Explaining the high skill of Reservoir Computing methods in El Niño prediction

Francesco Guardamagna[1,2], Claudia Wieners[1,2], and Henk A. Dijkstra[1,2]

[1]Institute for Marine and Atmospheric research Utrecht, Department of Physics, Utrecht University, Utrecht, the Netherlands
[2]Center for Complex Systems Studies, Utrecht University, Utrecht, the Netherlands

Correspondence: Francesco Guardamagna <f.guardamagna@uu.nl>

Abstract. Accurate prediction of the extreme phases of the El Niño Southern Oscillation (ENSO) is important to mitigate the socioeconomic impacts of this phenomenon. It has long been thought that prediction skill was limited to a 6 months lead time. However, Machine Learning methods have shown to have skill at lead times up to 21 months. In this paper we aim to explain for one class of such methods, i.e. Reservoir Computers (RCs), the origin of this high skill. Using a Conditional Nonlinear Optimal Perturbation (CNOP) approach, we compare the initial error propagation in a deterministic Zebiak-Cane (ZC) ENSO model and that in an RC trained on synthetic observations derived from a stochastic ZC model. Optimal initial perturbations at long lead times in the RC involve both sea surface temperature and thermocline anomalies which leads to a decreased error propagation compared to the ZC model, where mainly thermocline anomalies dominate the optimal initial perturbations. This reduced error propagation allows the RC to provide a higher skill at long lead times than the deterministic ZC model.

1 Introduction

The El Niño Southern Oscillation (ENSO) phenomenon, driven by ocean-atmosphere interactions in the tropical Pacific, is one of the biggest sources of interannual climate variability (Neelin et al., 1998). The full ENSO cycle shows an irregular period of 2-7 years. During its warm (El Niño) and cold (La Niña) phases, ENSO strongly affects the climate all over the globe through well-known teleconnections (McPhaden et al., 2006), increasing the incidence of extreme weather events like global droughts (Yin et al., 2022) and tropical cyclones (Wang et al., 2010). ENSO can therefore have a substantial impact on the worldwide economy (Liu et al., 2023a), and accurate and reliable forecasts are necessary to mitigate its socioeconomic consequences.

For this reason, ENSO modeling and forecasting have been a central topic of extensive research, which thanks to the contribution of the Tropical Ocean–Global Atmosphere program, led to the development of a complete hierarchy of models. This hierarchy includes conceptual models (Jin, 1997; Suarez and Schopf, 1988; Takahashi et al., 2019; Timmermann et al., 2003), intermediate complexity models (Zebiak and Cane, 1987; Battisti and Hirst, 1989) and Global Climate Models (Planton et al., 2021). Many of these classical dynamical models can reasonably forecast ENSO up to a lead time of 6 months, with a correlation between predictions and observations larger than 0.5 (Barnston et al., 2012), but their skill rapidly decreases for longer lead times.

In recent years, the application of Machine Learning (ML) techniques for predicting ENSO has significantly advanced (Bracco et al., 2024). Ham et al. (2019a) showed that Convolutional Neural Networks (CNNs) trained with CMIP5 and reanalysis data could obtain reasonable skill up to lead times of about 17 months. Hu et al. (2021) advanced the CNN approach by integrating dropout and transfer learning with a residual CNN, obtaining a good performance up to a lead time of 21 months. Long Short-Term Memory (LSTM) networks, able to exploit the temporal dynamics present in the training data, have also been successfully applied to ENSO forecasting (Xiaoqun et al., 2020). More recent studies have combined LSTM with other methods like Graph Neural Networks (Jonnalagadda and Hashemi, 2023), CNNs (Mahesh et al., 2019) and AutoEncoders (Jonnalagadda and Hashemi, 2023) to create hybrid models boosting the performance, as they are able to capture both the spatial and temporal dynamics present in the data. Reservoir Computer (RC) methods, a special class of Recurrent Neural Networks (RNNs), have shown optimal performance in predicting ENSO (Hassanibesheli et al., 2022). The RC offers a good balance between performance and model simplicity, which enhances explainability and facilitates analysis of model predictions. Moreover, like other RNN-based models, the RC offers the possibility of generating a self-evolving system that does not rely on external inputs (Guardamagna et al., 2024). This characteristic is crucial to understanding the internal dynamics of the RC and the evolution of errors over time during forecasting.

All these new tools provide more accurate forecasting skills than classical dynamical models, especially for longer lead times, and seem to be able to circumvent the "Spring Predictability Barrier" (SPB). The SPB (Webster and Yang, 1992; Lau and Yang, 1996) has been identified and documented across all the ENSO's dynamical models hierarchy from conceptual models (Jin and Liu, 2021a, b; Jin et al., 2021) to comprehensive GCMs (Duan and Wei, 2013). In particular, in the intermediate-complexity Zebiak and Cane (ZC) model (Zebiak and Cane, 1987), the SPB has been rigorously studied and quantified using the Conditional Nonlinear Optimal Perturbation (CNOP) framework (Mu et al., 2007). This tool has been applied to investigate the sensitivity of the ZC model to both initial conditions (Duan et al., 2013b) and model parameters (Yu et al., 2014) uncertainties. Thus, the ZC model is an excellent testbed to analyze why ML algorithms can have skill beyond the SPB, providing good predictions even when initialized during boreal spring.

In this paper, we aim to explain the good performance of RC methods in ENSO prediction. Specifically, we will compare the evolution of optimal initial perturbations, determined using the CNOP approach, between the RC (trained with synthetic observations from the stochastic ZC model) and the deterministic ZC model. In section 2, we shortly describe the ZC model and the CNOP technique, focusing on the changes introduced to adapt them to our analysis; in addition, the RC approach is briefly presented. In section 3 we first assess the performance of the RC, and then present results of the CNOP analysis for both the RC approach and the ZC model. A summary and discussion of the results follows in section 4.

2   Models and Methods

2.1   Zebiak and Cane (ZC) model

The ZC model is an intermediate complexity ENSO model that describes the evolution of anomalies with respect to a prescribed seasonal mean climatological state across the tropical Pacific. The state vector of this model consists of two-dimensional fields

of sea surface temperature, thermocline depth, oceanic and atmospheric velocities, and the atmospheric geopotential. For a complete description of the model's components and equations, we refer the reader to Zebiak and Cane (1987). We will use both the original deterministic ZC model and a stochastic ZC model following the approach described in Roulston and Neelin (2000). In this stochastic version, only noise in the zonal wind-stress field is applied as follows. First, a linear regression model relating SST anomalies and surface zonal wind-stress anomalies was constructed empirically from observations using the ORAS5 dataset (Copernicus Climate Change Service, 2021) over the period between 1961 and 1991 with a time step of 10 days (corresponding to the ZC model time step). Next, the variability explained by this linear model was subtracted from the total zonal wind-stress field to obtain the residual zonal wind-stress anomalies. The first EOF of this residual (Fig. A1) shows a strong component over the eastern Pacific. In Feng and Dijkstra (2017), the first two EOFs were included, where the second EOF captures the westerly wind bursts, but to keep the spatial noise structure simple, we only included the first EOF. Finally, the principal component (PC) related to the first EOF was fitted to a first-order autoregressive model:

$$x_{t+1} = ax_t + b\epsilon_t, \tag{1}$$

where ϵ_t is a white noise term following a Gaussian distribution with zero mean and unit variance ($\epsilon_t \sim N(0,1)$), while a and b are the fitted parameters. This fitted first-order autoregressive model was used during integration to generate a different (random) zonal wind-stress anomaly pattern at each time step.

There is still a debate on whether the Pacific climate state is in a ~~supercritical or subcritical~~ subcritical or supercritical regime (Kessler, 2002; Guardamagna et al., 2024). This distinction hinges on whether ENSO variability is a damped oscillation excited by stochastic forcing (subcritical) or occurs as a sustained oscillation or limit cycle (supercritical)~~or is a damped oscillation excited by stochastic forcing (subcritical). Hence~~. In the supercritical case, ENSO behavior is strongly influenced by nonlinearities, which arise from three main sources in the ZC model: heat advection, wind stress anomalies, and subsurface water temperature variations (Duan et al., 2013a). Given this ongoing debate, we study here both regimes, which can be easily distinguished in the ZC model by varying a single parameter. Following Tziperman et al. (1994), we use a parameter r_d in the drag coefficient $C_d = r_d C_d^0$, where C_d^0 is the standard value in the ZC model. Given the zonal and meridional wind velocities $\mathbf{u_a} = (u_a, v_a)$, the ZC model computes the wind stress (τ_x, τ_y) acting on the ocean surface according to the bulk formula:

$$(\tau_x, \tau_y) = \rho_{air} r_d C_d^0 |\mathbf{u_a}| (u_a, v_a), \tag{2}$$

where ρ_{air} is the air density, and $r_d = 1$ in the original model configuration (Zebiak and Cane, 1987). With increasing r_d, the ZC model generates a larger wind-stress response to sea surface temperature anomalies, intensifying the coupling strength between ocean and atmosphere.

In the deterministic version of the ZC model, an initial anomaly on the seasonal background state rapidly decays for $r_d = 0.79$ (Fig. B2a). In contrast, for $r_d = 0.8$, ENSO variability occurs as a periodic solution with a $\sim 4\,years$ period (Fig. B2b). Hence, the Hopf bifurcation bounding the two regimes is located between $r_d = 0.79$ and $r_d = 0.8$; here we choose $r_d = 0.77$ as a value in the subcritical regime and $r_d = 0.9$ in the supercritical regime. When noise is introduced, the ZC model's ENSO is phase-locked in the winter season (Fig. A3) for both $r_d = 0.77$ and $r_d = 0.9$. The SPB is identified with the initial month

90   corresponding to the fastest decrease in autocorrelation in eastern Pacific SST anomalies (Jin and Liu, 2021a). According to this definition, the ZC model shows a clear SPB in May for both $r_d = 0.77$ and $r_d = 0.9$ (Fig. A4). All these aspects make the ZC model a good testbed for understanding why the RC can circumvent the SPB, both in the subcritical and supercritical regime.

2.2   Reservoir Computer

95   Although the procedure to generate a RC has been well described elsewhere (Pathak et al., 2018), we briefly summarise the approach here, also introducing our notation. Given an input signal $u(n) \in R^{N_u}, n = 1, .., N_t$, where N_t is the total number of time steps and a given output signal $y^{target}(n) \in R^{N_y}$, the RC has to learn how to estimate an output signal $y(n) \in R^{N_y}$ as similar as possible to $y^{target}(n)$. To do that during the training procedure, an error measure $E(y, y^{target})$ is minimized, for which we choose a common measure for regression problems: the Mean Squared Error (MSE) defined by

$$100 \quad \mathrm{E}(y, y^{target}) = \frac{1}{N_y} \sum_{i=1}^{N_y} \left(\frac{1}{N_t} \sum_{n=1}^{N_t} (y_i(n) - y_i^{target}(n))^2 \right). \tag{3}$$

   Before the training procedure, the input data $u(n)$ are nonlinearly expanded into a higher dimensional so-called reservoir space, generating in this way a new signal $x(n) \in R^{N_x}$. This new representation of the data also contains temporal information and is based on the following update equations:

$$\tilde{x}(n) = \tanh \left(W^{in} u(n) + W x(n-1) \right), \tag{4a}$$

$$x(n) = (1 - \alpha) x(n-1) + \alpha \tilde{x}(n), \tag{4b}$$

where the hyperbolic tangent ($\tanh$) is applied component wise. Including a non-linear activation function such as $\tanh$ in the update equations enables the RC to estimate nonlinear relationships among the input variables, in contrast to less sophisticated models like the Linear Regressor, which can only capture linear relationships. This gives the RC an advantage in scenarios where nonlinearities play a significant role. The two matrices $\cancel{W_{in} \in R^{N_x \times N_u}} W^{in} \in R^{N_x \times N_u}$ and $W \in R^{N_x \times N_x}$ are gener-

105   ated randomly according to chosen hyperparameters. The non-zero elements of W and $\cancel{W_{in}} W^{in}$ are sampled from a uniform distribution over the range $[-a, a]$. The sparse matrix W derives from a random network with mean degree $< k >$, while $\cancel{W_{in}}$ W^{in} is a dense matrix. The quantity $\alpha \in (0, 1]$ in (4b) is the leaking rate. The output layer is defined as $y(n) = W^{out} x(n)$, where $W^{out} \in R^{N_y \times N_x}$, and during the training procedure only the weights of W^{out} are estimated by minimising $E(y, y^{target})$ through a linear regression procedure. We use a ridge regression to avoid overfitting, leading to the loss function $\mathcal{L}$:

$$110 \quad \mathcal{L}(W^{out}) = E(y, y^{target}) + \epsilon \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} (W_{i,j}^{out})^2. \tag{5}$$

The hyperparameters are given by the dimension of the reservoir (N_x), the spectral radius of the matrix W (ρ), the sparsity of W's connections $< k >$, the input scaling a and the leaking rate α. Given an input sequence $u(n) = y^{target}(n)$, the RC is trained by determining W^{out} from the sequence $y(n) = u(n+1) = y^{target}(n+1)$, using the loss function (5).

After training, the RC can be transformed into an autonomous evolving dynamical system to be used for prediction (Pathak et al., 2018). Thereto feedback connections between the outputs at time step n and the inputs at the subsequent time step are introduced. In this way, a model is generated that autonomously evolves in time, according to

$$x(n+1) = (1-\alpha)x(n) + \alpha \tanh(W x(n) + W^{in} u(n+1)), \tag{6a}$$

$$u(n+1) = y(n) = W^{out} x(n), \tag{6b}$$

where $x(n)$ and $x(n+1)$ are the reservoir states at time step n and $n+1$, while $y(n)$ is the output at time step n and $u(n+1)$ is the input at the subsequent time step $n+1$. This property of the RC allows us to make predictions similar to classical dynamical systems. Consequently, we can study how an initial perturbation evolves in the RC.

In the results below, the input vector u consists of the following feature variables: the NINO3 index, the thermocline depth anomalies h_W and h_E averaged over the regions 5°N-5°S ×120°E-180°E, and 5°N-5°S ×180°E-290°E, respectively and the zonal surface ~~wind speed~~ wind-speed anomalies τ_C averaged over the area 5°N-5°S × 145°E-190°E. ~~We chose to use zonal surface wind speed instead~~ Instead of directly using ~~the wind stress, as the noise in the Stochastic ZC model is applied directly to the~~ zonal surface wind-stress anomalies, zonal surface wind-speed anomalies have been used as a proxy. The two variables are inherently correlated through the bulk formula (2) and therefore convey similar information. However, a key distinction arises from how noise is introduced in the ZC model, specifically in the form of random zonal wind-stress anomalies. This leads to random local fluctuations in the zonal ~~wind stress (see Section 2.1). This strategy helps mitigate the risk of~~ wind-stress signal, which are inherently difficult for the RC to ~~overfit the noise present in the training data. In addition~~ predict and reproduce. In contrast, the surface wind-speed anomaly signal is smoother and more predictable, making it easier for the RC to learn and generalize efficiently.

In addition to the previous variables, a sine signal with a ~~period of 12 months representing~~ 12-month period was included to represent the seasonal cycle ~~was added such that $N_u = 5$,~~ such that $N_u = 5$. Although a combination of sine and cosine signals is required to identify uniquely each month of the year, we found that including both made little difference in performance. Therefore, to minimize the number of input variables and reduce the complexity of the learned function, we decided to use only the sine signal. The output vector consists of the same variables as in the input except for the sine signal, hence $N_y = 4$. In self-evolving mode, the sine signal encoding the seasonal cycle is provided as an external input rather than generated directly by the RC.

