ITAT 2016 Proceedings, CEUR Workshop Proceedings Vol. 1649, pp. 102–109
http://ceur-ws.org/Vol-1649, Series ISSN 1613-0073, © 2016 N. Jajcay, M. Paluš

# Statistical modelling in climate science

Nikola Jajcay[1,2] and Milan Paluš[1]

[1] Dept. of Nonlinear Dynamics and Complex Systems, Institute of Computer Science, Academy of Sciences of the Czech Republic
[2] Dept. of Atmospheric Physics, Faculty of Mathematics and Physics, Charles University in Prague

*Abstract:* When it comes to modelling in atmospheric and climate science, the two main types of models are taken into account – dynamical and statistical models. The former ones have a physical basis: they utilize discretized differential equations with a set of conditions (boundary conditions + present state as an initial condition) and model the system's state by integrating the equations forward in time. Models of this type are currently used e.g. as a numerical weather prediction models. The statistical models are considerably different: they are not based on physical mechanisms underlying the dynamics of the modelled system, but rather derived from the analysis of past weather patterns. An example of such a statistical model based on the idea of linear inverse modelling, is examined for modelling the El Niño – Southern Oscillation phenomenon with a focus on modelling cross-scale interactions in the temporal sense. Various noise parameterizations and the possibility of using a multi-variable model is discussed among other characteristics of the statistical model. The prospect of using statistical models with low complexity as a surrogate model for statistical testing of null hypotheses is also discussed.

## 1 Modelling in climate science

Climate models, which rely on the use of quantitative methods to simulate interactions in the climate system, are one of the most important tools to predict and asses future climate projections or to study the climate of the past. In general, two types of models are mainly used: dynamical models and statistical models. The base for a dynamical model is a set of discretized differential equations which are integrated forward in time from the present state, posing as an initial condition. The most prominent example of the usage of dynamical models is without doubt a general circulation model (GCM hereafter). It employs a mathematical model of circulation of the planetary atmosphere and oceans, therefore it uses the Navier-Stokes equations on a rotating sphere (describing a motion of viscous fluid) with thermodynamic terms for energy sources and sinks. The above described model is used in numerical weather prediction, to infer the reanalysis datasets of the past climate and for future climate projections in climate model intercomparison projects CMIP3 [1] and CMIP5 [2].

The uncertainties of the forecast arisen from the GCM models are usually classified into two types: the first one is related to the initial errors (errors in determining the "true" present state of the climate), while the second one is due to the model errors [3] and these are intrinsic. The problem with initial errors is usually tackled by considering an ensemble of model forecasts (instead of just one realization - integration from single initial state), starting with slightly different initial conditions. The model errors are intrinsically connected with the exponential error growth emerging from the chaotic behaviour related to nonlinearities in discretized equations [4]. This limits the predictability of such GCMs to 6-10 days maximum (e.g. [5]).

### 1.1 Statistical models

The second kind of models used in climate science are statistical models. In their design, they are considerably different than the dynamical models in the sense that they are not based on physical mechanism underlying the dynamics of the modelled system, but rather derived from the analysis of past weather patterns. Probably the most used concept is that of inverse stochastic model [6], where the model is designed, then estimated using past data and, finally, stochastically integrated forward in time to obtain the prediction. The disadvantages connected to this type of models consist of the selection of variables that capture the system we are trying to model. Other possible issue could be the non-stationarity of the modelled system - since the statistical model does not involve the underlying physical mechanisms, just the interaction between subsystems (ignoring hidden variables), the model estimated on some subset of the past data may not correctly capture all possible states of the system. In other words, the training period of the past data used to estimate the statistical model may not capture the full phase space of the modelled system.

The motivation for building a statistical model for particular phenomenon, apart from its forecasting, would be to scale down the complexity of the problem. When we find some e.g. nonlinear interactions in the observed data, and we are interested in uncovering the mechanisms, constructing a models of different complexity and seeking such interactions in them would help to expose the mechanisms and shed some light on the problem.

