 over the Equator to be 240° and 120°, respectively, we find that the MJO presents a faster propagation speed of nearly 12.4 m s$^{-1}$ over phases 5–8 (from western Pacific to Africa approximately) and a slower speed of nearly 4.6 m s$^{-1}$ over phases 1–4 (Indian Ocean to western Pacific region). This behaviour can also be appreciated from the top panels in Fig. 9.



| #LM | Min | Max | Mean | Sd |
|---|---|---|---|---|
| 1834 | 57.9 | 77.4 | 65.9 | 3.0 |
| 1002 | 50.9 | 65.1 | 57.6 | 2.2 |

**Fig. 7 Distribution of periods II.** Same as Fig. 2 for the El Niño (purple) and La Niña (turquoise) years.

The shaded areas in each ($\pi$/8) sub-phase of Fig. 10 represent the minimum and maximum propagation times among the set of initial conditions considered. This allows us to identify regions in which the non-linearity has significant effects and in which regions it is insignificant. We see that for *EN.M*2 and *EN.M*3 the MJO's behaviour depends weakly on the initial condition at $\frac{\pi}{2}$, $\pi$, and $\frac{3\pi}{2}$ approximately. However, there is a strong dependency on initial conditions over $\frac{\pi}{4}$, $\frac{7\pi}{8}$, $\frac{5\pi}{4}$, and $\frac{15\pi}{8}$. These strongly non-linear regions coincide with the position of the local minima and maxima, which could indicate that over these regions atmospheric non-linear processes may be happening, ultimately defining the MJO's local speed of propagation; namely, where it accelerates and decelerates.

Finally, we note that the La Niña representative models *LN.M*2 and *LN.M*3 show a well-defined oscillatory behaviour around the time it takes for the *LN.M*1 to propagate across each sub-division of MJO phases. Both models are basically in phase, presenting two minima at $\frac{\pi}{4}$ and $\frac{5\pi}{4}$, and two maxima at $\frac{3\pi}{4}$ and $\frac{7\pi}{4}$. In contrast to the behaviour of El Niño models, we find that there is no distinction in the characteristic times for phases 1–4 and 5–8 in La Niña models.

## DISCUSSION

Our results show that the inferred climatological and ENSO-dependent models can be grouped into distinct classes, where each class contains models with the same structure of non-null coefficients, while showing variability in the values taken by these coefficients, ultimately determining the equations of motion. Nevertheless, all models across these classes have a common harmonic oscillator core dynamics. This harmonic oscillator captures some basic and expected behaviours of the MJO's evolution, such as its eastward propagation and characteristic time of evolution. In addition, each class contains models with small linear or non-linear corrections to the harmonic oscillator that account for other important and observed behaviours in MJO's evolution and its dependency on El Niño and La Niña years. We discuss here, first the results obtained for the climatological

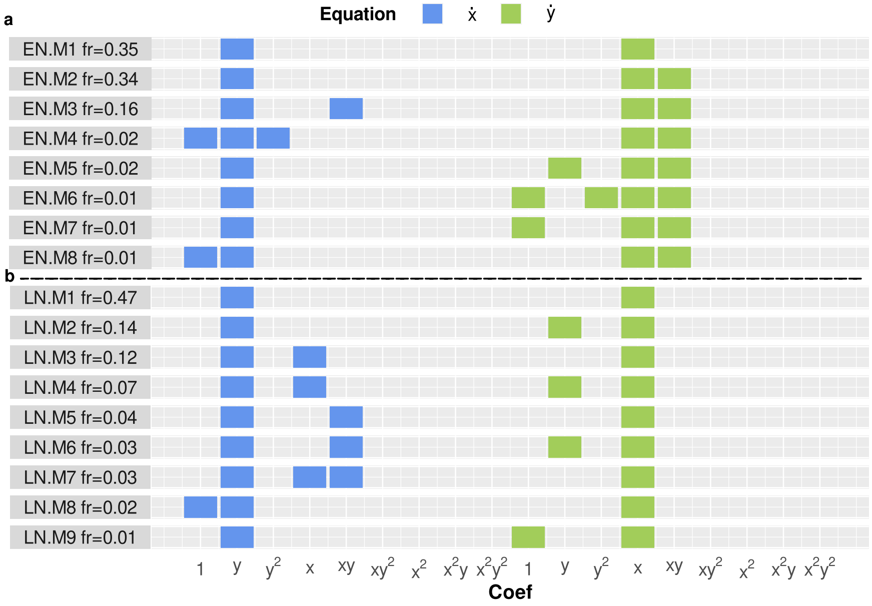


**Fig. 8 Model's structure II.** Structure of inferred models as in Fig. 4 but for (**a**) El Niño and (**b**) La Niña years.