2.3 CNOP computation

Our implementation of the CNOP methodology follows the one described by Duan et al. (2013b). Let $M_{t_0,t}$ be the propagator of a nonlinear model from initial time t_0 to a chosen end time t_e. We indicate v_0 as the initial perturbation ~~super-imposed~~ superimposed on the model's background state V_0 at time t_0. For a selected norm $||.||$, an initial perturbation $v_{0\delta}$ is defined as

a CNOP if and only if:

$$J(v_0) = ||M_{t_0,t_e}(V_0 + v_0) - M_{t_0,t_e}(V_0)||, \tag{7a}$$

$$J(v_{0\delta}) = \max_{C(v_0) \leq \delta} J(v_0), \tag{7b}$$

where $C(v_0)$ is the constraint condition and $M_{t_0,t}(V_0)$ represents the model state at time t when the integration starts from the background state V_0 at time t_0. In Duan et al. (2013b), an initial perturbation is applied to all the grid points over the tropical area and the constraint condition to the initial perturbation amplitude $C(v_0)$ is defined as:

$$C(v_0) = \sqrt{\sum_{i,j}[(w_T^{-1}T'_{i,j})^2 + (w_h^{-1}h'_{i,j})^2]}, \tag{8}$$

140   where $T'_{i,j}$ and $h'_{i,j}$ are the initial sea surface temperature anomalies (SSTA) and thermocline depth anomalies, respectively, at grid point (i,j). The weights $w_T = 2°C$ and $w_h = 50$ m represent the characteristics scale of SST and thermocline depth anomalies, respectively.

As mentioned in section 2.2, the RC is trained using a limited feature vector. To ensure a fair comparison of CNOPs between those of the RC and the ZC models, the tropical area of the ZC model is divided into boxes and uniform perturbations are
145   applied over those boxes. Specifically, we apply a uniform SSTA perturbation T'_E over all the grid points in the NINO3 area (5°N-5°S × 210°E-270°E), a uniform thermocline depth perturbation h'_W to all the grid points in the area 5°N-5°S ×120°E-180°E and a uniform thermocline depth perturbation h'_E to all the grid points in the area 5°N-5°S × 180°E-290°E. The constraint condition can then be written as:

$$C(v_0) = \sqrt{(w_T^{-1}T'_E)^2 + (w_h^{-1}h'_E)^2 + (w_h^{-1}h'_W)^2}. \tag{9}$$

150   For both the RC and the ZC model, the objective function $J(v_0)$ in (7) has been defined as the Root Squared Error (RSE) between the perturbed and background trajectories. Specifically if we define the NINO3 index value at time t when the integration start from the initial state V_0 as NINO3(t, V_0), the objective function $J(v_0)$ is defined as:

$$J(v_0) = \sqrt{\sum_{t=t_0}^{t=t_N}(\text{NINO3}(t, (V_0 + v_0)) - \text{NINO3}(t, V_0))^2}, \tag{10}$$

where $t_N = t_e$. To solve the optimization problem associated with determining the CNOP, we use the gradient-free Cobyla
155   optimization algorithm (Powell, 1994). Since the Cobyla algorithm starts its optimization process from a random initial guess, we always perform 10 different realizations starting from 10 different initial guesses to select the CNOPs that shows the largest error propagation according to the value of $J(v_0)$; a detailed description of the COBYLA algorithm is reported in appendix B.

3   Results

In the results section, we will first explain the training and validation of the RC (section 3.1), demonstrate the forecasting skills
160   of the RC (section 3.2), also demonstrating the importance of the zonal surface wind velocity anomalies as a training variable. Next, in subsection 3.3, we present the results of the CNOP analysis for both the RC and deterministic ZC models.

3.1 Training and Validation of the RC

For both subcritical ($r_d = 0.77$) and supercritical ($r_d = 0.9$) regimes we first performed a simulation of 1000 years with the stochastic ZC model using a time step of 10 days. We will refer to these data as 'synthetic observations'. The NINO3 amplitudes of the supercritical case (Fig. 1b) are, as expected, about a factor 2 larger than those of the subcritical case (Fig. 1a). As mentioned in section 2.2., the 12 months period sine signal and the feature vector components h_W, h_E, τ_C, and NINO3 (extracted from the synthetic observation time series) are used to train the RC. To investigate the effect of τ_C on the performance of the RC, we also trained a second RC using only h_W, h_E, NINO3 and the sine signal. Before training the NINO3 and both h_W and h_E have been normalized by $w_T = 2°C$ and $w_h = 50m$, respectively. From the total 1000 years of synthetic observations, the first 300 years were discarded to avoid capturing any initial transient behavior. The next 500 years were used for training and validation (300 years for training and 200 years for validation), and the last 200 years were used for testing, ensuring an independent evaluation of the RC model performance. The training of the RC was described in section 2.2, where given an input sequence $u(n) = y^{target}(n)$, W^{out} is determined from the sequence $y(n) = u(n+1) = y^{target}(n+1)$, using the loss function (5).

To determine the performance of the RC, we use the RC in self-evolving mode (section 2.2) to make predictions using a time step of 10 days. When we let the RC self evolve, the only external information we provide is the value of the sine signal representing the current month of the year. All the other variables (NINO3, h_E, h_W, and τ_C when the latter is included as a training variable) are directly produced by the output of the RC and are not provided as external information during prediction. To evaluate the RC's performance over the entire 200 years of validation trajectories for different lead times, we adopt a rolling approach. For each time step ~~t~~ $t(n)$ in the validation data set, a RC trajectory with a specific lead time ~~was~~ is generated. The ~~initial reservoir state for each prediction was determined using 5-years of data prior to the time t (i.e. 180 steps). The~~ final values of each trajectory, corresponding to the lead time of interest, ~~were~~ are then concatenated to form a complete 200-years trajectory, say y_{full}. Before performing inference, we always determine the internal RC state using 5 years of data prior to the time step $t(n)$. Discarding the initial $x(n)$ RC states for $0 \leq n \leq n_{transient}$ is a common practice in Reservoir Computing. This step is necessary to mitigate the impact of initial transients caused by the arbitrary initialization of the reservoir state, typically set to $x(0) = 0$, or randomly initialized. In our case, we set $x(0) = 0$. This initialization creates an artificial starting state that is unlikely to recur once the reservoir dynamics stabilize. A warm-up period is, therefore, necessary to allow the RC to reach a stable dynamical regime. The length of this warm-up period depends on the RC's memory capacity and the specific learning task. Based on our experiments, we found that a 5-year warm-up period is sufficient to stabilize the reservoir dynamics and eliminate the effects of initial transients. As a result, the reservoir states corresponding to the 5 years of data preceding the time step $t(n)$ are used to initialize the RC internal state and then discarded. In our notation, this means discarding the $x(n)$ reservoir states for $t(n) - n_{transient} \leq n \leq t(n)$, where $n_{transient} = 180$, given our 10-day time step.

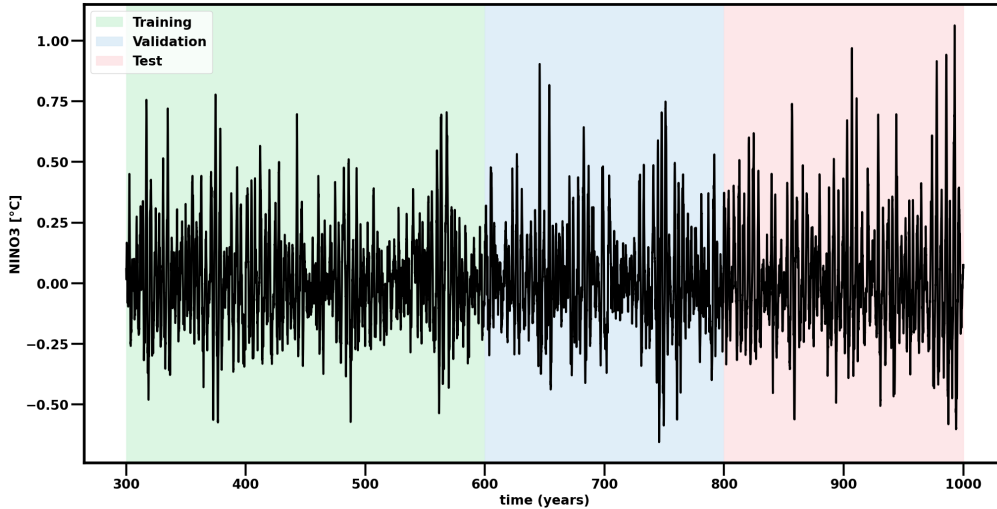
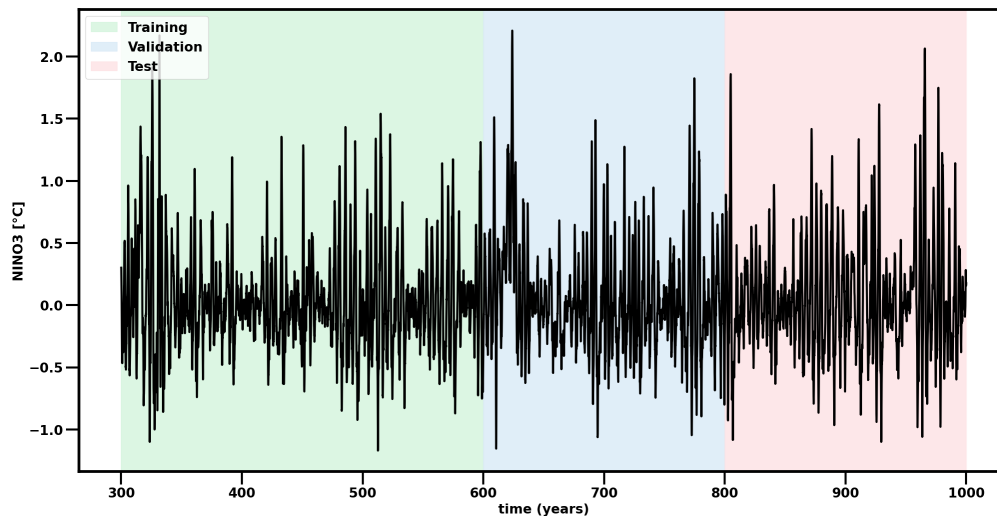
To identify the best set of hyperparameters, a separate validation procedure was conducted for each regime ($r_d = 0.77$ and $r_d = 0.9$) and for each set of training variables (including and excluding the zonal surface wind speed anomalies) using a Bayesian search. For each hyperparameter set, the RC model's 18-month lead time predictions were evaluated using

(a)

(b)

Figure 1. NINO3 index from the last 700 years of the stochastic ZC model simulations (synthetic observations) used to train, validate, and test the RC model: (a) $r_d = 0.77$, (b) $r_d = 0.9$.

the root mean square (RMS) error computed including all feature variables in y_{full}. This latter was done to ensure that the RC model could replicate the synthetic observations for all variables of interest rather than simply replicating the NINO3 index. Among all the different hyperparameters, the reservoir dimension N_x is one of the most significant for the RC's performance (Lukoševičius (2012); Verstraeten et al. (2007)). Increasing N_x expands the reservoir's state space, allowing for a richer and more complex high-dimensional representation of the input signal $u(t)$ (see Section 2.2). Additionally, larger N_x values increase the reservoir's memory capacity.

200

During our experiments, the Bayesian search has consistently converged on large N_x values, with a notable difference between the supercritical and subcritical regimes. In the supercritical regime, the optimal reservoir dimension is approximately 400, regardless of whether the variable τ_C is included during training. In the subcritical regime, the optimal dimension is larger, with $N_x = 476$ when τ_C is included and $N_x = 534$ when it is excluded. Table C1 in appendix C reports the optimal N_x values identified for each regime and input-variables configuration, as well as the optimal values for all other RC's hyperparameters (see Section 2.2).

After validation, we evaluated the RC model~~'s performance was evaluated on the 200-years~~'s performance on the 200-year test set using ~~the identified best hyperparameter set~~these best hyperparameter sets, as described next.

3.2 RC performances

Figure 2 presents the mean and standard deviation of the Anomaly Correlation Coefficient (ACC) of 50 different RC's prediction trajectories and the target NINO3 index from the 200-years test dataset, computed at a monthly time step (so averaged over three model time steps). We evaluated the RC's ability to replicate the monthly NINO3 index rather than the 10-day time step index used for training, since this is the common approach for assessing the performance of ENSO forecasting models. As the reservoir is generated by random W and W^{in} values, each RC needs to be retrained first (using the 300 years data set) as described in section 2.2 and hence multiple RCs are used for evaluating the ACC. Again a rolling approach (as for the validation data set, see (section 3.1) was used for the test set and hence the ACC is determined using the 200-years vector y_{full}.

In the supercritical regime (Fig. 2b), the RC model performs better when zonal surface ~~wind speed~~ wind-speed anomalies τ_C are included as a training variable, though its performance is also acceptable even when τ_C is excluded. On the other hand, in the subcritical regime (Fig. 2a), the RC performance for longer lead times (9 to 18 months) improves when τ_C is excluded during training. ~~This is due to the fact that the RC model is overfitting the noisein the subcritical regime when τ_C is included in the feature vector. In the stochastic ZC model, the noise is introduced as~~ In this regime ENSO is primarily driven by atmospheric noise, introduced in the Zebiak and Cane model in the form of random zonal wind-stress burst (see section 2.1)~~, and~~. When the model is initialized from ENSO neutral conditions, optimal atmospheric noise patterns can trigger transient growth of perturbations, provided the initial conditions are favorable. Conversely, if a perturbation is already developing, subsequent noise patterns can either reinforce or damp its evolution. The variable τ_C is therefore particularly useful for predicting the short-term variability of ENSO, as it provides critical information about external forcing that influences early perturbation dynamics. Accordingly in the subcritical regime the ~~oscillation is purely noisedriven. Although we use zonal wind speed instead of directly use the wind stress to reduce noise sensitivity, the RC still tends to overfit the noise, reducing its predictive performance. The~~ RC achieves better performances at shorter lead times (3-6 months), when τ_C is included. At longer lead times (9-18 months), improved predictive performances requires the RC to rely more on system internal dynamics rather than the short-term influence of stochastic noise. Including τ_C during training can lead to overfitting, causing the model to focus excessively on short-term noise patterns instead of learning the internal system dynamics. In the supercritical regime, nonlinearities play an essential role (see Section 2.1). In the ZC model, these nonlinearities arise from three main sources: heat advection, wind stress anomalies, and water temperature variations (Duan et al. (2013a)), all of which influence the evolution

of ENSO. This characteristic is reflected in the RC model's performance, which exhibits improved predictive skill at both short and long lead times when τ_C is included during training, underscoring the importance of the nonlinear effects introduced by this variable in this regime.

Overall, the RC performs better in the supercritical regime, achieving an ACC of 0.8 at a 12-month lead time when zonal
240 surface wind speed anomalies ~~were~~ are included during training. In ~~contrast, in~~ the subcritical regime, the RC model achieves an ACC of 0.75 at a 12-month lead time when τ_C is excluded during training.

To better appreciate the performance of the RC model, we also compared it with a simple Linear Regressor as a benchmark; results are also included in Fig. 2. ~~In the supercritical regime the Linear Regressor also performs worse when τ_C is~~
~~excluded during training, further demonstrating the importance of this variable. In contrast,~~ The comparison between the LR
245 and RC reveals that the performance improvement achieved by adopting the RC model is not drastic. However, the results still demonstrate a clear and consistent advantage in using the RC, both in the subcritical ~~regime the Linear Regressor performs~~
~~similar whether τ_C is included or excluded during training, reinforcing the idea that this variable is not essential in this regime.~~
~~The RC outperforms the Linear Regressor both~~ and supercritical regimes, and regardless of whether the variable τ_C is included during training. This improvement stems from the nonlinear activation function used in the RC (see section 2.2), which enables
250 it to capture nonlinear relationships between input variables something the LR model, limited to linear approximations, cannot achieve. This advantage is particularly clear in the supercritical ~~and subcritical regimes, whether τ_C is included or excluded~~
~~during training~~regime, where the RC model provides a more significant performance increase compared to the subcritical regime. This is ~~due to the capability of the RC to capture the nonlinear behavior in~~ expected, as nonlinearities play a more prominent role in the supercritical regime (see Section 2.1). Moreover, the relatively small performance gap between the LR
255 and the RC can be attributed to the ~~dynamics of the stochastic ZC model~~ZC model being a model of intermediate complexity in which ENSO is a weakly nonlinear phenomena (e.g. all wave dynamics in ocean and atmosphere is linear in the model). The ZC model's data exhibit simpler dynamics than real-world observations or simulations with more complex General Circulation Models (GCMs). In such cases, the performance advantage of the RC over the LR is expected to be more pronounced.