In the following sections, the inverse stochastic model for forecasting the El Niño - Southern Oscillation (ENSO hereafter) phenomenon is built following [7], with the focus on various noise parametrizations and possible use of multiple variables.

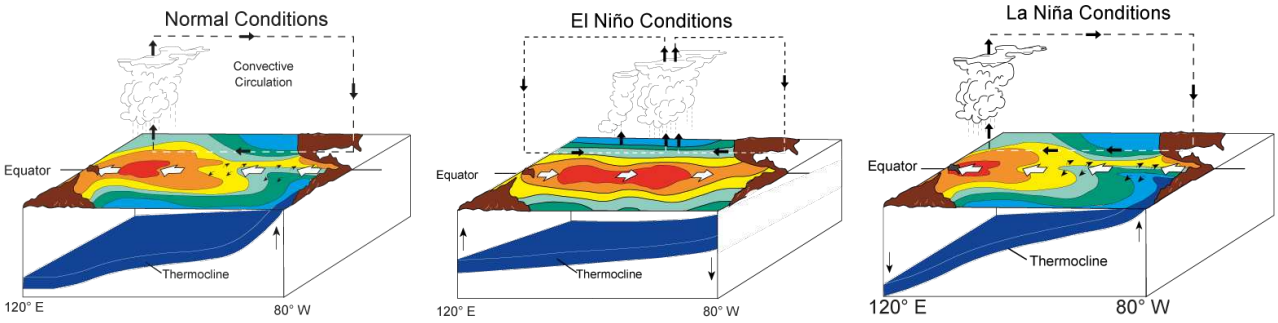


Figure 1: ENSO phenomenon, its phases and mechanisms: (left) neutral, (center) positive and (right) negative. Figures taken from [10].

## 2 Data-based ENSO model

The ENSO phenomenon exhibits strong interannual climate signal and has a great economic and societal impact. It originates from the coupled ocean-atmosphere dynamics of the tropical Pacific [8], but has a strong influence on circulation and air-sea interaction also outside the tropical belt through teleconnections associated with it [9].

The ENSO phenomenon expresses itself as a sea surface temperature (SST hereafter) anomaly and exists in three distinct phases - the neutral, positive (El Niño) and negative (La Niña). The basic physical mechanisms for each of the phases are depicted in Fig. 1. The normal state of the equatorial Pacific (Fig. 1 left) is warm SST in the western basin, near Australia and cold SST in the eastern basin, near the coast of Peru. Above the warm water in the west, the deep convection takes place, where warm and moist air is ascending to the border of troposphere, creating an area of low atmospheric pressure and area of persistent precipitation. From the upper part of the troposphere, the air is moving eastward and then it descends already as cold and dry, creating an area of high atmospheric pressure above the eastern equatorial Pacific. From this basin, the air is blowing westward on the surface, in agreement with the trade winds, finishing the circulation loop known as the Walker circulation. The easterly surface air flow triggers the oceanic surface current to flow poleward, effectively removing water from the surface, thus the water needs to be replaced and this is due to the upwelling, where in the equatorial area, the water is upwelled from roughly 50 meters depth to the surface. Since the thermocline (a border between cold deep ocean and warm surface ocean) is located below 50 meters in the west, the upwelled water is warm, but in the eastern Pacific the thermocline level is above the 50 meters, thus the upwelled water is cold, creating the cold SST in the east and warm SST in the west.

The warm phase of ENSO (Fig. 1 center) creates a warm SST anomaly in the eastern Pacific, acting to weaken the Walker circulation, to move the area of persistent precipitation eastward, to diminish the difference between eastern and western Pacific surface pressures and to level off the thermocline. Reversely, the negative phase of ENSO (Fig. 1 right) is acting to strengthen the Walker circulation, to move the area of persistent precipitation even more westward, the differences in surface pressure is now larger and the thermocline is even more tilted. The ENSO tends to naturally oscillate between these three phases without a distinct period (there is no distinct peak in ENSO signal's spectrum) and the reasons why are still largely unknown.