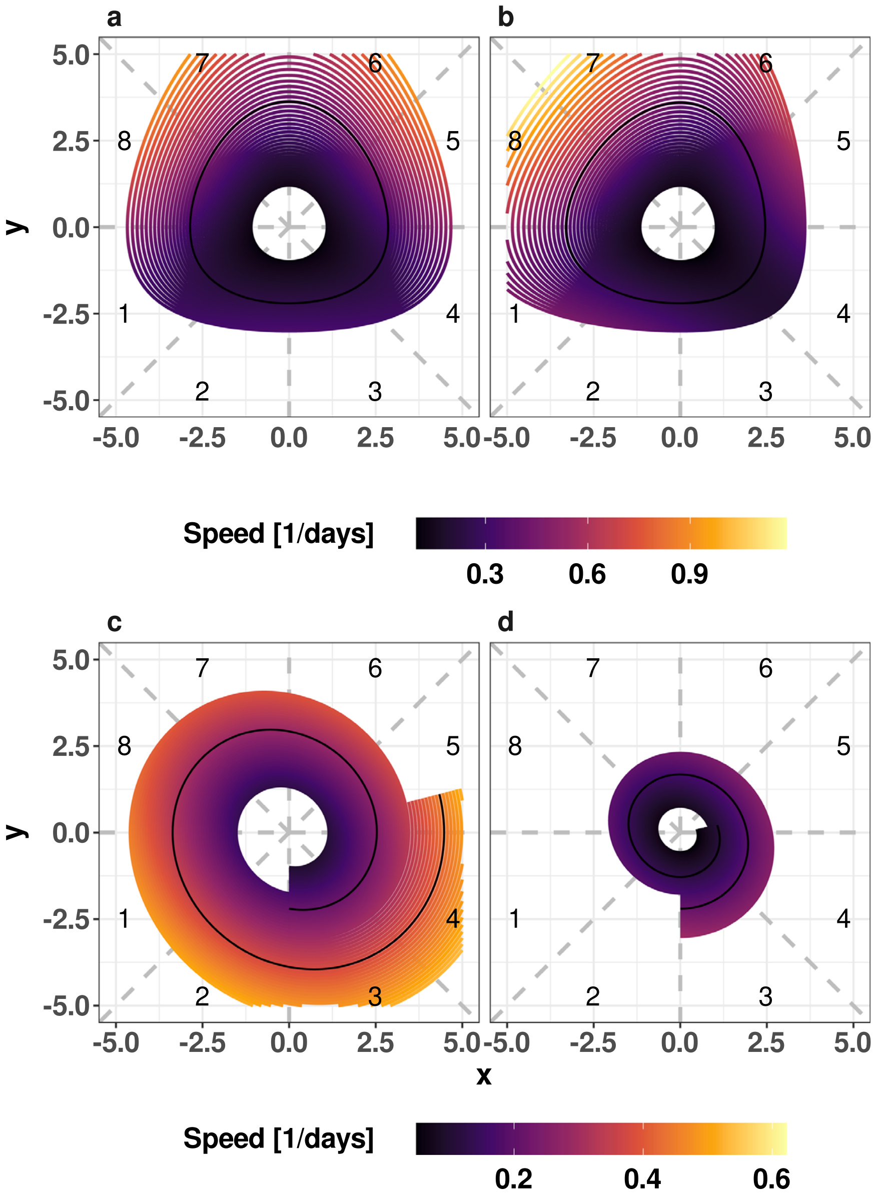


**Fig. 9 Model's trajectories II.** Same as in Fig. 5 for El Niño (top panels (**a**, **b**)) and La Niña (bottom panels (**c**, **d**)) years. For El Niño [La Niña], we show the phase space of *EN.M*2 [*LN.M*2] on the left and *EN.M*3 [*LN.M*3] on the right.

models, and then those corresponding to ENSO modulated models.

By repeatedly applying SINDy to $m$ random samples taken from all the available data (i.e., the whole December to March period from 1979 to 2021), we find that most inferred models are linear. These linear dynamical systems are oscillators with periods within the IS time-scale (see Fig. 2). In particular, the mean period from all of the linear models we get from 1000 independent sampling realisations (each with $m$ samples) is approximately 60 days, which is within the expectations for the MJO's evolution. We show that this result holds independently of the number $m$ of random samples taken to construct the climatological models, implying that SINDy is a robust methodology to model MJO's evolution. On the other hand, the spread of the distribution of periods from the linear climatological models depends on $m$, becoming narrower as $m$ is increased.

In order to fix $m$, one can attempt to follow ref. [32], where the authors apply SINDy to synthetic signals from noiseless dynamical systems and study the dependence of the inferred models (which should recover the original system generating the signals) with the number of random samples relative to the characteristic period of the system. They show that for a sampling ratio of $f_s = 2^6 = 64$ samples/period, one should measure nearly two periods in order to correctly infer the underlying dynamical system. In this work, we have daily data and the inferred characteristic period of the models is close to 60 days, thus, our sampling ratio is $f_s \approx 2^6$ samples/period. Since the underlying dynamical system is unknown and it is noisy, we choose to have $m = 2^8 = 256$ random samples, which nearly covers four periods of the MJO models.