The ability of the RC model to mitigate the SPB is demonstrated in Fig. 3. This figure presents the normalized mean
260 absolute error (MAE) between the median NINO3 of 50 different RC's predictions and the corresponding target values from the synthetic observations test dataset (see section 3.1) at various lead times and for both the RC initialized before the SPB in March, April, and May, and after the SPB in September, October, and November. As a comparison benchmark, the normalized MAE for the Linear Regressor (LR) predictions is also included for the same initialization months. Additionally, to ensure a fair comparison between the subcritical and supercritical regimes, all RC and LR predictions and the corresponding target values
265 have been normalized by the standard deviation of the 200-years synthetic observations test dataset (0.47 for the supercritical regime and 0.24 for the subcritical regime) before computing the MAE. In Fig. 3, we present results for the different input variable configurations for both the subcritical and supercritical cases. Specifically, the variable τ_C is excluded from the input variables in the subcritical regime but included in the input variables in the supercritical regime.

In both the subcritical and supercritical regimes, the RC model outperforms the LR also in terms of mean absolute error,
270 regardless of the initialization period. However, to a certain extent, it is still affected by the SPB, which occurs in May in the ZC

(a) subcritical regime   (b) supercritical regime

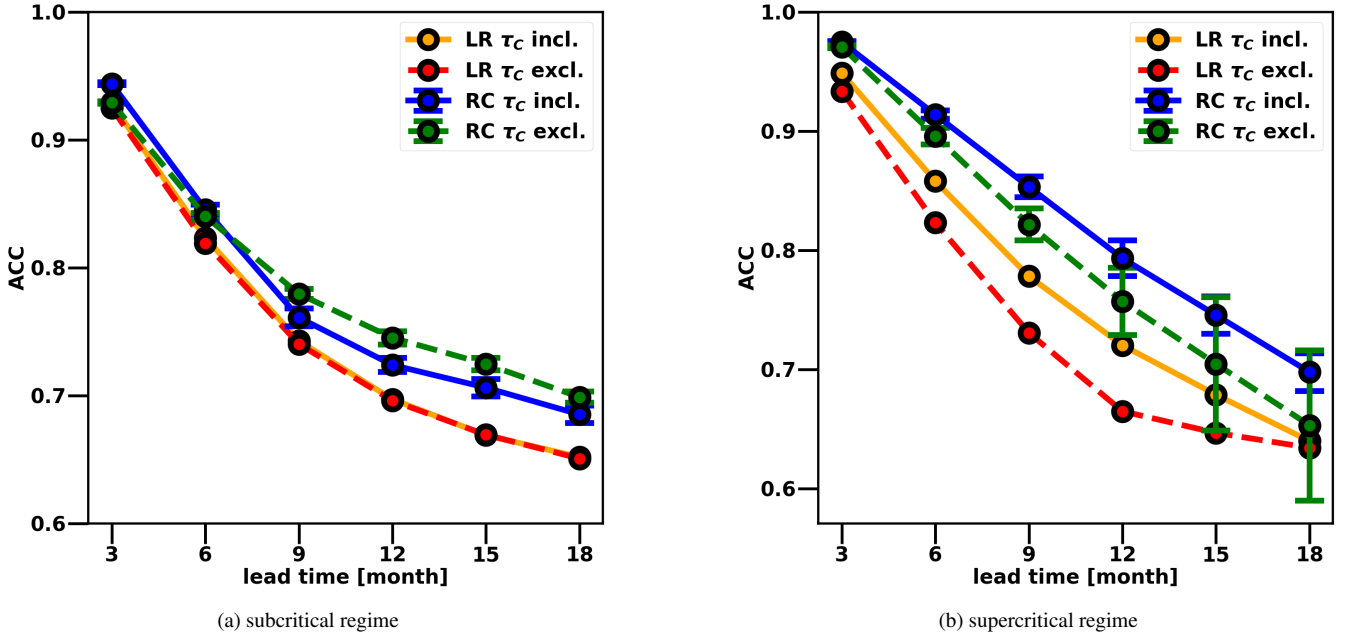



Figure 2. Mean and standard deviation of the Anomaly Correlation Coefficient (ACC) of 50 different RC model realizations and the 200-years synthetic observations for the NINO3 index, computed at a monthly time step. Results are shown for the 2 regimes: (a) subcritical ($r_d = 0.77$) and (b) supercritical ($r_d = 0.9$), with zonal surface wind speed anomalies (τ_C) either included or excluded during training. Results from the Linear Regressor are also included for comparison

model (as discussed in Section 2.1). On the other hand, the RC model demonstrates a clear ability to mitigate the effects of the SPB compared to the LR. This can be most clearly seen when comparing the pre-spring initialization performance of the two models at 3-months and 6-months lead times. In the supercritical regime, with pre-SPB initialization, the RC model achieves a normalized MAE of 0.2 at 3-months lead time and 0.35 at 6-months lead time, while the LR shows a higher normalized

275   MAE of 0.3 at 3-months lead time and 0.5 at 6-months lead time. In the subcritical regime, with pre-SPB initialization, the RC achieves a MAE of 0.34 at 3-months lead time and 0.5 at 6-months lead time, while the LR shows a MAE of 0.36 at 3-months lead time and 0.54 at 6-months lead time. In the supercritical regime, the RC shows a larger performance improvement, compared to the subcritical regime where the difference in performance with respect to the LR is less evident. Moreover, also in terms of normalized MAE, the RC performs better in the supercritical regime than in the subcritical one. ~~This result aligns~~

280   ~~with expectations, as non-linearities play a more important role in the supercritical regime.~~

(a) subcritical regime   (b) supercritical regime

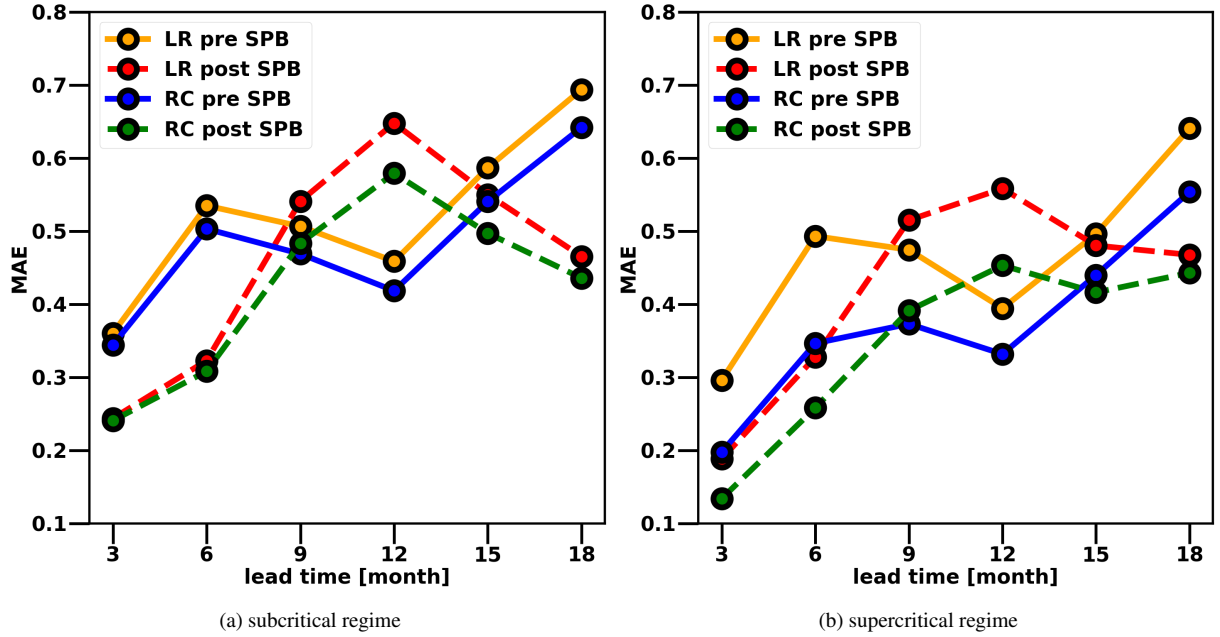



Figure 3. Normalized mean absolute error (MAE) between the median of 50 different RC realizations' predictions and the 200-years synthetic observation test set, for the NINO3 index computed at a monthly time step, and with the RC initialized both before (March, April, May) and after (September, October, November) the SPB: (a) $r_d = 0.77$, zonal surface wind speed anomalies excluded during training. (b) $r_d = 0.9$, zonal surface wind speed anomalies included during training. Results from the Linear Regressor (LR) are also included for comparison. Both predictions and target values have been normalized by the standard deviation of the 200 years synthetic observations test dataset (0.47 for the supercritical regime and 0.24 for the subcritical regime), before computing the MAE.

3.3 CNOP analysis

For both the RC model and the deterministic ZC model and for both $r_d = 0.77$ (subcritical regime) and $r_d = 0.9$ (supercritical regime), we computed the CNOPs for different lead times using the last 50 years of the 200-years synthetic observations test dataset as initial conditions (cf. section 3.1). This choice has been made to balance computational efficiency and statistical significance. The CNOP computations using the Cobyla algorithm are highly computationally expensive, and 50 years of data is sufficient to obtain statistically significant results. By selecting the last 50 years of the 200-year test period, we also ensure complete statistical independence between the training and test data. For the RC model, perturbations were directly applied to NINO3 and mean thermocline depth anomalies (h_E and h_W). In contrast, for the deterministic ZC model, a uniform perturbation was applied over three different boxes in the Pacific for both SSTA and thermocline depth anomalies (as described in section 2.3).

(a) subcritical regime   (b) supercritical regime

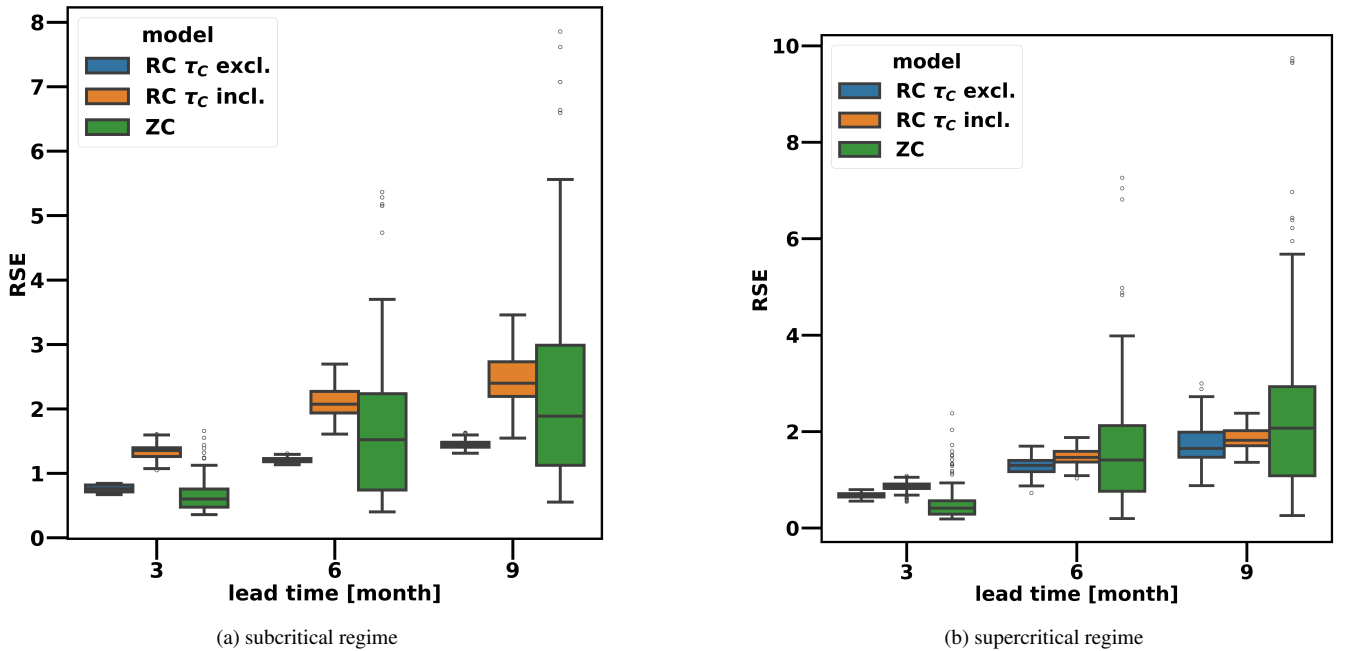



Figure 4. Distribution of the normalized RSE distances between the perturbed and unperturbed trajectories for different lead times when CNOPs are applied, taking as initial conditions the months [March,April,May] for (a) $r_d = 0.77$ and (b) $r_d = 0.9$. The boxes indicate the interquartile range (IQR), the range within which the central 50% of data points are located. The whiskers extend to the minimum and maximum values within 1.5 times the IQR from the first and third quartile. The central line corresponds to the median. All the RSE distances have been normalized by the standard deviation of the NINO3 index extracted from the 50 years of synthetic observations considered for the CNOP computation (0.29 for the subcritical regime and 0.56 for the supercritical regime).

Before computing the CNOPs for the RC model, we identified and saved the best-performing RC realization out of 50 for each combination of lead time, r_d value, and training variables set based on the forecasting skill for the 200-years synthetic observations test period (see section 3.1). The top-performing realization (for each combination of lead time, r_d value and

training variables) was then considered for the CNOP computation. This was done to avoid biases related to the random
initialization of the RC. We computed the CNOPs for lead times 3, 6, and 9 months (the optimization time considered during the CNOPs computation), focusing on a single constraint value $\delta = 0.05$ and a specific forecast initialization season just before the SPB, encompassing March, April, and May. The value of δ corresponds to a maximum NINO3 perturbation of 0.1°C (section 2.3) or a maximum h_e or h_w perturbation of 2.5m. For the deterministic ZC model, the same procedure to compute the CNOP was used (see section 2.3). To quantify the divergence between two trajectories caused by the CNOPs, for both the RC and the ZC models, we computed the Root Square Error (RSE) distance between the perturbed and unperturbed ~~trajectories~~ NINO3 trajectories, as defined in ~~(10). For the deterministic ZC model, the same procedure to compute the CNOP was used (see section 2.3)~~[10]. This distance was used to estimate each model's sensitivity to initial condition perturbations, given a specific initial state. To make a fair comparison between the subcritical and supercritical regimes, all the RSE distances obtained have been normalized by the standard deviation of the 50 years of NINO3 synthetic observations considered for the CNOPs computation (0.29 for the subcritical regime and 0.56 for the supercritical regime).

In the supercritical regime (Fig. 4b), the RC model is more susceptible to initial perturbations at shorter lead times. However, at a 6-months lead time, the RC model's sensitivity to initial perturbations becomes, on average, smaller when τ_C is excluded during training and similar to that of the deterministic ZC model when τ_C is included during training (see Table D2). At 9-months lead time, the RC's sensitivity to initial perturbations is on average smaller for both τ_C included and excluded. At both 6- and 9-months lead times, the deterministic ZC model's sensitivity results show a much wider distribution than the RC, regardless of whether τ_C is included or excluded as a training variable. In the subcritical regime (Fig. 4a), the RC model becomes more susceptible to perturbations than the deterministic ZC model when τ_C is included as a training variable. Conversely, when this variable is excluded, the RC model shows less sensitivity to perturbations than the deterministic ZC model. This difference is likely because including τ_C as a training variable causes the RC model to learn more the noise component of synthetic observations. Since ENSO variability in the subcritical regime is highly affected by noise, including these anomalies during training leads to a system with a larger error propagation.