The important aspect of ENSO is that its positive phase - El Niño is generally characterized by a larger magnitude than its negative phase - La Niña. This statistical skewness is one of the indicators that, at least to some extent, the dynamics of ENSO involves nonlinear processes [11]. At the same time, the most detailed numerical dynamical models seem to severely underestimate this nonlinearity [12], hence the quality of the forecast is not satisfactory.

From the reviews of statistical models for ENSO forecasting before 2000 [13] it was clear, that majority of models were still linear, but lately the nonlinear models are getting more attention (e.g. [14]). In the following, we describe easy-to-interpret nonlinear model for ENSO forecasting.

### 2.1 Inverse models

The concept of inverse stochastic models are used as the starting point in developing the ENSO model. Let $\mathbf{x}(t)$ be the state vector of anomalies, so $\mathbf{x}(t) = \mathbf{X}(t) - \overline{\mathbf{X}}$, where $\mathbf{X}(t)$ is the climate state vector (could be multi- or univariate climate observations e.g. temperature, pressure etc. or a PCA time series from eigen-decomposition of some climate field) and $\overline{\mathbf{X}}$ is its time-mean. The evolution of anomalies could be expressed as

$$\dot{\mathbf{x}} = \mathbf{L}\mathbf{x} + \mathbf{N}(\mathbf{x}) \tag{1}$$

where $\mathbf{L}$ is a linear operator, $\mathbf{N}$ represents the nonlinear terms and dot denotes time derivative.

The simplest type of inverse models is linear inverse models (LIM, [6]). By assuming, in eq. (1), that $\mathbf{N}(\mathbf{x})dx \approx \mathbf{T}\mathbf{x}\,dt + d\mathbf{r}^{(0)}$, where $\mathbf{T}$ is the matrix describing linear feedbacks of unresolved (hidden) processes on $\mathbf{x}$ and $d\mathbf{r}^{(0)}$ is a white-noise process, eq. (1) could be written as

$$d\mathbf{x} = \mathbf{B}^{(0)}\mathbf{x}dt + d\mathbf{r}^{(0)}, \;\; \mathbf{B}^{(0)} = \mathbf{L} + \mathbf{T}. \tag{2}$$

The matrix $\mathbf{B}^{(0)}$ and the covariance matrix of the noise $\mathbf{Q} \equiv \langle \mathbf{r}^{(0)} \mathbf{r}^{(0)T} \rangle$ can be directly estimated from the observed statistics of $\mathbf{x}$ by multiple linear regression [15]. The state vector $\mathbf{x}$, or predictor-variable vector, consists of amplitudes of corresponding principal components (PCA analysis [16] yields spatial patterns - empirical orthogonal functions and its respective time series - principal components), while the vector of response variables contains their tendencies $\dot{\mathbf{x}}$.

### 2.2 Nonlinear multilevel model

The assumptions of linear, stable dynamics and of additive white-noise used to construct LIMs are only valid to certain degree of approximation. In particular, the stochastic forcing $\mathrm{d}\mathbf{r}^{(0)}$ typically involves serial correlations, and, in addition, the matrices $\mathbf{B}^{(0)}$ and $\mathbf{Q}$ obtained from the data exhibit substantial dependence on the lag, that was used to fit them [17]. The two modifications of the basic inverse model, that address both nonlinearity and serial correlations are taken into account, as in [18].

The first modification is obtained by assuming polynomial, rather than linear form of $\mathbf{N}(\mathbf{x})$ in eq. (1), in particular, a quadratic dependence. The $i^{\text{th}}$ component $N_i(\mathbf{x})$ could be written as

$$N_i(\mathbf{x}) \approx \left( \mathbf{x}^T \mathbf{A}_i \mathbf{x} + \mathbf{t}_i \mathbf{x} + c_i^{(0)} \right) \mathrm{d}t + \mathrm{d}r_i^{(0)} \qquad (3)$$

The matrices $\mathbf{A}_i$ represent the blocks of a third-order tensors, while the vectors $\mathbf{b}_i^{(0)} = \mathbf{l}_i + \mathbf{t}_i$ are the rows of the matrix $\mathbf{B}^{(0)} = \mathbf{L} + \mathbf{T}$ (as in eq. (2)). These objects, as well as components of the vector $\mathbf{c}^{(0)}$, are estimated by multiple polynomial regression [19].