From the results shown in Figs. 5 and 6, we note that the most frequently inferred model structure is the harmonic oscillator, *C.M*1. This is the most basic model that can account for MJO's fundamental behaviour: its eastward propagation and the IS time-scale of evolution. However, to gain further insights about other MJO behaviours, one needs to consider corrections to the dynamics of the harmonic oscillator. For example, several works[3,33,34] have shown that the MJO's speed has a non-uniform

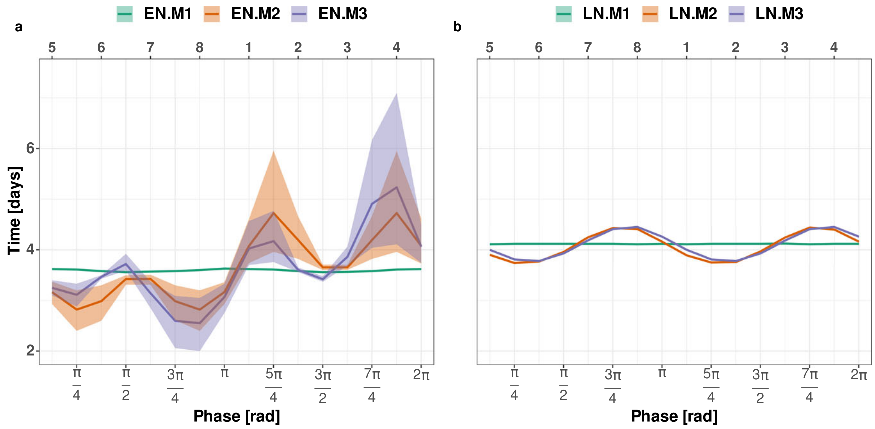


**Fig. 10 Propagation times II.** MJO's time of propagation for El Niño (**a**) and La Niña (**b**) models. The mean time over initial conditions is shown in solid line while the minimum and maximum times are represented with the shadow colour-bands. The time is calculated for each $\pi/8$ sub-phase, and each sub-phase is identified through the usual polar angle measured from the *x*-axis. The corresponding MJO phase is shown on top of each panel.

behaviour, with slow characteristic speeds over the Indian Ocean region and fast speeds over Pacific Ocean region. The acceleration – typically seen to the east of the Maritime Continent—can be understood as a decopuling of the convective and circulatory components of the MJO. The decoupling occurs due to the MJO entering the cold waters of the eastern Pacific Ocean and a corresponding reduction of the moist convergence to the east of MJO's convective centre[15,16,28].

Figure 6 shows that according to the linear *C.M*1 model, the MJO spends equal amounts of time (~15 days) over each pair of phases, resulting in a characteristic global speed of 7.7 m s$^{-1}$. However, for the non-linear *C.M*2 model, the total time over each pair of phases depends on the position: for the low phases 1–4, the time averaged over initial conditions is ~32 days, while for the high phases 5–8, it is ~28 days. Given that the phases are not uniformly distributed along the zonal direction, this time difference is more noticeable in their respective speeds. Considering a longitudinal span of 120° [240°] for the phases 1–4 [5–8], we get speeds of nearly 4.8 m s$^{-1}$ and 11.0 m s$^{-1}$ for the Africa to Maritime Continent region and Western Pacific to Africa region, respectively. This shows that, even though the non-linear perturbation in *C.M*2 is small, it accounts for the MJO's non-uniform speed of propagation.

Overall, the non-linear model *C.M*2—which is the *C.M*1 model with a small non-linear correction—adequately represents MJO's period, eastward propagation, and captures the local dependency of MJO's speed, showing large and small values over the Pacific and Indian Ocean regions, respectively.

There is extensive literature reporting changes in the MJO's behaviour during El Niño and La Niña years[18–21,34]. It has been shown that during La Niña years, the MJO has a more intense convective centre over the Indian Ocean and Maritime Continent regions, and a quasi-stationary behaviour with a slow eastward propagation speed of nearly 1.5 m s$^{-1}$ to 5.0 m s$^{-1}$. On the other hand, findings show that for El Niño years the convective centre is weaker, the eastward speed of propagation is larger (with values between 5.0 m s$^{-1}$ to 8.0 m s$^{-1}$ over the Indian Ocean), and there is enhanced activity over the Central Pacific as a result of the MJO being able to propagate further east beyond the Maritime Continent. These changes are believed to be associated with the sea surface temperature distribution and the corresponding Walker Circulation low frequency variability[15,16,20,21], which is mainly controlled by ENSO.

From the set of linear models that we infer for the MJO during ENSO years, we find that the distribution of periods (from the different data realisations that generate the models) have a mean characteristic period of 58 days for El Niño and 66 days for La Niña years (Fig. 7). We find that the difference between these periods and their variances is significant—we perform a two-tailed mean difference test and a Leven's test of variance homogeneity. The corresponding mean global speed values are ~8.0 m s$^{-1}$ and ~7.0 m s$^{-1}$, which imply that the MJO is faster during El Niño years. Although this difference in speed is not as large as some values reported in the literature[20,21,34], we highlight that it is still captured here by low dimensional linear models, which can be improved when including non-linear corrections.

The total set of MJO models conditioned to El Niño and La Niña years carry important implications.

Firstly, we note that the number of models inferred for El Niño years—and therefore the variability of models needed to describe the MJO's behaviour—is larger than the number of models inferred for La Niña years: 2894 versus 2136. This implies that MJO's behaviour during El Niño years is more complex, hence, more difficult (a priori) to predict. This complexity during El Niño years can also be seen by analysing the fraction of linear vs non-linear models: while we find that the MJO is predominantly non-linear during El Niño (1892 non-linear models out of 2894), it generally presents a linear behaviour during La Niña (1834 linear models out of 2136). This distinction is crucial in determining the non-uniform regional speed over the MJO's phases during El Niño years, compared to the uniform behaviour shown during La Niña years (as shown in Fig. 10).