Previous studies (Mu et al., 2007) have quantified the SPB in the deterministic ZC model using the CNOP framework, revealing that the deterministic ZC model is particularly sensitive to initial perturbations when initialized just before the boreal spring season. Our results support this finding, showing that the deterministic ZC model exhibits a stronger sensitivity to initial condition perturbations when initialized close to the SPB than when it is initialized later in the year (see Table D1). This also holds for summer initialization (not shown) in June, July, and August, where the models show results similar to spring initialization in March, April, and May, with the RC mitigating sensitivity to initial perturbations similarly. This behavior is due to the proximity of the summer season to the SPB. The CNOP cost function evaluates the distance between the entire perturbed and unperturbed trajectories (see 10), taking all months into account. When the deterministic ZC model integration is initialized just before the SPB, the number of months affected by the SPB is maximized, and at longer lead times (6 and 9 months), we observe a pronounced increase in the sensitivity to initial conditions perturbations compared to when the model is initialized in the autumn and winter seasons. This effect is also found when comparing the sensitivities of autumn and winter initializations. Compared to an autumn-initialized trajectory, an integration initialized in winter has crossed the SPB before at

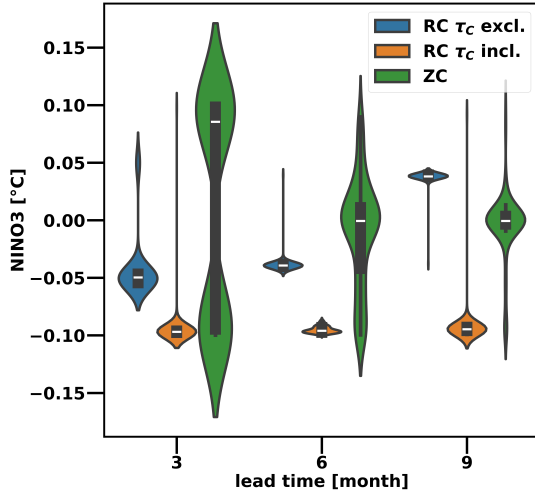
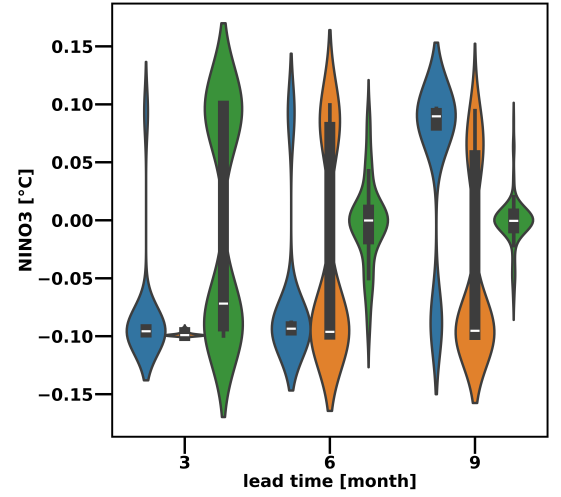
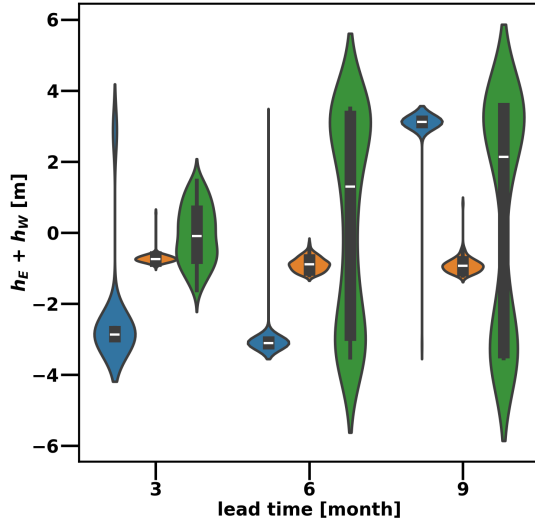
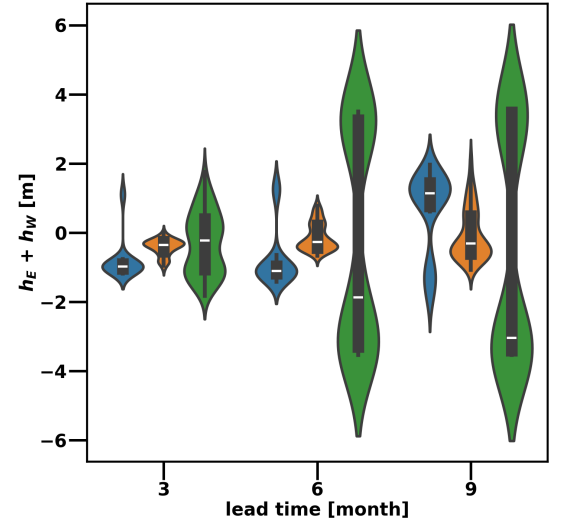



Figure 5. Violin plots showing the distribution of the CNOPs obtained for both the NINO3 Index and $h_E + h_W$ (sum of the thermocline anomalies for both the western and eastern Pacific). (a)-(c) $r_d = 0.77$ (b)-(d) $r_d = 0.9$. In both cases, $\delta = 0.05$, the period considered corresponds to the last 50 years of the 200 years synthetic observation test dataset, and the months [Mar, Apr, May] are taken as initial conditions.

a 9-months lead time and the sensitivity to initial perturbations for an integration initialized in winter is, on average, larger than for the autumn initialization.

On the other hand, when the RC is initialized later than the SPB, it exhibits a sensitivity to initial perturbations similar to that found when it is initialized just close to the SPB (see Table D2). As the number of months affected by the SPB increases (at 6- and 9-months lead time, with a forecast initialized in spring), the RC effectively reduces both the average sensitivity to initial condition perturbations and the width of sensitivity results' distribution compared to the ZC model, consequently

335   decreasing the number of events strongly sensitive to initial conditions perturbations. The only exception is the RC trained, including the variable τ_C in the subcritical regime, which consistently has a greater sensitivity to initial condition perturbations than the deterministic ZC model, for reasons already mentioned above. Moreover the inclusion of this variable decreases the performance of the RC in the subcritical regime at longer lead times (see section 3.2). These results demonstrate that the RC model effectively mitigates the sensitivity to initial condition perturbations at long lead times (6, 9 months) when a forecast is

340   initialized just before the SPB, compared to the ZC model, for which the spring season corresponds to the strongest sensitivity to initial perturbations. This capability explains why the RC model can reduce the effects of the SPB, delivering skillful predictions at long lead times.

(a)   (b)

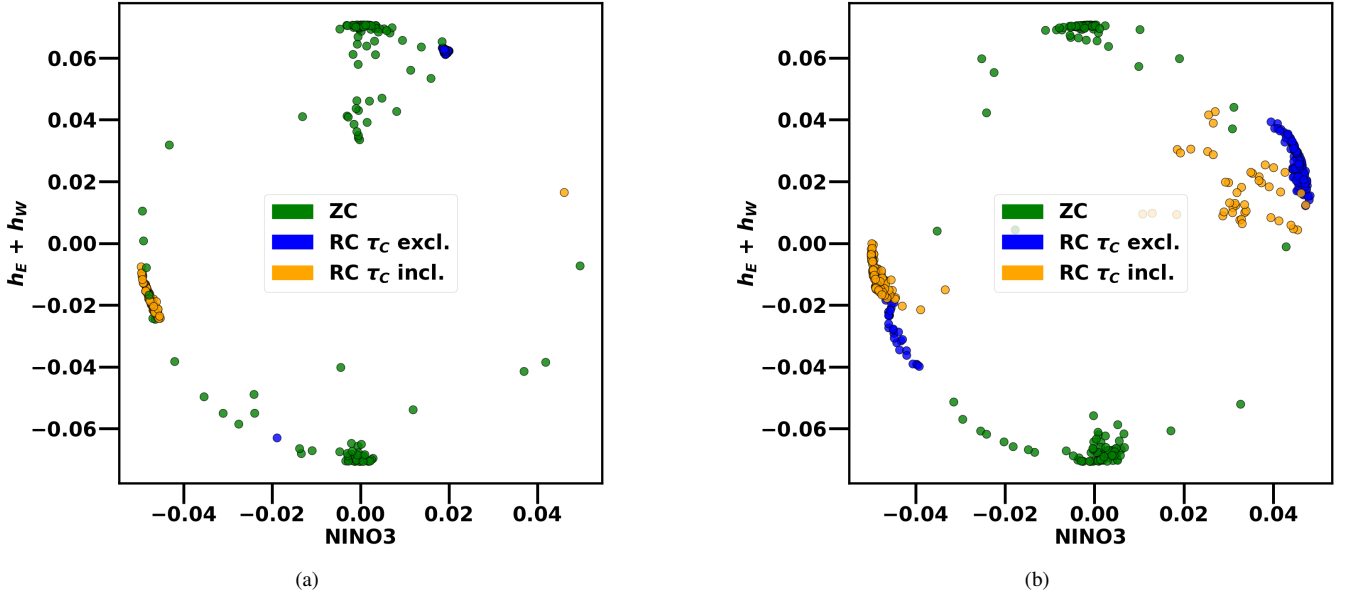



Figure 6. Scatter plot of the CNOPs in the normalized NINO3 index, $h_E + h_W$ anomaly plane. (a) $r_d = 0.77$ (b) $r_d = 0.9$. In both cases, $\delta = 0.05$, the period considered corresponds to the last 50 years of the 200-years synthetic observation test dataset, the months [Mar, Apr, May] are taken as initial conditions and the lead time considered is 9 months. The NINO3 index and $h_E + h_W$ anomalies have been normalized dividing by 2°C and 50m, respectively.

Figure 5 shows the estimated CNOPs for both the ZC and RC models when initialized just before the SPB in March, April, and May. The estimated CNOPs are presented for the NINO3 index and the sum of the thermocline perturbations in the eastern

345   and western Pacific ($h_E + h_W$). The ZC model's sensitivity to initial NINO3 perturbations decreases as the forecasting lead time increases. In contrast, perturbations to the thermocline depth become increasingly crucial for optimal perturbation growth

at longer lead times. This is true for both $r_d = 0.77$ and $r_d = 0.9$. On the other hand, the CNOPs of the RC have different behavior, both for the supercritical and subcritical regimes. The RC is sensitive to quite different initial perturbations leading to less variability in the error propagation compared to the ZC model. This is supported by Fig. 6, which shows the distribution of the CNOPs in the (NINO3, h_E+h_W) plane for a 9 months lead time. For visualization purposes, the initial anomalies for NINO3 and (h_E+h_W) have been normalized dividing by $2°C$ and 50m, respectively (see section 2.3). The CNOPs for the RC and the ZC models show a strongly different distribution~~, with perturbations in~~ . The RC model consistently exhibits greater sensitivity to NINO3 ~~being more dominant at longer lead timesfor the RC model than for the deterministic ZC model~~ perturbations at both short (3 months) and long (6 and 9 months) lead times, while the ZC model shows increasing sensitivity to initial thermocline depth perturbations as the optimization time extend. In the ZC model, ENSO variability is ~~highly affected by the~~ strongly influenced by thermocline feedback (~~Zebiak, 1987~~Zebiak and Cane (1987)). The RC reduces this sensitivity ~~and consequently reduces the error propagation . It is worth noticing that~~ at longer lead times helping to mitigate error propagation over time.

Another notable characteristic, visible in Figure 5, is that the ZC model exhibits a highly symmetrical distribution for the h_E+h_W CNOPs at a 9-month lead time, as well as for the NINO3 CNOPs at a 3-month lead time. In contrast, the RC model shows a more biased distribution for the estimated NINO3 CNOPs, particularly in the subcritical regime and at shorter lead times (3 months) in the supercritical regime. At longer lead times (9 months) in the supercritical regime, the RC model still shows a relatively symmetrical distribution for the estimated NINO3 CNOPs.

To better understand the origin of these differences, we classify the initial conditions analyzed (before applying the CNOPs), spanning the months [March, April, May] into three groups based on the initial eastern Pacific sea surface temperature anomalies (SSTA). Specifically, in the supercritical (subcritical) regime, we define an initial state as *positive* if the corresponding initial NINO3 index is larger than 0.2 (0.1)°C, ~~when zonal surface wind speed anomalies are not~~ *negative* if the initial NINO3 index is smaller than -0.2 $(-0.1)°C$, and *neutral* if -0.2 $(-0.1)°C \leq$ NINO3 ≤ 0.2 $(0.1)°C$. The initial states have been classified in a different way for the supercritical and subcritical regime to account for the fact in the subcritical regime the amplitudes of NINO3 is a factor 2 smaller compared to the supercritical regime see section 3.1.

As shown in Figures D2c and D2d in Appendix D, when the ZC model's SSTA initial condition in the eastern Pacific is characterized by a positive (negative) anomaly, the optimal initial perturbation for h_E+h_W for a 9-month optimization period is negative (positive), if the model is initialized before the event's peak. This limits the propagation of the anomaly, leading to a weaker El Niño (La Niña) event if the model is initialized prior to the event's peak. Conversely, if the model is initialized after the event's peak, the negative (positive) thermocline perturbation drives a faster and steeper return to neutral conditions.

When the ZC model is initialized from neutral conditions, the sign of the estimated CNOPs for h_E+h_W depends on the reference trajectory. If the system would have transitioned into an El Niño event, the optimal h_E+h_W is negative, thus dampening the positive anomaly; if it would have transitioned into a La Niña event, the optimal h_E+h_W is positive, suppressing the negative anomaly.

The same reasoning applies to the 3-month lead-time NINO3 optimal perturbations: a negative (positive) NINO3 perturbation weakens the near-term growth of an initially positive (negative) eastern Pacific SSTA.

Regarding the RC model, as shown in Figures D3c and D3d, the estimated CNOPs for the NINO3 index at a 9-month lead time in the supercritical regime exhibit a fairly symmetrical distribution overall. However, significant differences emerge depending on whether the variable τ_C is included during training, ~~the RC shows a weaker sensitivity to initial~~ as well as in comparison with the ZC model. When τ_C is included, and the RC is initialized before the peak of an event, the optimal NINO3 perturbation is positive (negative) for initial conditions characterized by positive (negative) eastern Pacific SSTA, thereby reinforcing the initial anomaly and leading to larger El Niño (La Niña) events. In contrast, when the RC is initialized after the peak of a positive (negative) event, the optimal NINO3 perturbation reverses sign, resulting in a faster and steeper return to neutral conditions.

Neutral initial conditions tend to prefer negative NINO3 perturbations, resulting in stronger La Niña and weaker El Niño events relative to the reference trajectories. When τ_C is not included during training, positive SSTA initial conditions still favor positive optimal NINO3 perturbations~~at longer lead times compared to both the supercritical case and the subcritical case when zonal surface wind speed anomalies are included as~~, yielding stronger El Niño events when the model is initialized before the peak. In contrast, for events characterized by negative initial eastern Pacific SSTA, the optimal NINO3 perturbations are positive, thereby mitigating the development of the initial negative anomaly over time. Moreover, neutral initial conditions exhibit a stronger tendency toward positive optimal NINO3 perturbations in the absence of τ_C.

In the subcritical regime, the optimal NINO3 perturbation at a ~~training variable. However, it~~ 9-month lead time exhibits a marked asymmetry. As illustrated in Fig. D3a and D3b, including τ_C during training leads to consistently negative optimal perturbations across all three categories of initial conditions, whereas excluding τ_C results in positive optimal perturbations for every category. At shorter lead times (3 months), in both the subcritical and supercritical regimes and regardless of whether τ_C is included during training, the RC model clearly prefers negative optimal NINO3 perturbations. These results highlight the influence of τ_C on the estimated CNOPs, particularly in the subcritical regime. Further evidence of τ_C's impact in both supercritical and subcritical regimes appears when examining another result shown in Fig. 5, where we find that excluding zonal surface wind-speed anomalies during training in the subcritical regime causes the RC model to exhibit greater sensitivity to thermocline depth perturbations at longer lead times compared to the supercritical regime and to the subcritical regime when τ_C is included during training. However, even in this scenario, the RC remains less sensitive to thermocline depth anomalies than the ZC model. In the subcritical regime, ENSO variability is primarily driven by atmospheric noise, introduced as stochastic wind stress forcing. This noise affects the thermocline slope, activating mechanisms that lead to the development of perturbations. When the variable τ_C is included during training, the RC explicitly learns the relationship between wind-stress anomalies and thermocline adjustments, and the state of the surface winds is provided as an initial condition. Consequently, smaller thermocline perturbations can be amplified by wind-stress anomalies, resulting in larger deviations from the unperturbed trajectory. As a result, the optimal initial perturbations primarily target the NINO3 index, favoring negative NINO3 perturbations, while the optimal $h_E + h_W$ perturbations are only slightly negative and closer to zero, following the same trend as the NINO3 CNOPs. The impact of these optimal perturbations varies based on the initial conditions. When the RC is initialized with positive or slightly positive (neutral) eastern Pacific SSTA, negative NINO3 and $h_E + h_W$ perturbations suppress or weaken the evolution of positive initial anomalies, resulting in a faster return to neutral

conditions. Conversely, when the RC is initialized with negative or slightly negative (neutral) eastern Pacific SSTA, negative NINO3 perturbations amplify the evolution of negative anomalies, leading to either a stronger La Niña event or a longer persistence of La Niña conditions if the RC is initialized after the peak.