The second modification, considering the serial correlations in residual forcing, is due to the multilevel structure of our model. In particular, consider the $i^{\text{th}}$ component of the first, main level of the inverse stochastic model

$$\mathrm{d}x_i = \left( \mathbf{x}^T \mathbf{A}_i \mathbf{x} + \mathbf{b}_i^{(0)} + c_i^{(0)} \right) \mathrm{d}t + \mathrm{d}r_i^{(0)}, \qquad (4)$$

where $\mathbf{x} = \{x_i\}$ is the state vector and matrices $\mathbf{A}_i$, vectors $\mathbf{b}_i^{(0)}$ and the components $c_i^{(0)}$ of the vector $\mathbf{c}^{(0)}$ as well as the components $r_i^{(0)}$ of the residual forcing vector $\mathbf{r}^{(0)}$ are determined by the least squares. The additional model level is added to express the known increments $\mathrm{d}\mathbf{r}^{(0)}$ as a linear function of an extended state vector $[\mathbf{x}, \mathbf{r}^{(0)}]$. We estimate this level's residual forcing again by the least squares. More levels are added the same way, until the $L^{\text{th}}$ level's residual, $\mathbf{r}^{(L+1)}$, becomes white in time, and its lag-0 correlation matrix converges to constant, hence

$$\begin{aligned}
\mathrm{d}r_i^{(0)} &= \mathbf{b}_i^{(1)}[\mathbf{x}, \mathbf{r}^{(0)}]\mathrm{d}t + r_i^{(1)}\mathrm{d}t, \\
\mathrm{d}r_i^{(1)} &= \mathbf{b}_i^{(2)}[\mathbf{x}, \mathbf{r}^{(0)}, \mathbf{r}^{(1)}]\mathrm{d}t + r_i^{(2)}\mathrm{d}t, \\
&\cdots \\
\mathrm{d}r_i^{(L)} &= \mathbf{b}_i^{(L+1)}[\mathbf{x}, \mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(L)}]\mathrm{d}t + r_i^{(L+1)}\mathrm{d}t \qquad (5)
\end{aligned}$$

The eqs. (4) and (5) describe a wide variety of processes in a fashion that explicitly accounts for the modeled process $\mathbf{x}$ feeding back on the noise statistics. The linear multilevel model is obtained by assuming $\mathbf{A}_i \equiv 0$ and $\mathbf{c}^{(0)} \equiv 0$ in eq. (4). Details of the methodology and further discussion could be found in [7].

It is well known, that the extreme ENSO events tend to occur in boreal winter. From several ways to include this phase locking to the annual cycle, the alternative approach used here is to include seasonal dependence in the dynamical part of the first level. Namely, we assume the matrix $\mathbf{B}^{(0)}$ and vector $\mathbf{c}^{(0)}$ to be periodic, with period $T = 12$ months:

$$\begin{aligned}
\mathbf{B}^{(0)} &= \mathbf{B}_0 + \mathbf{B}_s \sin(2\pi t/T) + \mathbf{B}_c \cos(2\pi t/T), \\
\mathbf{c}^{(0)} &= \mathbf{c}_0 + \mathbf{c}_s \sin(2\pi t/T) + \mathbf{c}_c \cos(2\pi t/T) \qquad (6)
\end{aligned}$$

In this case, the whole record is used to estimate four seasonal-dependent coefficients. The model is trained in the leading EOF (empirical orthogonal function) space [16] of tropical Pacific SST anomalies. The optimal number of state-vector components and the degree of nonlinearity has to be assessed by cross-validation. The parameters in this paper were used as in [7].