Secondly, we show (Fig. 8) that the MJO models during El Niño and La Niña years can be grouped into 8 and 9 classes of models (containing nearly 92% and 93% of the total number of models), respectively. When comparing the two most frequently appearing non-linear models for El Niño years (*EN.M*2 and *EN.M*3 classes) and the corresponding linear models for La Niña years (*LN.M*2 and *LN.M*3 classes), we note that all of them present corrections to the harmonic oscillator's basic structure that act as an 'energy-like' source/sink (Fig. 9), i.e., dissipative terms. The main difference is that for the *EN.M*2 and *EN.M*3 models, these corrections are position dependent, changing sign according to the location in phase space. Specifically, they add and subtract equal amounts of energy along a given trajectory, resulting in close trajectories. In contrast, for La Niña models these source/sink terms have definite signs. Therefore, they add energy in the *LN.M*2 class or subtract energy in the *LN.M*3 class. This is inconsistent with MJO's behaviour, which is known to have an initiation, growth, and decay life-cycle[3,34].

We believe that the *LN.M*2 and *LN.M*3 classes represent the growing and decaying stages of MJO's evolution, but neither of them is valid over all phase space. In this sense, we note that the number of models in each class is similar (with frequencies of ~0.14 and ~0.12, respectively), so it is possible that a combination of them could create a self-consistent model valid over all phase space. Moreover, a linear combination of $LN.M2$ and $LN.M3$ would leave the characteristic times of propagation over each sub-phase unchanged (as shown in the right panel of Fig. 10, both models have near identical propagation times).

Finally, Fig. 10 shows two important results. The most relevant is that MJO's non-linear models during El Niño years present different mean speeds over two large regions: a fast propagation over the Pacific Ocean to Africa's region and a slow propagation over the Indian Ocean to Maritime Continent regions. This regional difference is not seen during La Niña years, for which $LN.M2$ and $LN.M3$ show the same mean time of propagation. This acceleration is a known characteristic of MJO's evolution, but according to our models, it only occurs during El Niño years and not during La Niña years.

The other result is that within these two large regions, spanning from phases 5 to 8 and from 1 to 4, the non-linear models for El Niño years show respectively two local minima and two local maxima for the times of propagation (panel a in Fig. 10). Moreover, the evolution over these regions is strongly dependent on the initial condition, which implies that the system's non-linearity is significant over these locations. In particular, we note that the first minima and maxima (located at $\pi/4$ and $5\pi/4$) as well as the second pair of minima and maxima (located at $7\pi/8$ and $15\pi/8$) are $\pi$ apart from each other. Given that the MJO is nearly a wavenumber 1 process – which is captured by using OMI's first two principal components – each of these pairs correspond to the convective and dry centres passing through the same location. The first pair, corresponds to the convective/dry centre passing over the Maritime Continent region, while the second pair correspond to the convective/dry centre passing over the South America and Congo convective regions. We hypothesise that when the MJO passes over these regions, its convective and dry centres interact in opposite ways with the climatological convective centres, giving rise to the fast and slow non-linear behaviours.

On the other hand, the linear models for La Niña years show (right panel in Fig. 10) one minima and one maxima over the 5–8 phases and 1–4 phases, respectively. However, these corresponds to small departures from the mean value, and are related to the shape of the trajectories of these linear models and not to a possible interaction such as with the El Niño non-linear models. We note that in order to assess the possible local interactions – whether linear or non-linear—we would need to incorporate data beyond OMI's two principal components, since these mainly represent the MJO's large scale variability. Consequently, these results leave open questions regarding the regional interaction responsible for the acceleration and deceleration of the MJO along its evolution, as well as the differences between the MJO models we infer during El Niño and La Niña years. Future work could also focus on finding how the closed trajectories generated by our models relate to circumnavigating MJO events[35–38].

To summarise, in this work we apply a recently developed methodology (known as SINDy) in order to look for two-dimensional data-driven models of the MJO and to diagnose its evolution during the warm season of the Austral hemisphere, i.e., December to March. We consider an OMI data-set as SINDy's input, where the OMI is a principal component index based solely on the outgoing longwave radiation field and serves as a proxy for MJO's convective activity. We then look for models of the MJO's evolution that best fit $m$ random samples of the OMI data from 1979 to 2021 – which we name as climatological models—or the years within that range which have El Niño events or La Niña events—ENSO modulated models. In each case, a set of models is constructed for the corresponding OMI data by taking $N = 1000$ realisations of the random samples, which ensures sufficient statistical power to analyse the resultant model structures and generated trajectories that represent the MJO's evolution.

We show that both types of MJO inferred models—climatological or ENSO modulated—can be grouped into disjoint classes according to the structure of their equations of motion, i.e., according to their non-null coefficients. We also show that some structures appear more frequently than others (Figs. 4 and 8), focusing our analysis in the most frequent models, which are the models most likely to be inferred by a random sampling of the OMI data.

Our main results show that irrespective of the years considered, the MJO can be robustly represented as a harmonic oscillator that captures MJO's eastward propagation and its characteristic time of evolution. We find that these harmonic oscillators present periods of 60, 58 and 66 days for the climatology, El Niño, and La Niña models, respectively. This shows that the MJO is faster during El Niño years compared to La Niña years, a behaviour which has been previously found in the literature[19–21,34].