420 When τ_C is not included, the RC learns only the direct relationship between NINO3 and thermocline depth anomalies, without explicit knowledge of how wind anomalies influence thermocline slope adjustments. As a result, in the absence of wind-forcing information, larger initial thermocline perturbations are required to induce significant error growth over time. Under these conditions, the optimal initial perturbations consist of both positive NINO3 and positive $h_E + h_W$ perturbations.

The effect of these optimal perturbations again depends on the initial conditions. When the RC is initialized with positive or 425 slightly positive eastern Pacific SSTA, the combination of positive NINO3 and $h_E + h_W$ perturbations significantly enhances the evolution of positive anomalies, leading to a stronger El Niño event if the RC is initialized before the peak. If the RC is initialized after the peak, these perturbations extend the duration of El Niño conditions by delaying the decay of warm anomalies.

For negative or slightly negative (neutral) initial conditions, the positive optimal perturbations effectively suppress the further 430 development of negative anomalies, once again accelerating the return to neutral conditions.

In the supercritical regime, thermocline depth perturbations are not purely activated by atmospheric noise but also emerge due to the internal instability of the system. The influence of stochastic atmospheric noise is weaker than in the subcritical case. This means that even if the model does not explicitly account for the effect of wind-stress forcing on the thermocline slope, strong initial thermocline depth anomalies are not needed to maximize error propagation. However, even in the supercritical 435 regime, when τ_C is not included, the RC remains more sensitive to initial ~~NINO3 perturbations than the deterministic ZC model.~~ thermocline perturbations at longer lead times than when τ_C is included, but the difference is much less pronounced than in the subcritical regime. Also the differences in terms of the distribution of the optimal perturbations per initial conditions category is less pronounced compared to the subcritical regime and is only evident for the negative and neutral initial conditions (not for the positive ones).

4   Summary and Discussion

440

Relatively limited research has been carried out to understand the underlying reasons for the strong performance of ML prediction models in ENSO prediction, in particular their apparent ability to reduce error propagation and overcome the Spring Predictability Barrier (SPB) as deduced from dynamical models. In previous studies, explainable AI techniques like Layerwise Relevance Propagation (LRP) have been used to identify and estimate which patterns in the data are exploited by Machine 445 Learning (ML) methods to make specific ENSO predictions (Ham et al., 2019b; Rivera Tello et al., 2023), or to explore teleconnections of ENSO (Ito et al., 2021; Liu et al., 2023b). The LRP technique has also been extended to the Echo State Network (ESN) framework to investigate the importance of the leaking rate parameter α and the ESN's robustness to random input perturbations while performing a El Niño/La Niña binary classification task (Landt-Hayen et al., 2022). ~~Our work focused on a single ML technique, i.e.,~~ In a recent study (Qin et al. (2024)), an approach similar to ours is followed, using

the CNOP framework to estimate optimal initial perturbations for a U-net deep learning model trained on both reanalysis data and simulations from various CMIP6 Global Circulation Models (GCMs). They validated their results with the GFDL CM2p1 numerical model, showing that the deep learning models and numerical models exhibit a similar error evolution over time for the same initial conditions and superimposed perturbations. However, their automatic differentiation–based optimization algorithm only applies to deep learning architectures, preventing them from determining whether the deep learning and numerical models show the same optimal initial perturbations, leading to the largest error propagation.

In our study, we address this limitation by employing a gradient-free optimization algorithm (Cobyla), enabling a fair comparison between a Reservoir Computing (RC) ~~and showed~~ model and the Zebiak and Cane (ZC) numerical model. Our results indicate that the RC model ~~'s~~ effectively reduces error growth from optimal initial perturbations compared to the ZC model, offering a plausible explanation for its higher predictive skill ~~at long lead times can be attributed to its capacity to limit the growth of initial uncertainties over time, compared to the dynamical Zebiak-Cane (ZC) model~~. It is important to clarify that our study does not aim to analyze the observed ENSO behaviour. The objective of our study is only to demonstrate that the RC can mitigate error propagation resulting from initial conditions perturbations more effectively than a classical dynamical numerical model. Such an analysis and comparison is impossible using real-world observations as we do not know the evolution operator of the real-world system and hence cannot determine the CNOP. By focusing on the ZC model data, we analyze the RC model's behavior and learned dynamics within a well defined case.

We first demonstrated that the RC, when trained on data from the stochastic ZC model (acting as synthetic observations), exhibits good predictive skill up to an 18-month lead time and hence effectively overcomes the SPB problem both in the subcritical and supercritical regimes. In the supercritical regime, the RC model performs better when zonal surface wind speed anomalies are included during training while in the subcritical regime the RC actually performs better for longer lead times (9 to 18 months) when the zonal surface wind speed anomalies are excluded. While this result may depend on the implementation of the wind-stress noise (Feng and Dijkstra, 2017) which we restricted here mostly to the eastern Pacific (by using only the first EOF of the residual wind-stress field), the reason is that the RC is overfitting the noise in the subcritical regime.

Previous studies have also noticed strong predictive performances when applying the RC to ENSO forecasting. For instance, Hassanibesheli et al. (2022) achieved high prediction skills (ACC > 0.8) up to a lead time of 14 months when training the RC with the observed NINO3 and NINO3.4 indexes, decomposed into a low-frequency and high-frequency components. Their performance is comparable to ours at long lead times but our model performs better at shorter lead times (3-6 months). Additionally, like in our study, they found that their approach could mitigate the SPB problem. However, care must be taken when comparing our findings with their results due to substantial differences in the data used for training, the training variables considered, and the implementation of the forecasting framework.

After the RC's performance analysis, we investigated the propagation of errors in initial conditions in boreal spring (just before the SPB) for both the RC and deterministic ZC models using the Conditional Nonlinear Optimal Perturbation (CNOP) approach (Duan et al., 2013b). In the supercritical regime, the RC can significantly reduce error propagation in particular at longer lead times (6-9 months). In the subcritical regime, the RC is less susceptible to perturbations compared to the ZC model, when surface ~~wind~~ wind-speed anomalies are excluded during training and more susceptible when they are included. The

485 reduced sensitivity of AI models to small initial perturbations has also been found in Selz and Craig (2023). In that study, the inability of the AI frameworks to reproduce the "butterfly effect" of the atmosphere was considered a limitation, as it prevents the generation of large ensembles due to inadequate error growth properties. However, as noted in Selz and Craig (2023), this limitation can be mitigated by training multiple models with different random seeds to generate a confidence interval. We argue that this reduced sensitivity to initial perturbations is not a disadvantage but is what enables AI models to extend the

490 predictability horizon of a system, allowing them to maintain higher predictive skills at longer lead times. The actual CNOPs have quite a different pattern for the ZC and RC cases; the CNOP pattern of the ZC resembles the one obtained in earlier papers (Duan et al., 2013b) with a dominant response in the thermocline field for longer lead times, but in the CNOP pattern of the RC also a strong sea surface temperature component is present.

The thermocline anomalies are important for error propagation on the longer time scales, in particular in the ZC model in
495 which the ENSO variability is highly affected by the thermocline feedback (Zebiak and Cane, 1987). Hence, effectively, the RC model reduces the components in the thermocline anomalies and hence reduces error propagation. While we restricted to only particular cases, as we only used one value of parameter in the constraint condition δ and we allowed only one EOF in the ZC wind-stress noise, we think that the modification of the dynamical behavior in the RC (with respect to the ZC) to change the spatio-temporal properties of the error propagation is the key explanation for the superior skill of the RC on long

500 lead times and the reason for being able to overcome the SPB. ~~The same possibly also holds for other ML techniques, such as CNN approaches (Ham et al., 2019b) and the methodology used here (using the CNOP approach) is~~ As a final remark, we specify that this mechanism is proposed to explain the RC's superior performance within a relatively short-term prediction horizon, as opposed to the decadal timescales required to assess the long-term dynamics and statistics of ENSO. Developing an emulator that captures the long-term dynamics and statistics of a system is an entirely different task, necessitating distinct model

505 architectures, hyperparameter configurations, and evaluation criteria compared to those adopted in our study, like assessing how effectively a model captures the intrinsic nonlinearities of the system. In recent years, various Machine Learning models have demonstrated impressive predictive skills up to a ~~promising way to determine whether this is the case.~~ 21-month lead time without necessarily capturing all the underlying physical processes of ENSO. The results shown here for the Reservoir Computer can be extended to other Machine Learning models, potentially explaining their predictive skills up to nearly two

510 years lead time, far beyond the SPB.

Code availability. All data and code used in this study are available at this link https://zenodo.org/records/15006826

References

Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S., and DeWitt, D. G.: Skill of Real-Time Seasonal ENSO Model Predictions during 2002–11: Is Our Capability Increasing?, Bull. Amer. Meteor. Soc., 93, 631–651, 2012.

Battisti, D. S. and Hirst, A. C.: Interannual Variability in a Tropical Atmosphere-Ocean Model: Influence of the Basic State, Ocean Geometry and Nonlinearity ., Journal of the Atmospheric Sciences, 46, 1687–1712, https://doi.org/10.1175/1520-0469(1989)046<1687:IVIATA>2.0.CO;2, 1989.

Bracco, A., Brajard, J., Dijkstra, H. A., Hassanzadeh, P., Lessig, C., and Monteleoni, C.: Machine Learning for the Physics of Climate, Nature Reviews Physics, https://doi.org/10.1038/s42254-024-00776-3, 2024.

Copernicus Climate Change Service: ORAS5 global ocean reanalysis monthly data from 1958 to present, 2021.

Duan, W. and Wei, C.: The 'spring predictability barrier' for ENSO predictions and its possible mechanism: Results from a fully coupled model, International Journal of Climatology, 33, 1280–1292, https://doi.org/10.1002/joc.3513, 2013.

Duan, W., Yu, Y., Xu, H., and Zhao, P.: Behaviors of nonlinearities modulating the El Niño events induced by optimal precursory disturbances, Climate Dynamics, 40, 1399–1413, 2013a.

Duan, W., Yu, Y., Xu, H., and Zhao, P.: Behaviors of nonlinearities modulating the El Niño events induced by optimal precursory disturbances, Climate Dynamics, 40, 1399–1413, 2013b.

Feng, Q. Y. and Dijkstra, H. A.: Climate network stability measures of El Niño variability, Chaos, 27, 035 801–15, 2017.

Guardamagna, F., Wieners, C., Fang, X., and Dijkstra, H. A.: Detection of limit cycle signatures of El Niño in models and observations using reservoir computing, Journal of Physics: Complexity, 5, 015 016, https://doi.org/10.1088/2632-072X/ad2699, 2024.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature Publishing Group, pp. 1–17, 2019a.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature, 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7, 2019b.

Hassanibesheli, F., Kurths, J., and Boers, N.: Long-term ENSO prediction with echo-state networks, Environmental Research: Climate, 1, 011 002, https://doi.org/10.1088/2752-5295/ac7f4c, 2022.

Hu, J., Weng, B., Huang, T., Gao, J., Ye, F., and You, L.: Deep Residual Convolutional Neural Network Combining Dropout and Transfer Learning for ENSO Forecasting, Geophysical Research Letters, 48, https://doi.org/10.1029/2021GL093531, 2021.

Ito, M., Karamperidou, C., Sadowski, P., Camargo, S., Lee, C.-Y., and Patricola, C.: Explainable Artificial Intelligence for Insights into the Relationship between ENSO and Tropical Cyclone Genesis, in: AGU Fall Meeting Abstracts, vol. 2021, pp. GC42A–07, 2021.

Jin, F.-F.: An Equatorial Ocean Recharge Paradigm for ENSO. Part I: Conceptual Model, Journal of the Atmospheric Sciences, 54, 811 – 829, https://doi.org/10.1175/1520-0469(1997)054<0811:AEORPF>2.0.CO;2, 1997.

Jin, Y. and Liu, Z.: A Theory of the Spring Persistence Barrier on ENSO. Part I: The Role of ENSO Period, Journal of Climate, 34, 2145 – 2155, https://doi.org/10.1175/JCLI-D-20-0540.1, 2021a.

Jin, Y. and Liu, Z.: A theory of Spring Persistence Barrier on ENSO. Part II: Persistence Barriers in SST and ocean heat content, Journal of Climate, pp. 1–30, https://doi.org/10.1175/JCLI-D-20-0820.1, 2021b.

Jin, Y., Liu, Z., and McPhaden, M. J.: A Theory of the Spring Persistence Barrier on ENSO. Part III: The Role of Tropical Pacific Ocean Heat Content, Journal of Climate, 34, 8567 – 8577, https://doi.org/10.1175/JCLI-D-21-0070.1, 2021.

Jonnalagadda, J. and Hashemi, M.: Long Lead ENSO Forecast Using an Adaptive Graph Convolutional Recurrent Neural Network, Engineering Proceedings, 39, https://doi.org/10.3390/engproc2023039005, 2023.

Kessler, W. S.: Is ENSO a cycle or a series of events?, Geophysical Research Letters, 29, 40–1–40–4, https://doi.org/https://doi.org/10.1029/2002GL015924, 2002.

Landt-Hayen, M., Kröger, P., Claus, M., and Rath, W.: Layer-wise Relevance Propagation for Echo State Networks applied to Earth System Variability, https://arxiv.org/abs/2210.09958, 2022.

Lau, K.-M. and Yang, S.: The Asian monsoon and predictability of the tropical ocean–atmosphere system, Quarterly Journal of the Royal Meteorological Society, 122, 945–957, https://doi.org/https://doi.org/10.1002/qj.49712253208, 1996.

Liu, Y., Cai, W., Lin, X., Li, Z., and Zhang, Y.: Nonlinear El Niño impacts on the global economy under climate change, Nature Communications, 14, 5887, 2023a.

Liu, Y., Duffy, K., Dy, J. G., and Ganguly, A. R.: Explainable deep learning for insights in El Niño and river flows, Nature Communications, 14, https://doi.org/10.1038/s41467-023-35968-5, 2023b.

Lukoševičius, M.: A Practical Guide to Applying Echo State Networks, pp. 659–686, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-35289-8_36, 2012.

Mahesh, A., cody Evans, M., Jain, G., Castillo, M., Lima, A. R., Lunghino, B. D., Gupta, H., Gaitan, C. F., Hunt, J., Tavasoli, O., and Brown, P. T.: Forecasting El Niño with Convolutional and Recurrent Neural Networks, https://api.semanticscholar.org/CorpusID:209371718, 2019.

McPhaden, M. J., Zebiak, S. E., and Glantz, M. H.: ENSO as an Integrating Concept in Earth Science, Science, 314, 1740–1745, https://doi.org/10.1126/science.1132588, 2006.

Mu, M., Xu, H., and Duan, W.: A kind of initial errors related to "spring predictability barrier" for El Niño events in Zebiak-Cane model, Geophysical Research Letters, 34, https://doi.org/https://doi.org/10.1029/2006GL027412, 2007.

Neelin, J., Battisti, D. S., Hirst, A. C., Jin, F.-F., Wakata, Y., Yamagata, T., and Zebiak, S. E.: ENSO Theory, Journal of Geophysical Research, 103, 14,261–14,290, 1998.

Pathak, J., Wikner, A., Fussell, R., Chandra, S., Hunt, B. R., Girvan, M., and Ott, E.: Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model, Chaos, 28, 041 101, 2018.

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., McGregor, S., McPhaden, M. J., Power, S., Roehrig, R., Vialard, J., and Voldoire, A.: Evaluating Climate Models with the CLIVAR 2020 ENSO Metrics Package, Bulletin of the American Meteorological Society, 102, E193 – E217, https://doi.org/10.1175/BAMS-D-19-0337.1, 2021.