## 3 Results

In this section, the brief results are presented of how the statistical model is able to simulate the ENSO signal. The skill of the model is determined in the sense of basic linear ENSO metrics such as the amplitude of the ENSO signal, the seasonality (since the seasonality is important aspect of ENSO dynamics) and finally, the power spectrum of ENSO signal. The model is employed as described in the previous section, the matrices and vectors are estimated from the previous data and then the model is integrated to obtain the time series of same length as the training data. Since the model is stochastic (forced by a white noise), we employed an ensemble of 20 members. Each member is integrated with slightly different initial conditions and these members are referred to as realizations.

The basic ENSO metric is its amplitude, which could be characterized by the standard deviation of SST anomalies averaged over Nino3.4 box (bounded by 5°S - 5°N and 120°W - 170°W). In Fig. 2 we can see the ENSO amplitude as derived from the Nino3.4 index [20] (thick black line), along with 20 realizations from the data-based ENSO model, both linear and quadratic (gold for linear, red for quadratic).

As can be seen, the linear model slightly overestimates the ENSO amplitude, while the quadratic model slightly underestimate the ENSO amplitude. From the spread of the ensemble members we could infer that the model is sensitive to initial conditions and the forcing. Still, the ensemble averages for both models are within reasonable distance from the data borderline, therefore in this aspect the model performs adequately.

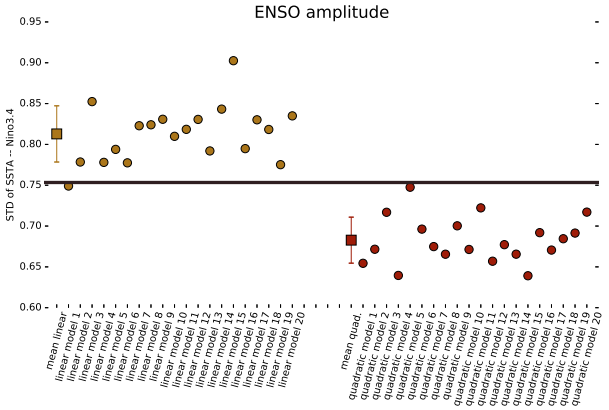


Figure 2: ENSO amplitude as standard deviation of SST anomalies in data (black line) and in 20 realizations of linear (gold) and quadratic (red) model.

Other metric connected with ENSO amplitude is its seasonality. As written above, the ENSO phenomenon exhibits seasonal changes in variance, with elevated variance in winter months and lower variance in spring and summer months. This can be also seen in Fig. 3, where the monthly variance is plotted for the data and for both models. Both models are capable of modelling higher variance in winter months and drop in variance through spring and summer, although the difference in variance is higher in data than in both models. Still, the ensemble averages are reasonably close to the data.

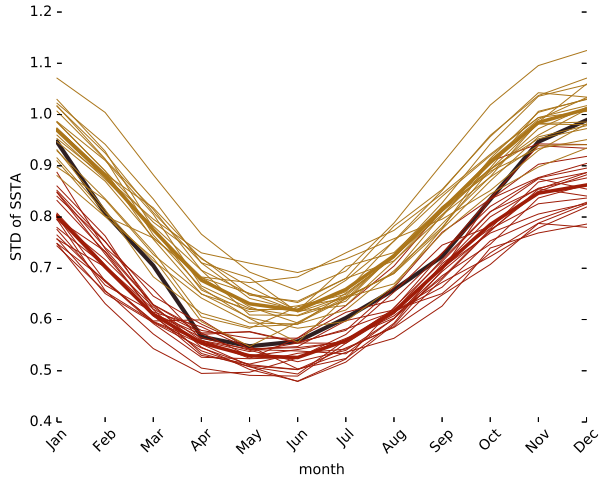


Figure 3: ENSO seasonality as standard deviation per month in data (black curve) and in 20 realizations of linear (gold) and quadratic (red) model. Thicker lines represent the mean over 20 realizations in the respective model.