The remaining models have other linear or non-linear terms upon the basic structure of the harmonic oscillator, although these new terms are at least one order of magnitude smaller. This means that the MJO can be modelled as a harmonic oscillator with small linear or non-linear corrections. Our results show that these corrections allow the models to include other known behaviours of the MJO, such as the varying local speed of propagation, whilst maintaining the eastward propagation and characteristic time of evolution (with small modifications). Particularly, during El Niño years, we find that the MJO is most frequently (65%) modelled as a non-linear oscillator, while during La Niña years, it is most frequently (86%) modeled as a linear oscillator. We believe that this difference could be a consequence of the different mean states through which the MJO propagates during El Niño and La Niña years. That is, during La Niña years, the mean state is characterised by an intensification of the Walker Cell and the convective activity over the Maritime Continent, whilst for El Niño years, this structure is shifted eastward resulting in a more noticeable perturbation of the climatological mean-field. Nevertheless, further research is needed in order to clarify the reason behind the different dynamics.

For the climatological models, the next most frequently inferred model – aside the harmonic oscillator ($C.M1$)—is a non-linear oscillator ($C.M2$) belonging to the family of Lienard's systems. From numerical simulations, we see that this model has self-sustained oscillations near the phase space's origin, where the non-linear term injects and subtracts the same amount of energy over different regions of the phase space, resulting in close trajectories and a varying speed of propagation. The representative model for the $C.M2$ class of models, presents a larger period than the representative harmonic oscillator ($C.M1$ class of models) that depends on the initial conditions (contrary to the linear models), and a non-uniform speed of propagation, with faster propagation (11.0 m s$^{-1}$) over the Pacific Ocean region (phases 5–8) and slower (4.8 m s$^{-1}$) over the Indian Ocean region (phases 1–4). This acceleration of the MJO when it passes the Maritime Continent has been reported before refs. [3,33,34].

For the El Niño models, the most frequent non-linear models $EN.M2$ and $EN.M3$ have almost identical evolution as the non-linear climatological model $CM.2$. They also present a region over the phase space with a self-sustained oscillatory behaviour characterised by a non-uniform speed of propagation, which results in fast and slow regions of propagation when the MJO's convective centre is located over the Pacific and Indian Oceans, respectively. However, the non-linear term is stronger in these models than in the $CM.2$ models. As a result, MJO's acceleration is larger for the $EN.M2$ and $EN.M3$ models, with characteristic speeds of 4.6 m s$^{-1}$ and 12.4 m s$^{-1}$ over the slow and fast regions. Moreover, we find that these models present two slow regions over the Indian Ocean (maxima in Fig. 10), and two fast regions over the Pacific Ocean (minima in Fig. 10) where the first minima and maxima, and second minima and maxima, are separated by $\pi$. We interpret this behaviour to be the result of a non-linear interaction between the convective and dry centres of the MJO with—possibly—the climatological convective centres located over the Maritime Continent region, and over the South America to Africa region.

For the La Niña models, *LN.M*2 and *LN.M*3 are dissipative linear oscillators that increase or decrease their amplitudes as they evolve. These behaviours can be interpreted as the growing and decaying stages of the MJO, respectively. However, we note that neither of these models are suitable to describe MJO's evolution by themselves. Consequently, we believe that a combination of them is required in order to get a reasonable model for MJO's evolution over any region of the phase space.

Finally, we note that our work sets the grounds for this type of approach to be further investigated either for the MJO or other climate system phenomena.

## METHODS

### Dataset

To characterise the time evolution of the MJO convective centres, we use the outgoing longwave radiation (OLR) MJO index (OMI) introduced by ref. [25]. The index is defined via the first two principal components (PC1 and PC2), which results from projecting the 20–96 days filtered OLR field onto the spatial patterns of the first two empirical orthogonal functions of the 30–96 days filtered OLR field with only eastward propagating wave numbers. The OMI index resulting from these procedures shows a more regular and smoother behaviour than the typical Wheeler and Hendon index[39]; this smoother behaviour benefits SINDy method as it looks for the deterministic component of the system under study. Also, the use of OMI is advisable when MJO's convective centres are the focus of interest[25]. We analyse MJO's behaviour through the OMI index for all days from December to March (DJFM), within the period from 1979 to 2021.

### Sparse identification of nonlinear dynamics

Sparse identification of non-linear dynamics (SINDy) is a method recently introduced in ref. [24] which addresses the problem of seeking the fundamental dynamics of a system from empirical data. It iteratively solves a multiple regression problem with a sparsity promoting hard thresholding procedure.

Consider an autonomous dynamical system of dimension $d$ with the general form,

$$\dot{\mathbf{X}}(t) = \mathbf{f}(\mathbf{X}(t)), \tag{8}$$

where $\mathbf{X}(t), \dot{\mathbf{X}}(t) \in \mathbf{R}^d$ represents the state of the system at time t and its corresponding velocity, and $\mathbf{f}: \mathbf{R}^d \to \mathbf{R}^d$ is the function defining the dynamic and time evolution of the system.