Powell, M. J. D.: A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation, pp. 51–67, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-015-8330-5_4, 1994.

Qin, B., Yang, Z., Mu, M., Wei, Y., Cui, Y., Fang, X., Dai, G., and Yuan, S.: The first kind of predictability problem of El Niño predictions in a multivariate coupled data-driven model, Quarterly Journal of the Royal Meteorological Society, 150, 5452–5471, https://doi.org/https://doi.org/10.1002/qj.4882, 2024.

Rivera Tello, G. A., Takahashi, K., and Karamperidou, C.: Explained predictions of strong eastern Pacific El Niño events using deep learning, Scientific Reports, 13, 21 150, 2023.

Roulston, M. S. and Neelin, J. D.: The response of an ENSO Model to climate noise, weather noise and intraseasonal forcing, Geophysical Research Letters, 27, 3723–3726, https://doi.org/https://doi.org/10.1029/2000GL011941, 2000.

Selz, T. and Craig, G. C.: Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?, Geophysical Research Letters, 50, e2023GL105 747, https://doi.org/https://doi.org/10.1029/2023GL105747, e2023GL105747 2023GL105747, 2023.

Suarez, M. J. and Schopf, P. S.: A Delayed Action Oscillator for ENSO., Journal of the Atmospheric Sciences, 45, 3283–3287, https://doi.org/10.1175/1520-0469(1988)045<3283:ADAOFE>2.0.CO;2, 1988.

Takahashi, K., Karamperidou, C., and Dewitte, B.: A theoretical model of strong and moderate El Niño regimes, Climate Dynamics, 52, 7477–7493, 2019.

590 Timmermann, A., Jin, F.-F., and Abshagen, J.: A Nonlinear Theory for El Niño Bursting., Journal of the Atmospheric Sciences, 60, 152–165, https://doi.org/10.1175/1520-0469(2003)060<0152:ANTFEN>2.0.CO;2, 2003.

Tziperman, E., Stone, L., Cane, M. A., and Jarosh, H.: El Niño chaos: overlapping of resonances between the seasonal cycle and the Pacific ocean-atmosphere oscillator, Science, 264, 72–74, 1994.

Verstraeten, D., Schrauwen, B., D'Haene, M., and Stroobandt, D.: An experimental unification of reservoir computing methods, Neural
595 Networks, 20, 391–403, https://doi.org/https://doi.org/10.1016/j.neunet.2007.04.003, echo State Networks and Liquid State Machines, 2007.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald,
600 A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Wang, B., Yang, Y., Ding, Q.-H., Murakami, H., and Huang, F.: Climate control of the global tropical storm days (1965–2008), Geophysical Research Letters, 37, https://doi.org/https://doi.org/10.1029/2010GL042487, 2010.

Webster, P. and Yang, S.: Monsoon and ENSO: Selectively Interactive Systems, Quarterly Journal of the Royal Meteorological Society, 118,
605 877 – 926, https://doi.org/10.1002/qj.49711850705, 1992.

Xiaoqun, C., Yanan, G., Bainian, L., Kecheng, P., Guangjie, W., and Mei, G.: ENSO prediction based on Long Short-Term Memory (LSTM), IOP Conference Series: Materials Science and Engineering, 799, 012 035, https://doi.org/10.1088/1757-899X/799/1/012035, 2020.

Yin, H., Wu, Z., Fowler, H. J., Blenkinsop, S., He, H., and Li, Y.: The Combined Impacts of ENSO and IOD on Global Seasonal Droughts, Atmosphere, 13, https://doi.org/10.3390/atmos13101673, 2022.

610 Yu, L., Mu, M., and Yu, Y.: Role of parameter errors in the spring predictability barrier for ENSO events in the Zebiak-Cane model, Advances in Atmospheric Sciences, 31, https://doi.org/10.1007/s00376-013-3058-3, 2014.

Zebiak, S. and Cane, M.: A model of El Nino-Southern Oscillation, Monthly Weather Review, 115, http://academiccommons.columbia.edu /catalog/ac%3A145611, 1987.

Appendix A: Zebiak and Cane model results

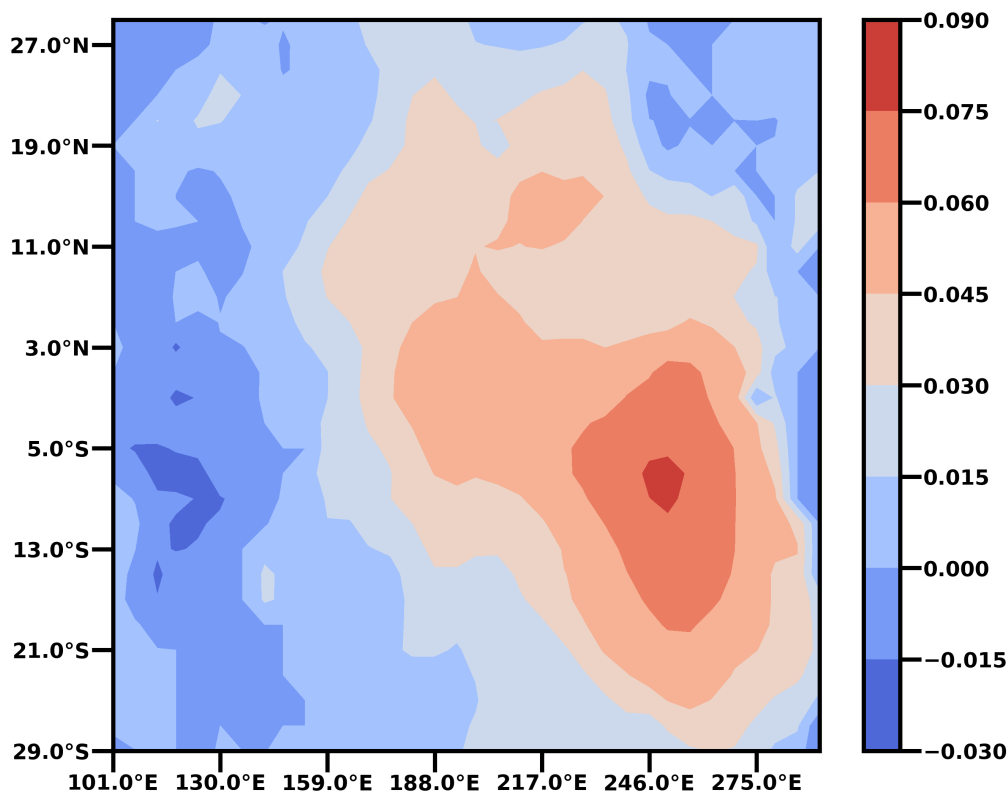



Figure A1. First EOF of the residual zonal wind stress anomalies as determined from the ORAS5 dataset (Copernicus Climate Change Service (2021)).

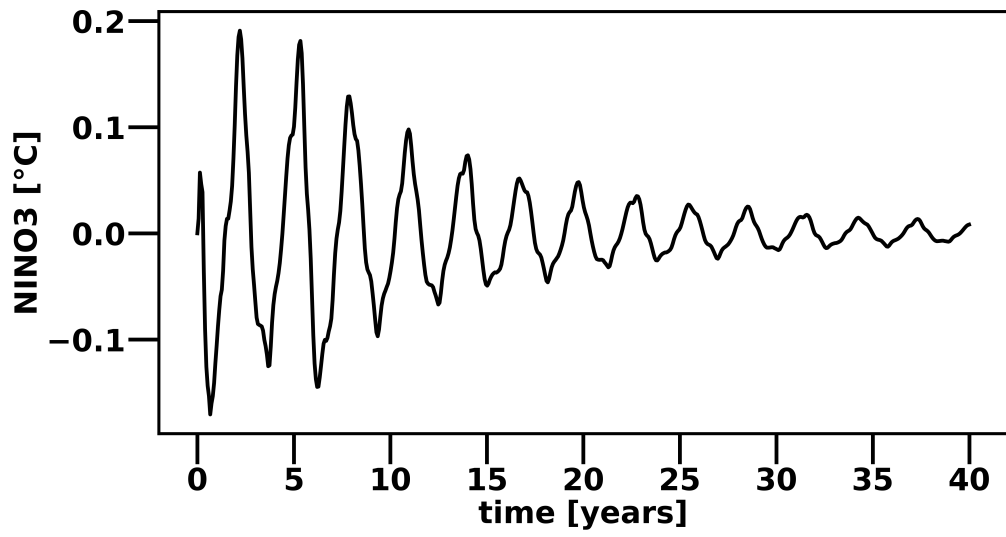
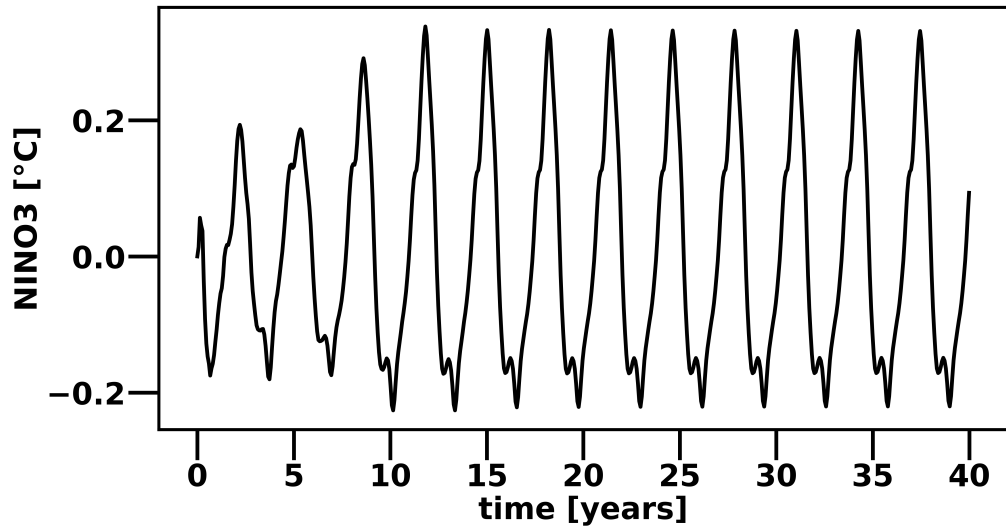



Figure A2. NINO3 index from the deterministic ZC model for (a) $r_d = 0.79$ and (b) $r_d = 0.8$.

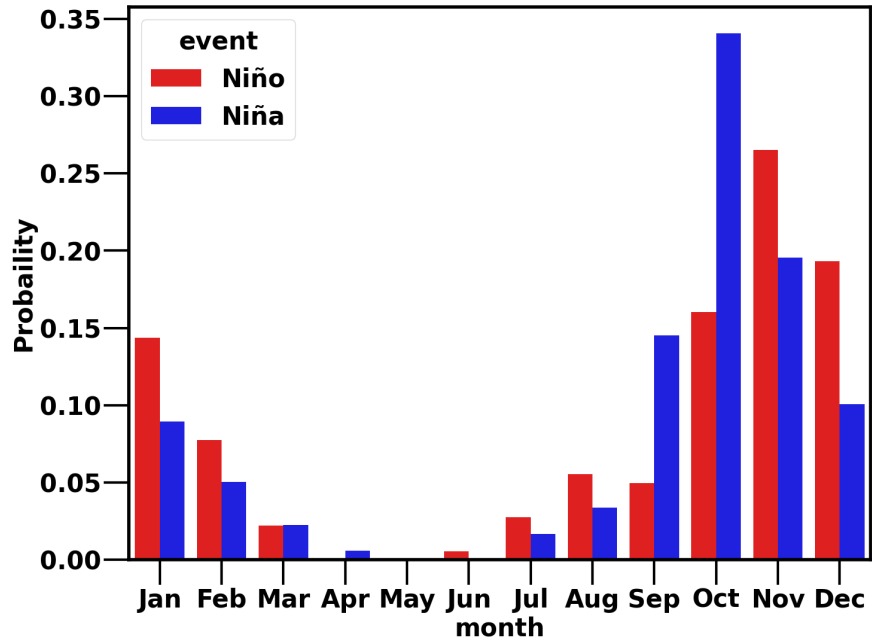
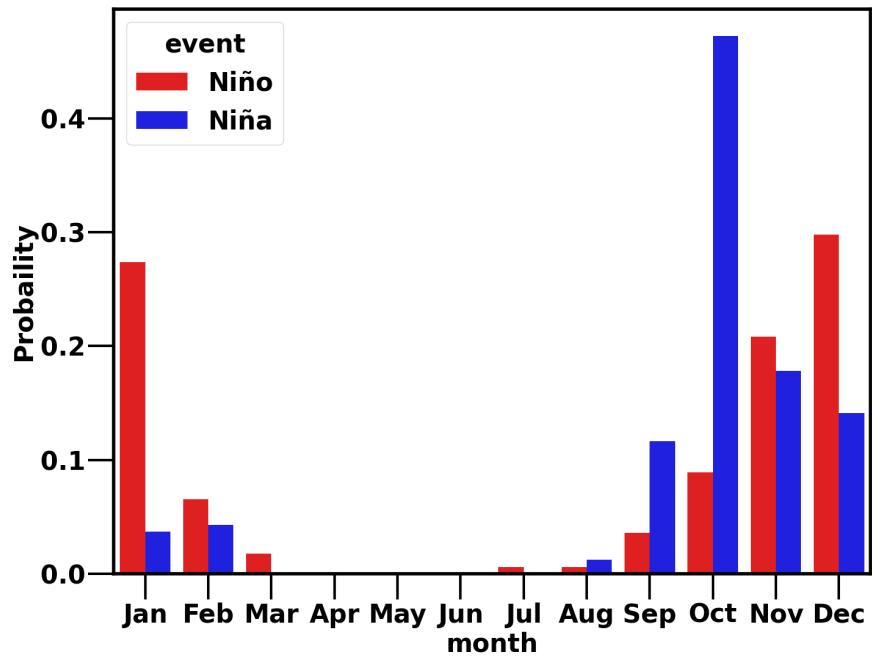



Figure A3. Frequency of the occurrence of La Niña and El Niño events for each calendar month. (a) $r_d = 0.77$ (b) $r_d = 0.9$. For both r_d values a stochastic ZC model realization of 1000 years has been considered.