The last metric taken into account was the power spectrum of Nino3.4 time series. The spectrum for the Nino3.4 data and both linear and quadratic model realizations can be seen in Fig. 4. The main peak in data occurs at roughly 5 year period, but still the ensemble averages for respective models are more flat in this area of frequencies. In the higher frequencies (around annual frequency and less) the power spectra are in agreement. In general, the spectra of modelled time series could be said to copy the actual Nino3.4 time series. The power spectra were computed using the Welch method [21].

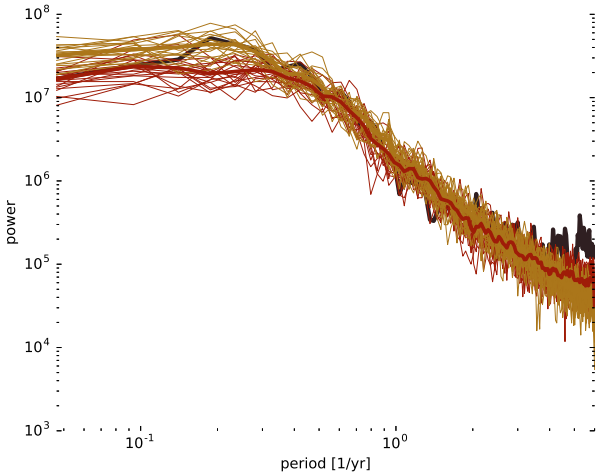


Figure 4: ENSO power spectra estimated using the Welch method in data (black curve) and in 20 realizations of linear (gold) and quadratic (red) model. Thicker lines represent the mean over 20 realizations in the respective model.

## 4 Noise parametrization in the model

The statistical model, once estimated, is integrated forward in time and forced by a noise - usually a realization of spatially correlated random process. In the most intuitive and basic case, the last level residuals' covariance matrix is estimated and decomposed using Cholesky factorization yielding a lower triangular matrix **R**. When the model is integrated, the random realization of white noise is multiplied by the matrix **R**, yielding spatially correlated white noise which is used as a random forcing in the model. The results for quadratic and linear ENSO models from the previous section were obtained using this simple noise parametrization, and the question is whether looking deeper into the residuals' structure could aid the model's performance.

### 4.1 Dependence on the system's state

First refinement for the noise parametrization arises from the concept of modelling climate processes which exhibit low-frequency variability (LFV). In this method, we find and select noise samples, snippets, from the past noise (residuals) which have forced the system during short time intervals that resemble the LFV phase just preceding the currently observed state, and then use these snippets (or information contained in them) to drive the current state into the future. For full methodology and discussion, see [22].

The found past noise snippets can be used in two different ways. The first one (as used in [22]) seeks various snippets from the past observations and then directly uses them to force the model as an ensemble. When e.g. we find 4 intervals which resemble the LFV phase, we integrate the model 4 times using all 4 noise snippets directly and than average over them. The second version (as used in our study) is to find, say, 100 samples of the past noise closest to the current state of the system, cluster them together and create covariance matrix from them. Afterwards, the Cholesky decomposition is used to obtain the matrix $\mathbf{R}$ and finally, the random white noise realization is multiplied by the matrix $\mathbf{R}$. Using this matrix, the spatial covariance of the forcing is dependent on the current state of the system. In both noise parametrizations, the current system state could be estimated in multiple ways: either using correlation of the SSA time series, or using the Euclidean distance in the subspace spanned by first few EOFs.

As can be seen in Fig. 5, although the amplitude statistics are not substantially shifted, the transient from high-variance winter period to low-variance spring and summer are better captured by the later model, with noise forcing conditioned on system's state. The power spectra for both models are practically the same (not shown).