Suppose that $m$ observations of the state variables $\mathbf{X}$ are collected over a set of times $t_1, t_2, \dots t_m$ and that the corresponding velocities can be either numerically estimated, or directly measured. The SINDy method, as well as other methodologies that seek for data driven models [40–42] of the form of eq. (8), relies on the fundamental hypothesis that the function $\mathbf{f}$ allows for a sparse representation over an -a priori- unknown base of real functions. In order to find this representation, suppose that we introduce a set of $n$ candidate functions,

$$\Theta(\mathbf{X}) = [1, \mathbf{X}, \mathbf{X}^2, \dots, \sin(\mathbf{X}), \cos(\mathbf{X}), \dots], \tag{9}$$

over which we expand linearly the function $\mathbf{f}$. We refer to this set as a dictionary, and its definition can be subjected to the specific knowledge of the system under study. Using the collected empirical data to evaluate each function in the set, and with the linear expansion of $\mathbf{f}$ over the $n$ functions in $\Theta$, we can express eq. (8) at each time $t_i$ as,

$$\dot{\mathbf{X}}(t_i) = \Xi(\mathbf{X}(t_i))\boldsymbol{\zeta}, \tag{10}$$

where $\boldsymbol{\zeta} \in \mathbf{R}^{nd}$ is an array conformed, for each of the $d$ state variables, by a vector of $n$ coefficients corresponding to the linear expansion. Note that $\Xi$ corresponds to the evaluation at any time of the set of functions $\Theta$. Eq. (10) can be written in matrix form as,

$$\dot{\mathbf{X}} = \Xi(\mathbf{X})\boldsymbol{\zeta}, \tag{11}$$

which corresponds to a multiple linear regression with $\Xi[\mathbf{X}]$ inputs, $\dot{\mathbf{X}}$ outputs, and $\boldsymbol{\zeta}$ as the array of coefficients to be found.

Generally, linear systems like the one in eq. (11) do not have a unique solution, or more yet, not even one; optimisation methods like least-squares are then needed in order to find an approximate solution to the problem. To avoid overfitted solutions, and in the spirit of findings simple parsimonious and interpretable dynamical systems, $l_1$-norm regularisation schemes like those in ref. [40,41,43] are employed. In order to deal with this, SINDy introduces a hard thresholding procedure in which a cutoff parameter $\lambda$ is used. The method seeks, iteratively, the least square solution of eq. (11) and sets to zero every coefficient found with an absolute value less than $\lambda$; this reduces, at each iteration, the dimensionality of the space of functions in which to look for the least square solution. The algorithm has been proved to converge to a sparse-like solution[44], and more detail and advancements on the subject can be found in a vast literature[24,45–48].

### Implementing SINDy on MJO

In order to apply SINDy to MJO data we introduce the variables $x = PC_2$ and $y = -PC_1$ as input data for eq. (11), and calculate -numerically- the corresponding velocities $\dot{x}$ and $\dot{y}$ as output data. The change of variables is made so that MJO's phases, $\phi = atan(-PC_1/PC_2)$, are in agreement with the usual geographical location of the convective centres when the Wheeler and Hendon index is used[25].

We work with a polynomial dictionary conformed by functions of the form $x^i y^j$, with $i$ and $j$ taking the integer values 0, 1, 2. This gives a total of 9 functions for each dimension, that goes from the constant function one as the one with smaller order, to $x^2y^2$ as the function with higher order. As a result we look for models of the general form,

$$\begin{cases} \dot{x} = \sum_{i=0}^{2}\sum_{j=0}^{2} a_{ij}x^i y^j \\ \dot{y} = \sum_{i=0}^{2}\sum_{j=0}^{2} b_{ij}x^i y^j \end{cases} \tag{12}$$

The selected dictionary is based on the fact that we will search for 2-dimensional dynamical systems which can have, as the most complex behaviour, a Hopff bifurcation, which is a structural change of the dynamics of the system from a limit cycle to a set of fixed points (or vice versa); the normal form of this type of bifurcation can be represented by the set of functions included in our dictionary.

We applied SINDy algorithm to MJO by taking $m$ number of randomly sampled data without replacement. As a first approach to finding a dynamical system, we do this taking samples from all *DJFM* years, and by repeating the process $N = 1000$ times. We test the sensibility of our results allowing $m$ to take the values $2^6, 2^7, 2^8, 2^9, 2^{10}$. As a second approach, we divide the period in El Niño and La Niña years, and repeat the former procedure with $m = 2^8$ random samples for each realisation.

### Model selection through Akaike information criterion

The threshold parameter $\lambda$ introduced by SINDy needs to be explored in order to evaluate the models found by the algorithm. If $\lambda$ is small enough, then no regressed coefficients will be set to zero and the solutions will coincide with the overfitted least square solution. If, on the other hand, $\lambda$ is large enough, all coefficients from the regression would be set to zero and no dynamical system would be found.

To take this into account, we explore the possible models obtained at each realisation allowing $\lambda^{-1}$ to take integer values between 1 and 100. In order to select a parsimonious model (or set of models) that better fits the data, we make use of the Akaike Information Criterion ($AIC$)[49]. $AIC$ is an estimate of the Kullback-Leibler Information[50], and it represents the expected relative distance between the fitted model and the unknown true model. For a linear regression problem like that in eq. (11) with independent and identically normally distributed residuals, and considering a finite sample of observations, AIC can be calculated as ref. [51]

$$AIC_c = m\log(RSS/m) + 2k + \frac{2k(k+1)}{m-k-1}, \tag{13}$$

where $AIC_c$ denotes $AIC$ corrected for finite sample, $RSS$ is the residual sum of squares, and $k$ is the total number of parameters of the model (including the residual's variance).