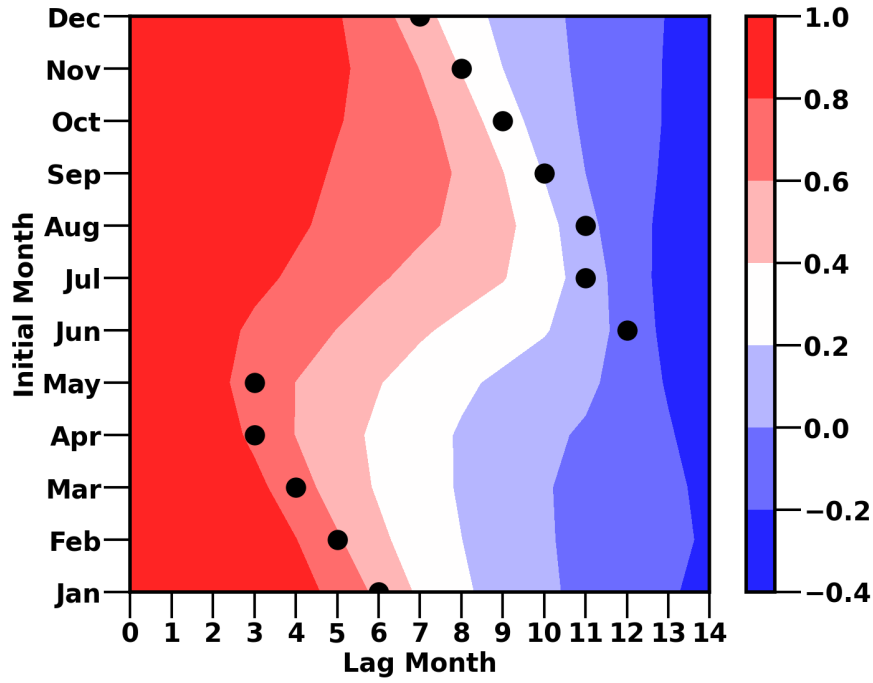
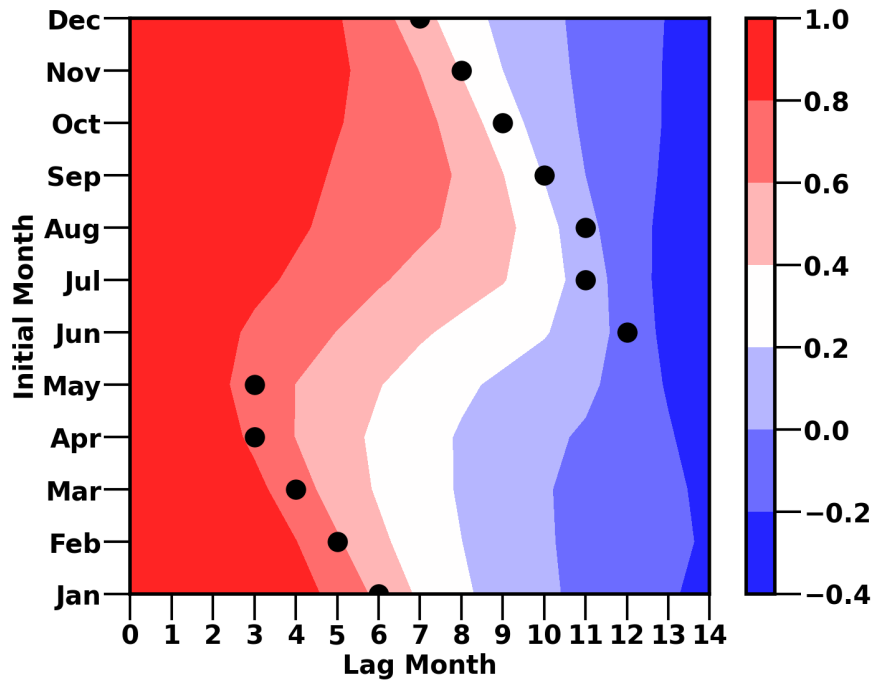



Figure A4. Level of autocorrelation of the NINO3 index for different starting months for (a) $r_d = 0.77$ and (b) $r_d = 0.9$. The black dots indicate the lag month corresponding to the maximum decrease in autocorrelation. In each case, a 1000 years stochastic ZC model realization, has been considered

Appendix B: COBYLA Algorithm

The Constrained Optimization BY Linear Approximations (COBYLA) algorithm (Powell (1994)) is a gradient-free optimization method designed to solve nonlinear constrained optimization problems. Given an objective function $J(x)$ to be minimized, where $x \in \mathbb{R}^n$, and a set of nonlinear constraints $C_i(x) \leq 0$, the algorithm starts from an initial guess x_0 and constructs an initial n-dimensional simplex, represented by a set of $n+1$ vertices $V^{(0)} = \{v_1^{(0)}, v_2^{(0)}, ..., v_{n+1}^{(0)}\} \subset \mathbb{R}^n$. Each vertex is defined as $v_i^{(0)} = x_0 + \rho_{\text{beg}} e_i$, where e_i is the i-th coordinate vector (a standard basis vector in $\mathbb{R}^n$), and ρ_{beg} is a specified initial trust-region radius that determines the initial simplex size.

At each iteration, the algorithm constructs a linear approximation of both the objective function and the constraints using a linear interpolation at the $n+1$ current simplex vertices $V^{(i)}$. It then identifies the worst-performing vertex, denoted as $v_{\text{worst}}^{(i)} = \arg\max_{v \in V^{(i)}} J(v)$. Once the worst-performing vertex is found, the algorithm formulates and solves a linear optimization problem within a trust region of radius ρ around $v_{\text{worst}}^{(i)}$ to generate a new candidate vertex $v_{\text{new}}^{(i+1)}$. If this new vertex improves the objective function, it replaces $v_{\text{worst}}^{(i)}$, updating the simplex structure. Otherwise, the trust-region radius is reduced, and the linear optimization problem is solved again.

The algorithm continues iterating until one of the stopping criteria is met. It terminates when the trust-region radius ρ falls below a predefined threshold ρ_{end}, when the change in the objective function is smaller than a specified tolerance ε, or when the number of iterations reaches the maximum allowed value N_{max}. COBYLA is particularly well-suited for nonlinear optimization problems with a small number of variables, especially when computing derivatives is challenging or infeasible. These characteristics make COBYLA an ideal choice for our analysis, as the ZC model's derivatives are impossible to compute, and our optimization problem involves only three variables (see Sections 2.3 and 3.3).

To validate the estimated CNOPs obtained using COBYLA for both the RC and ZC models, we first confirm that the CNOPs lie on the boundary of the sphere defined by the constraints. Next, we evaluate error propagation by applying multiple randomly chosen initial perturbations sampled from this boundary, to determine whether the CNOPs indeed correspond to the largest error growth. Figures B1 and B2 present these validation results. Specifically, Figure B1 shows a scatter plot of all estimated CNOPs for both models in the three-dimensional space defined by the normalized NINO3, h_E and h_W optimal initial perturbations. In contrast, Figure B2 illustrates, for a representative case, the divergence between perturbed and unperturbed trajectories resulting from both the CNOP and 50 random initial perturbations sampled along the constraint boundary. For illustration, results are provided for five different years, with both the RC and ZC models always initialized in April.

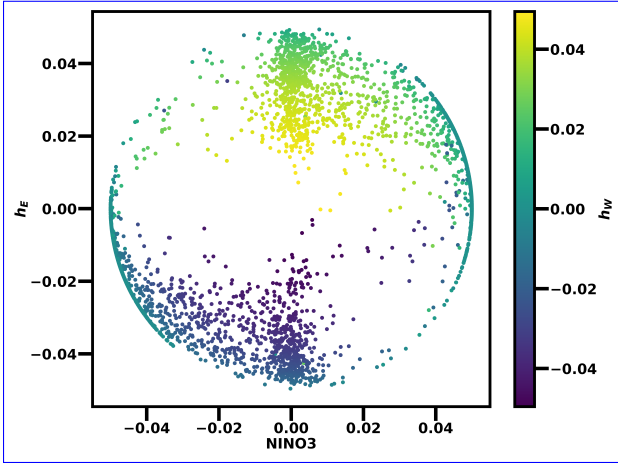
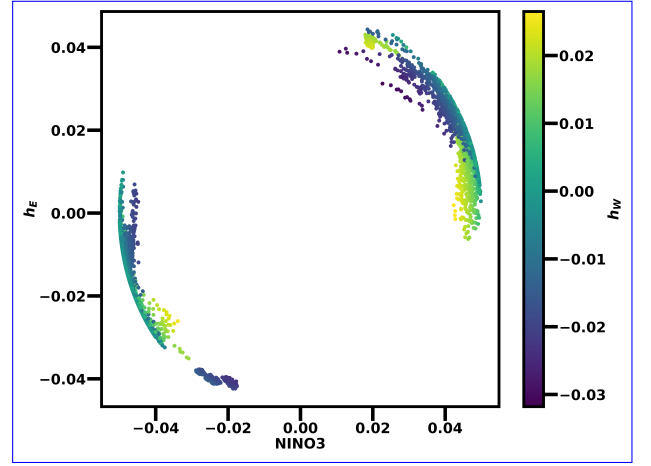
For our implementation, we adopted the COBYLA solver from the SciPy Python library (Virtanen et al. (2020)). Since COBYLA is inherently designed to solve minimization problems, while our objective is to maximize the distance between the reference and perturbed trajectories (see Sections 2.3 and 3.3), we account for this by minimizing $-J(x)$ instead of $J(x)$.

(a) ZC model   (b) RC model

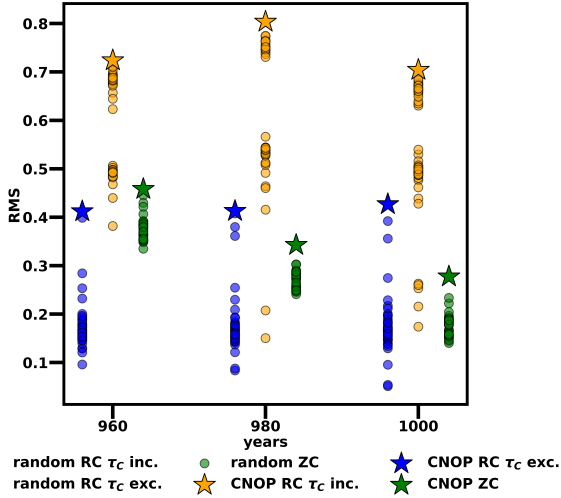
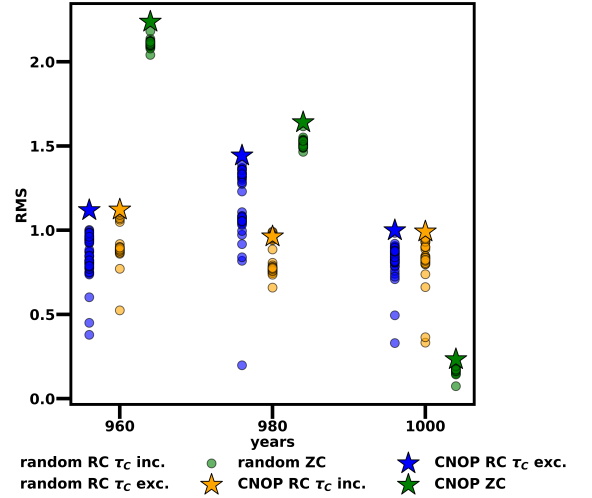
Figure B1. Scatter plot of all computed CNOPs in the normalized NINO3 index vs. [hE, hW] anomaly plane for (a) the ZC model and (b) the RC model for both τ_C included and excluded during training. The NINO3 index is normalized by 2°C, while the [hE, hW] anomalies are normalized by 50m.

(a) subcritical regime   (b) supercritical regime

Figure B2. Scatter plot of the RMS differences between perturbed and unperturbed trajectories for three model years, with both the RC and ZC models always initialized in April. Results are presented for the CNOP (i.e., the optimal initial perturbation) and for 50 random initial perturbations sampled from the boundary of the constraints.

 Appendix C: optimal RC hyperparameters

Input Variables	r_d	Best Hyperparameters
NINO3, h_E, h_W, τ_C	0.9	$N_x = 391, \rho = 0.84, <k> = 0.16,$ $a = 0.8, \alpha = 0.57$
NINO3, h_E, h_W	0.9	$N_x = 404, \rho = 1.07, <k> = 0.2,$ $a = 0.58, \alpha = 0.39$
NINO3, h_E, h_W, τ_C	0.77	$N_x = 476, \rho = 0.8, <k> = 0.1,$ $a = 0.63, \alpha = 0.6$
NINO3, h_E, h_W	0.77	$N_x = 534, \rho = 0.88, <k> = 0.1,$ $a = 0.38, \alpha = 0.17$

Table C1. Table showing the optimal RC model's hyperparameters sets, for the supercritical ($r_d = 0.9$) and subcritical ($r_d = 0.77$) regimes, and each input variable configuration (with and without the inclusion of τ_C).

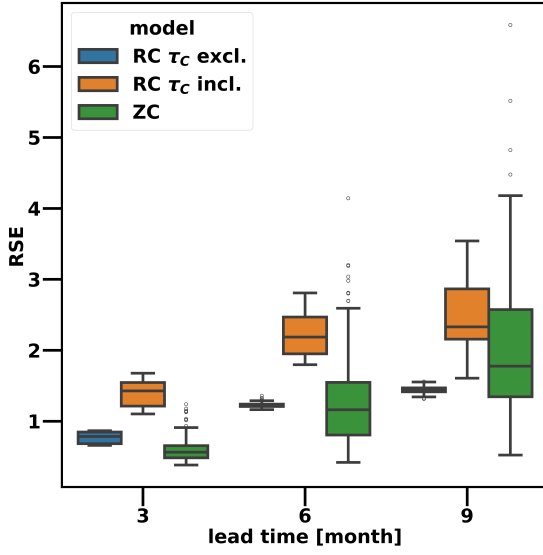
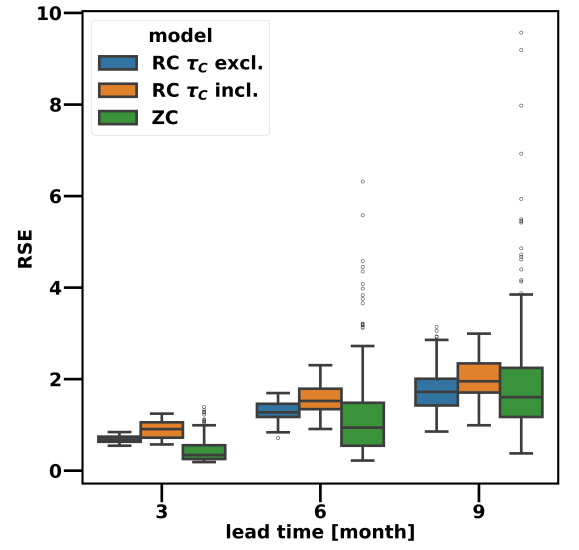
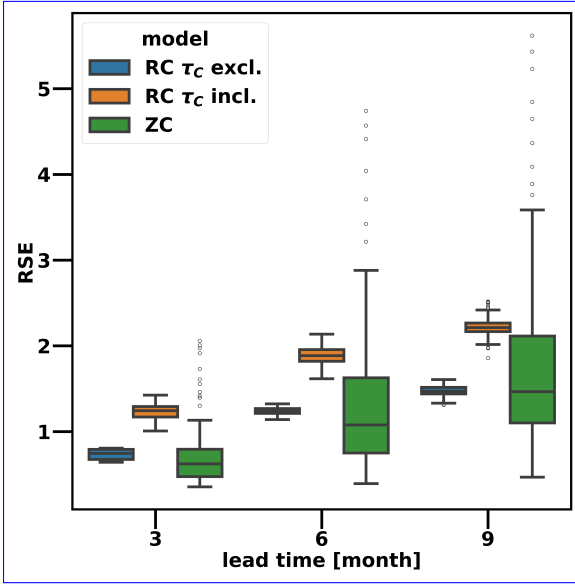
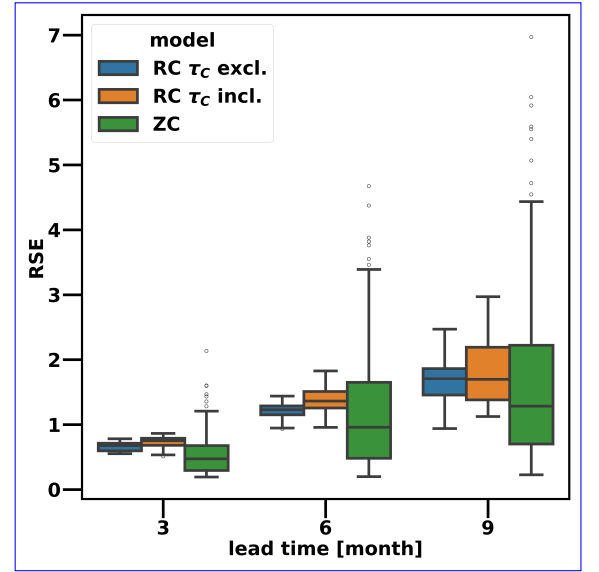
Appendix D: CNOPs

(a) subcritical regime

(b) supercritical regime

(c) subcritical regime

(d) supercritical regime

Figure D1. Distribution of the normalized RSE distances between perturbed and unperturbed trajectories for different lead times with the application of CNOPs, using initial conditions from: (a)-(b) [December,January,February], and (c)-(d) [September,October,November]. The left plots (a) and (c) display results for $r_d = 0.77$, while the right plots (b) and (d) show results for $r_d = 0.9$. The boxes indicate the interquartile range (IQR), the range within the central 50% of data points lie. The whiskers extend to the minimum and maximum values within 1.5 times the IQR from the first and third quartile. The central line corresponds to the median. The RSE distances are normalized by the standard deviation of the NINO3 index extracted from the 50 years of synthetic observations considered for CNOPs computation (0.29 for the subcritical regime and 0.56 for the supercritical regime).