### 4.2 Seasonal dependence of the forcing

Although the seasonal dependence of the model is captured in model's dynamics by fitting the seasonally dependent matrices $\mathbf{B}^{(0)}$ and $\mathbf{c}^{(0)}$ (recall eq. (6)), our analysis showed, that the last level's residuals still exhibit seasonally dependent amplitude. To address this issue, we computed the standard deviations for each month from the last level's residuals, then fitted the 5 harmonics of the annual cycle to capture the seasonal dependence, removed this dependence from the residuals, then estimated covariance matrix and subsequently the matrix $\mathbf{R}$ and finally generated spatially correlated white noise realization which was multiplied back by the requisite seasonal amplitude to account for the seasonally dependent amplitude of the forcing. The fitted harmonics of the annual cycle were selected as

$$P_i = \cos(2\pi it/T) + \sin(2\pi it/T),\ i = 1,\ldots,5 \qquad (7)$$

and then regressed on the seasonally varying standard deviation of the last level's residuals.

### 4.3 Using extended covariance matrix

The last modification to the noise is to use the extended covariance matrix instead of lag-0 covariance matrix. When evaluating system's state we do not take just the state closest to the current state of the model, but, say 5 consecutive months and construct the extended matrix out of this snippet. Then the matrix is decomposed using Cholesky factorization and used as a spatial correlation matrix $\mathbf{R}$ is random forcing generation.

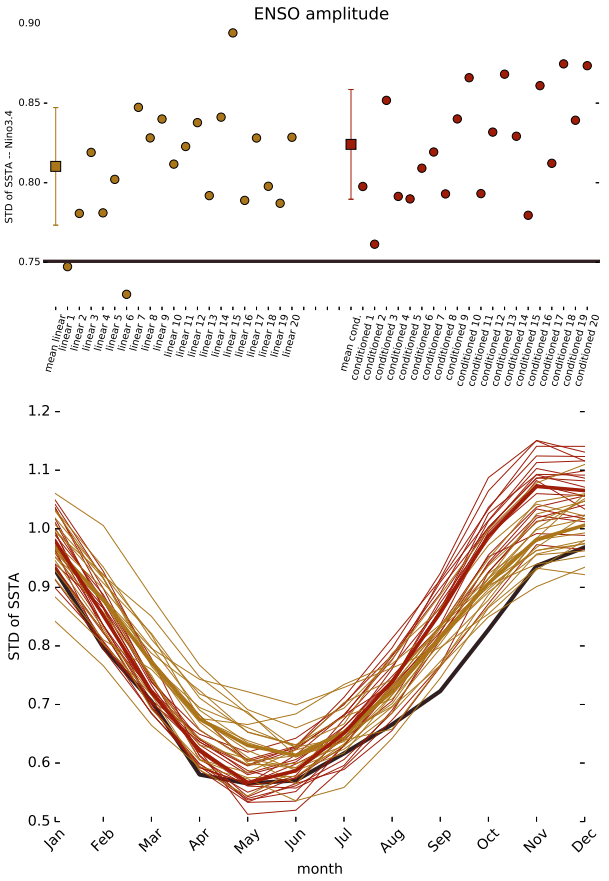


Figure 5: ENSO amplitude (upper) and seasonality (bottom) in data (black curve) and in 20 realizations of linear (gold) and linear with conditioned noise on the system's state (red) model. Thicker lines represent the mean over 20 realizations in the respective model.

The two latter modifications bring just a slight improvements into ENSO metrics (not shown), but could have more substantial advancements in modelling different atmospheric phenomena.

## 5 Synchronization and causality in the observed and modelled data

Better understanding of the complex dynamics of the atmosphere and climate is one of the challenges for contemporary science. Considering the climate system as a complex network of interacting subsystems [23] is a new paradigm bringing new data analysis methods helping to detect, describe and predict atmospheric phenomena [24]. A crucial step in constructing climate networks is inference of network links between climate subsystem [25]. Directed links determine which subsystems influence other subsystems, i.e. uncover the drivers of atmospheric phenomena. Inference of causal relationships from climate data is an intensively developing research

field, e.g. [26, 27]. Typically, a causal relation is sought between different variables or modes of atmospheric variability.