Given that the true model is unknown, we look for the fitted models with the smallest possible value of $AIC_c$. Introducing $AIC_c$ differences with respect to the minimum value, $\Delta AIC_c^i = AIC_c^i - AIC_{c,min}$, allow us to rank the relative support of all models with respect to the minimum ($\Delta AIC_c = 0$ is the model with most support). Following the rule introduced in ref. [51], models with $\Delta AIC_c$ between 0 and 2 have substantial support, between 4 and 7 have less support, and with values larger than 10 have essentially no support. Based on this rule, for each data realisation, we keep all models with $\Delta AIC_c < 7$ that are bi-dimensional. As a result, it is possible to select more than one model for each random sampling.

## DATA AVAILABILITY

OMI data is available at the National Oceanic and Atmospheric Administration Physical Science Laboratory (NOAA PSL) webpage [https://psl.noaa.gov/mjo/mjoindex/].

## CODE AVAILABILITY

A MatLab code for SINDy algorithm is available in ref. [24]. The underlying code for this study is available in [https://github.com/NikoDN87/SINDyMJO.git].

Received: 10 March 2023; Accepted: 15 November 2023;
Published online: 07 December 2023

## REFERENCES

1. Madden, R. A. & Julian, P. R. Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.* **29**, 1109–1123 (1972).
2. Alvarez, M. S., Vera, C. S. & Kiladis, G. N. MJO modulating the activity of the leading mode of intraseasonal variability in South America. *Atmosphere* **8**, 232 (2017).
3. Grimm, A. M. Madden–Julian oscillation impacts on South American summer monsoon season: precipitation anomalies, extreme events, teleconnections, and role in the MJO cycle. *Clim. Dyn.* **53**, 907–932 (2019).
4. Díaz, N. Barreiro, M. & Rubido, N. The distinct influence of two Madden–Julian trajectory classes on the South American dipole. *J. Clim.* **35**, 3487–3501 (2022).
5. Alvarez, M. S., Vera, CarolinaSusana, Kiladis, G. N. & Liebmann, B. Influence of the Madden–Julian oscillation on precipitation and surface air temperature in South America. *Clim. Dyn.* **46**, 245–262 (2016).
6. Jenney, A. M., Nardi, K. M., Barnes, E. A. & Randall, D. A. The seasonality and regionality of MJO impact on North American temperature. *Geophys. Res. Lett.* **46**, 9193–9202 (2019).
7. Manta, G., de Mello, S., Trinchin, R., Badagian, J. & Barreiro, M. The 2017 record marine heatwave in the Southwestern Atlantic shelf. *Geophys. Res. Lett.* **45**, 12–449 (2018).
8. Rodrigues, R. R., Taschetto, AndréaS., Sen Gupta, A. & Foltz, G. R. Common cause for severe droughts in South America and marine heatwaves in the South Atlantic. *Nat. Geosci.* **12**, 620–626 (2019).
9. Sobel, A. H. & Maloney, E. D. Effect of ENSO and the MJO on western North Pacific tropical cyclones. *Geophys. Res. Lett.* **27**, 1739–1742 (2000).
10. Camargo, S. J., Wheeler, M. C. & Sobel, A. H. Diagnosis of the MJO modulation of tropical cyclogenesis using an empirical index. *J. Atmos. Sci.* **66**, 3061–3074 (2009).
11. Pai, D. S., Bhate, J., Sreejith, O. P. & Hatwar, H. R. Impact of MJO on the intraseasonal variation of summer monsoon rainfall over India. *Clim. Dyn.* **36**, 41–55 (2011).
12. Taraphdar, S., Zhang, F., Leung, L. R., Chen, X. & Pauluis, O. M. MJO affects the monsoon onset timing over the Indian region. *Geophys. Res. Lett.* **45**, 10011–10018 (2018).
13. National Research Council et al. *Assessment of intraseasonal to interannual climate prediction and predictability*. (National Academies Press, 2010).
14. Krishnamurthy, V. Predictability of weather and climate. *Earth Sp. Sci.* **6**, 1043–1056 (2019).
15. Chen, G. & Wang, B. Circulation factors determining the propagation speed of the Madden–Julian oscillation. *J. Clim.* **33**, 3367–3380 (2020).
16. Wang, T. & Li, T. Factors controlling the diversities of MJO propagation and intensity. *J. Clim.* **34**, 6549–6563 (2021).
17. Kim, D., Maloney, E. D. & Chidong Zhang. MJO propagation over the maritime continent. *The multiscale global monsoon system*, 261–272 (World Scientific, 2020).
18. Gray, B. M. Seasonal frequency variations in the 40–50 day oscillation. *J. Climatol.* **8**, 511–519 (1988).
19. Goulet, L. & Duvel, Jean-Philippe A new approach to detect and characterize intermittent atmospheric oscillations: application to the intraseasonal oscillation. *J. Atmos. Sci.* **57**, 2397–2416 (2000).
20. Wei, Y. & Ren, Hong-Li Modulation of ENSO on fast and slow MJO modes during boreal winter. *J. Clim.* **32**, 7483–7506 (2019).
21. Suematsu, T. & Miura, H. Changes in the eastward movement speed of the Madden–Julian oscillation with fluctuation in the Walker circulation. *J. Clim.* **35**, 211–225 (2022).
22. Martin, Z. et al. The influence of the quasi-biennial oscillation on the Madden–Julian oscillation. *Nat. Revi. Earth. Environ.* **2**, 477–489 (2021).
23. Zhang, C., Adames, Á. F., Khouider, B., Wang, B. & Yang, D. Four theories of the Madden–Julian oscillation. *Rev. Geophys.* **58**, e2019RG000685 (2020).
24. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl Acad. Sci.* **113**, 3932–3937 (2016).
25. Kiladis, G. N. et al. A comparison of OLR and circulation-based indices for tracking the MJO. *Mon. Weather Rev.* **142**, 1697–1715 (2014).
26. Majda, A. J. & Stechmann, S. N. The skeleton of tropical intraseasonal oscillations. *Proc. Natl Acad. Sci.* **106**, 8417–8422 (2009).
27. Sobel, A. & Maloney, E. An idealized semi-empirical framework for modelling the Madden–Julian oscillation. *J. Atmos. Sci.* **69**, 1691–1705 (2012).
28. Vallis, G. K. Distilling the mechanism for the Madden–Julian oscillation into a simple translating structure. *Q. Jo. R. Meteorol. Soc.* **147**, 3032–3047 (2021).
29. Martin, Z., Barnes, E. & Maloney, E. Predicting the MJO using interpretable machine-learning models. (*Authorea Preprints*, 2022).
30. Silini, R., Barreiro, M. & Masoller, C. Machine learning prediction of the Madden–Julian oscillation. *npj Clim. Atmos. Sci.* **4**, 57 (2021).
31. Strogatz, S. H. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. (CRC Press, 2018).
32. Champion, K. P., Brunton, S. L. & Kutz, J. N. Discovery of nonlinear multiscale systems: sampling strategies and embeddings. *SIAM J. Appl. Dyn. Syst.* **18**, 312–333 (2019).
33. Tam, Chi-Yung & Lau, Ngar-Cheung Modulation of the Madden–Julian oscillation by ENSO: inferences from observations and GCM simulations. *J. Meteorol. Soc. Jpn. Ser. II* **83**, 727–743 (2005).
34. Pohl, B. & Matthews, A. J. Observed changes in the lifetime and amplitude of the Madden–Julian oscillation associated with interannual enso sea surface temperature anomalies. *J. Clim.* **20**, 2659–2674 (2007).
35. Kikuchi, K. & Takayabu, Y. N. Equatorial circumnavigation of moisture signal associated with the Madden–Julian oscillation (MJO) during boreal winter. *J. Meteorol. Soc. Jpn. Ser. II* **81**, 851–869 (2003).
36. Seo, K.-H. & Kim, K.-Y. Propagation and initiation mechanisms of the Madden-Julian oscillation. *J. Geophys. Res. Atmos.* **108**, 4384 (2003).
37. Matthews, A. J. Primary and successive events in the Madden–Julian oscillation. *Q. J. R. Meteorol. Soc.* **134**, 439–453 (2008).
38. Zhang, F., Taraphdar, S. & Wang, S. The role of global circumnavigating mode in the MJO initiation and propagation. *J. Geophys. Res.: Atmos.* **122**, 5837–5856 (2017).