Subcritical Regime ($r_d = 0.77$)

Start/Lead	3 Months	6 Months	9 Months
Spring	0.6	1.52	1.88
Winter	0.56	1.16	1.78
Autumn	0.62	1.08	1.47

Supercritical Regime ($r_d = 0.9$)

Start/Lead	3 Months	6 Months	9 Months
Spring	0.41	1.41	2.07
Winter	0.43 0.34	0.94	1.6
Autumn	0.53 0.48	0.96	1.28

Blue: Forecast crosses the SPB   **Orange**: Forecast does not cross the SPB

Table D1. Table showing the median of the normalized RSE distances between perturbed and unperturbed trajectories at various lead times, with CNOPs applied across different starting seasons. Only the Zebiak and Cane model is considered, with the top table representing the subcritical regime ($r_d = 0.77$) and the bottom table representing the supercritical regime ($r_d = 0.9$). All RSE distances are normalized by the standard deviation of the NINO3 index from the 50 years of synthetic observations considered for the CNOPs computation (0.29 for the subcritical regime and 0.56 for the supercritical regime).

Subcritical Regime, τ_C Included During Training

Model	Start/Lead	3 Months	6 Months	9 Months
RC	Spring	1.35 (0.14)	2.07 (0.33)	2.4 (0.54)
RC	Winter	1.42 (0.33)	2.19 (0.52)	2.33 (0.71)
RC	Autumn	1.24 (0.12)	1.89 (0.14)	2.21 (0.1)
ZC	Spring	0.6 (0.28)	1.52 (1.49)	1.88 (1.86)
ZC	Winter	0.56 (0.17)	1.16 (0.74)	1.78 (1.23)
ZC	Autumn	0.62 (0.32)	1.08 (0.88)	1.47 (1.01)

Subcritical Regime, τ_C Excluded During Training

Model	Start/Lead	3 Months	6 Months	9 Months
RC	Spring	0.76 (0.11)	1.2 (0.05)	1.44 (0.08)
RC	Winter	0.79 (0.16)	1.22 (0.03)	1.44 (0.06)
RC	Autumn	0.75 (0.12)	1.24 (0.05)	1.47 (0.08)
ZC	Spring	0.6 (0.28)	1.52 (1.49)	1.88 (1.86)
ZC	Winter	0.56 (0.17)	1.16 (0.74)	1.78 (1.23)
ZC	Autumn	0.62 (0.32)	1.08 (0.88)	1.47 (1.01)

Blue: Forecast crosses the SPB   **Orange**: Forecast does not cross the SPB

Supercritical Regime, τ_C Included During Training

Model	Start/Lead	3 Months	6 Months	9 Months
RC	Spring	0.86 (0.09)	1.46 (0.22)	1.82 (0.32)
RC	Winter	0.91 (0.33)	1.52 (0.44)	1.95 (0.64)
RC	Autumn	0.75 (0.11)	1.36 (0.25)	1.7 (0.8)
ZC	Spring	0.41 (0.28)	1.41 (1.36)	2.07 (1.85)
ZC	Winter	~~0.43~~ 0.34 (~~0.28~~0.3)	0.94 (0.94)	1.6 (1.07)
ZC	Autumn	~~0.54~~ 0.48 (~~0.46~~0.38)	0.96 (1.17)	1.28 (1.52)

Supercritical Regime, τ_C Excluded During Training

Model	Start/Lead	3 Months	6 Months	9 Months
RC	Spring	0.67 (0.07)	1.3 (0.23)	1.65 (0.52)
RC	Winter	0.68 (0.11)	1.27 (0.29)	1.72 (0.58)
RC	Autumn	0.68 (0.12)	1.23 (0.14)	1.7 (0.41)
ZC	Spring	0.41 (0.28)	1.41 (1.36)	2.07 (1.85)
ZC	Winter	~~0.43~~ 0.34 (~~0.28~~0.3)	0.94 (0.94)	1.6 (1.07)
ZC	Autumn	~~0.54~~ 0.48 (~~0.46~~0.38)	0.96 (1.17)	1.28 (1.52)

Blue: Forecast crosses the SPB   **Orange**: Forecast does not cross the SPB

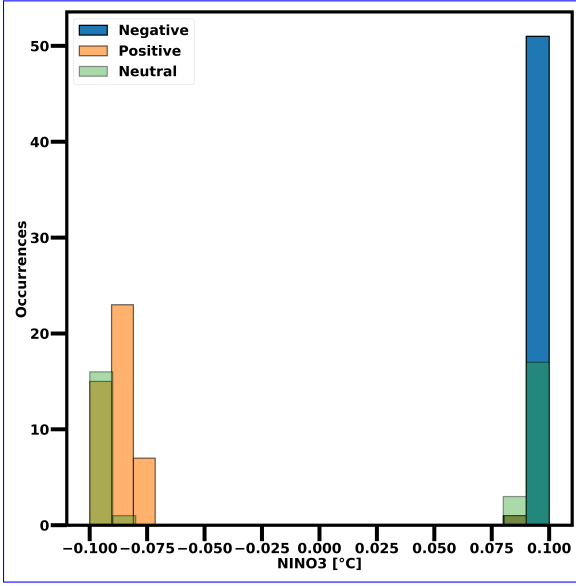
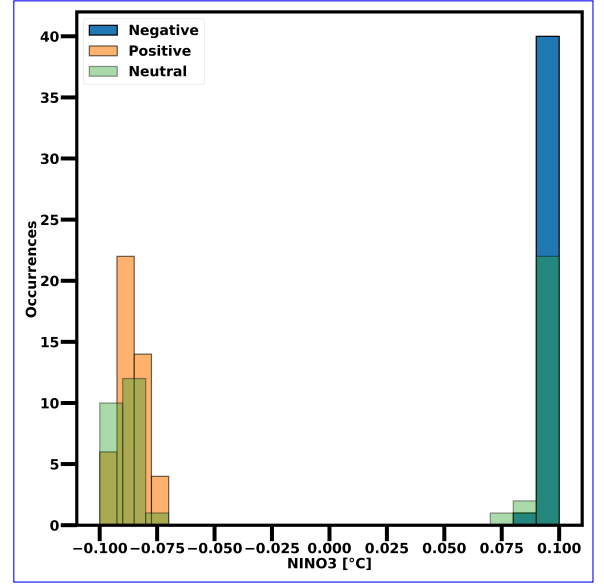
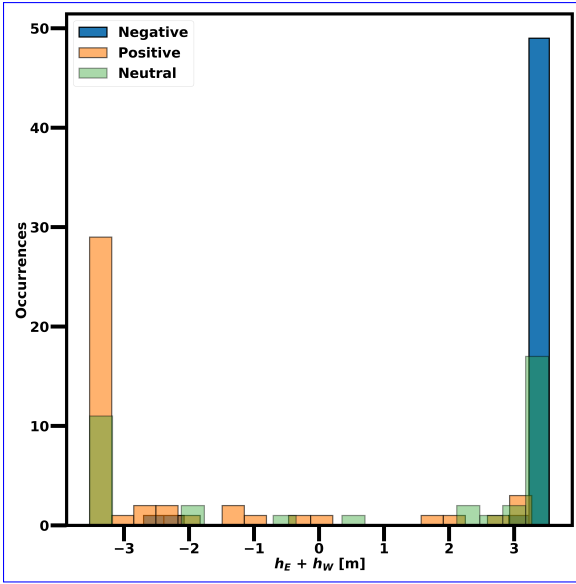
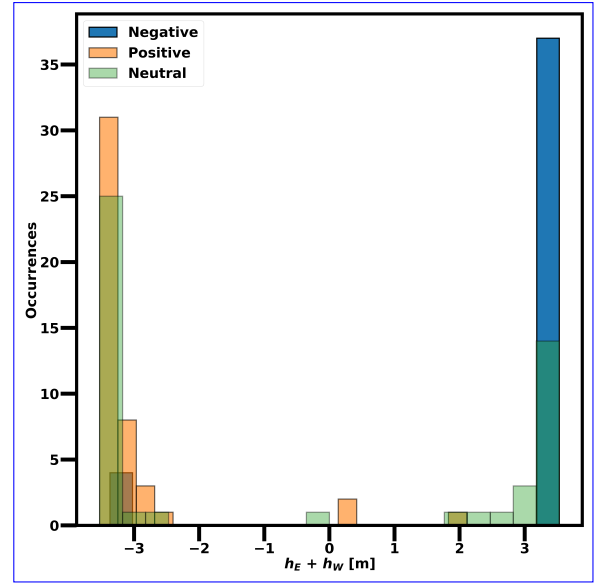
Table D2. Median (IQR) of the normalized RSE distances between perturbed and unperturbed trajectories at various lead times, with CNOPs applied across different starting seasons. Both ZC and RC models (trained with and without τ_C) are considered in the subcritical ($r_d = 0.77$) and supercritical ($r_d = 0.9$) regime. RSE distances are normalized by the standard deviation of NINO3 index from the 50 years of synthetic observations considered for CNOPs computation (0.29 for the subcritical regime and 0.56 for the supercritical regime). The interquartile range IQR is defined as the distance between the first and third quartile.

(a) subcritical regime lead time 3 months

(b) supercritical regime lead time 3 months

(c) subcritical regime lead time 9 months

(d) supercritical regime lead time 9 months

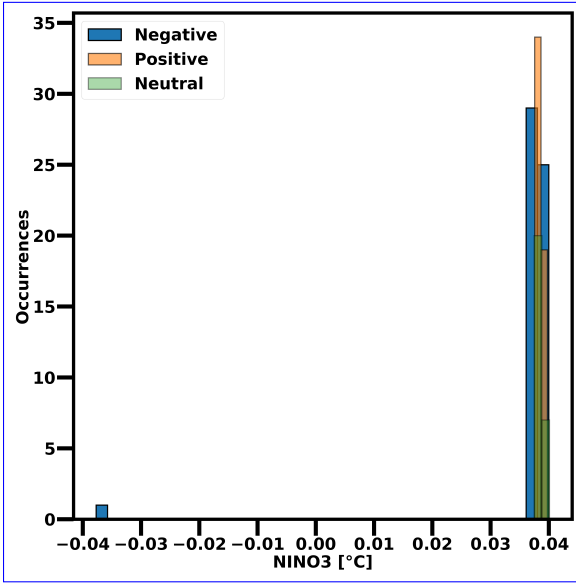
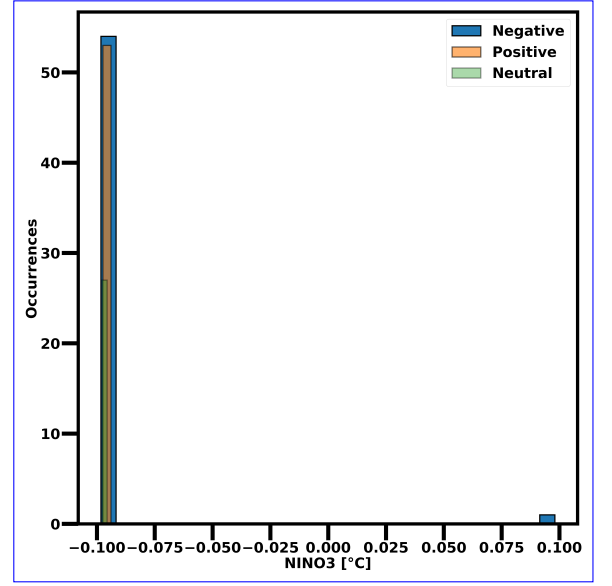
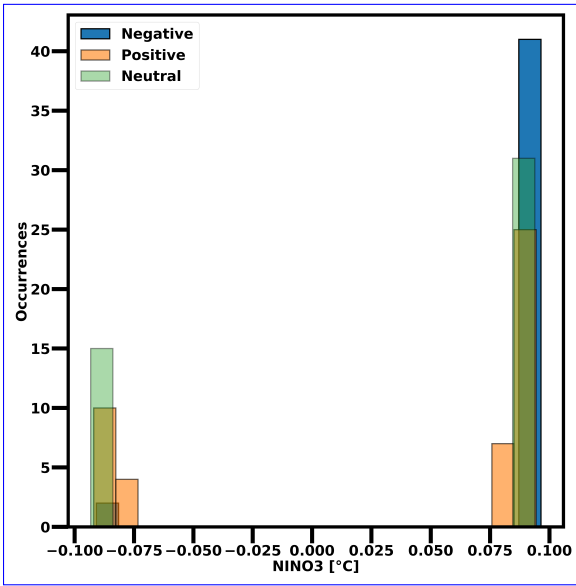
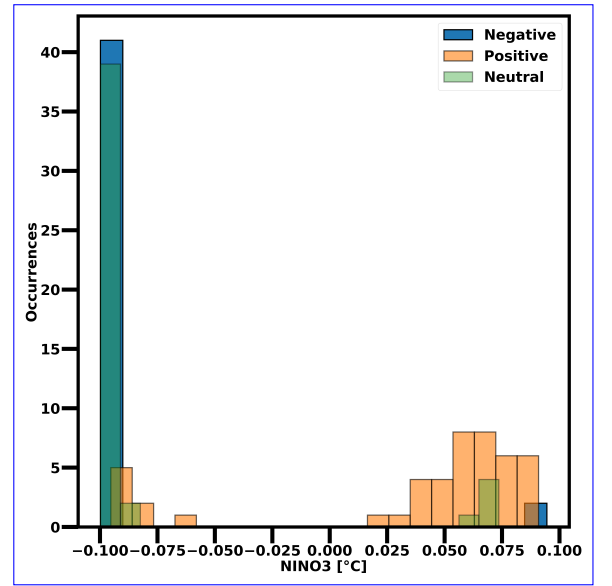
Figure D2. Distribution of the ZC model's CNOPS for positive, negative, and neutral eastern Pacific SSTA initial conditions. Panels (a)-(b) (top row) show the CNOPs obtained for the NINO3 index at a 3-month lead time, comparing subcritical and supercritical regimes, while panels (c)-(d) (bottom row) show the CNOPs obtained for $h_E + h_W$, the sum of the thermocline anomalies in the eastern and western Pacific at a lead time of 9 months. In the supercritical regime, positive, negative, and neutral eastern Pacific SSTA initial conditions are defined as NINO3 ≥ 0.2, NINO3 ≤ -0.2, and $-0.2 <$ NINO3 < 0.2, respectively. In the subcritical regime, positive, negative, and neutral SSTA initial conditions are defined as NINO3 ≥ 0.1, NINO3 ≤ -0.1, and $-0.1 \leq$ NINO3 < 0.1, respectively. In every case the months [March,April,May] (Spring Season) from the last 50 years of the synthetic data used for testing (see section 3.1) are taken as initial conditions, yielding a total of 150 initial conditions considered.

(a) subcritical regime τ_C not included

(b) subcritical regime τ_C included

(c) supercritical regime τ_C not included

(d) supercritical regime τ_C included

Figure D3. Distribution of the RC model's CNOPs for positive, negative, and neutral eastern Pacific SSTA initial conditions. Panels (a)–(b) (top row) display the CNOPs for the NINO3 index at a 9-month lead time under the subcritical regime, with and without τ_C included during training. Panels (c)–(d) (bottom row) show the CNOPs for the supercritical regime, also at a 9-month lead time, with τ_C either included or excluded. In the supercritical regime, positive, negative, and neutral initial conditions are defined as NINO3 ≥ 0.2, NINO3 ≤ -0.2, and $-0.2 < $ NINO3 < 0.2, respectively. In the subcritical regime, positive, negative, and neutral SSTA initial conditions are defined as NINO3 ≥ 0.1, NINO3 ≤ -0.1, and $-0.1 \leq$ NINO3 < 0.1, respectively. In every case the months [March,April,May] (Spring Season) from the last 50 years of the synthetic data used for testing (see section 3.1) **37** taken as initial conditions, yielding a total of 150 initial conditions considered.

Author contributions. All authors contributed to the design of this study. FG carried out all the computations and produced a first draft the paper. All authors contributed to the interpretation of the results and the final version of the paper.

Competing interests. The authors declare that no competing interests are present.

650 *Acknowledgements.* The work of Francesco Guardamagna, Claudia Wieners and Henk Dijkstra was supported by the Netherlands Organization for Scientific Research (NWO) under grant OCENW.M20.277.