39. Wheeler, M. C. & Hendon, H. H. An all-season real-time multivariate MJO index: development of an index for monitoring and prediction. *Mon. Weather Rev.* **132**, 1917–1932 (2004).
40. Candès, E. J., Romberg, J. & Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**, 489–509 (2006).
41. Donoho, D. L. Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006).
42. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
43. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **58**, 267–288 (1996).
44. Zhang, L. & Schaeffer, H. On the convergence of the Sindy algorithm. *Multiscale Model. Simul.* **17**, 948–972 (2019).
45. Mangan, N. M., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **2**, 52–63 (2016).
46. Quade, M., Abel, M., Nathan Kutz, J. & Brunton, S. L. Sparse identification of nonlinear dynamics for rapid model recovery. *Chaos: Interdiscip. J. Nonlinear Sci.* **28**, 063116 (2018).
47. Kaiser, E., Kutz, J. N. & Brunton, S. L. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proc. R. Soc. A* **474**, 20180335 (2018).
48. Mangan, N. M., Askham, T., Brunton, S. L., Kutz, J. N. & Proctor, J. L. Model selection for hybrid dynamical systems via sparse regression. *Proc. R. Soc. A* **475**, 20180534 (2019).
49. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*(eds Parzen, E., Kitagawa, G. & Tanabe, K.) 199–213 (Springer, 1998).
50. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Mathe. Stat.* **22**, 79–86 (1951).
51. Anderson, D & Burnham, K. Model Selection and Multi-model Inference. *A Practical Information-Theoretic Approach*. 2nd, (Springer: New York, NY, 2002).

## ACKNOWLEDGEMENTS

N.D. thanks the Comisión Academica de Posgrado (CAP), Universidad de la República, Uruguay. All authors thanks the Programa de Desarrollo de la Ciencias Básicas (PEDECIBA).

## AUTHOR CONTRIBUTIONS

N.D. implemented the methodology, analysed the data and derived the models. M.B. and N.R. supervised the findings of this work. All authors contributed to the idea of the study, discussed the results and worked on the final manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Correspondence** and requests for materials should be addressed to Nicolás. Rubido.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.