Paluš [28] has open another view at the complexity of atmospheric dynamics by uncovering causal relations or information flow between dynamics on different time scales in the same variable. Recently, phase-phase and also phase-amplitude interactions between dynamics on different temporal scales were observed in the ENSO dynamics (captured by the Nino3.4 index) using the approach as in [28]. Shortly, we use the continuous wavelet transform to the time series for particular time scales to obtain the instantaneous phase and amplitude of the oscillatory mode as

$$\psi(t) = s(t) + i\hat{s}(t) = A(t)e^{(i\phi(t))}, \quad (8)$$

$$\phi(t) = \arctan\frac{\hat{s}(t)}{s(t)}, \quad (9)$$

$$A(t) = \sqrt{s^2(t) + \hat{s}^2(t)}. \quad (10)$$

Then the time series of phase and / or amplitude are used to study the interactions. We adopt measures from information theory, namely mutual information and conditional mutual information, where the mutual information could be expressed as

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \quad (11)$$

where $p(\cdot)$ is the probability distribution or joint probability distribution and $X$ and $Y$ are our time series of either phase or amplitude derived from the ENSO SST data. Finally, the measures we are interested in could be written as:

- phase synchronization – $I(\phi_1(t);\phi_2(t))$,
- phase-phase causality – $I(\phi_1(t);\phi_2(t+\tau) - \phi_2(t)|\phi_2(t))$,
- phase-amplitude causality – $I(\phi_1(t);A_2(t+\tau)|A_2(t),A_2(t-\eta),A_2(t-2\eta))$,

### 5.1 Interactions in the data

As can be seen from Fig. 6, in ENSO dynamics captured by the Nino3.4 index, the synchronization of annual cycle with quasi-biennal and combination frequencies (frequencies that arise from the interactions between annual and the most prominent ENSO period) is observed. Also, the 4-6 year cycle of phase in ENSO dynamics influence the quasi-biennal range of the amplitude time series.

### 5.2 Interactions in the model

Our goal was to simulate the nonlinear cross-scale interactions in the model. This is important since it might help

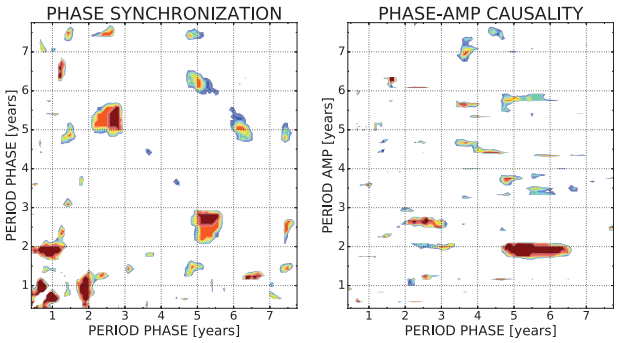


Figure 6: Phase synchronization (left) and phase-amplitude causality (right) in Nino3.4 time series. Shown is the significance (over 95$^{th}$ percentile against 500 Fourier transform surrogates) of k-nearest neighbours estimate of mutual information and conditional mutual information.

to uncover the mechanisms of these interactions and shed more light onto the dynamics of ENSO in general. We constructed the ENSO model and repeated the above analysis to modeled ensemble of the Nino3.4 time series.

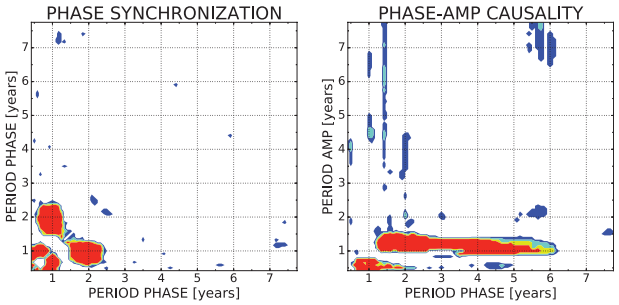


Figure 7: Phase synchronization (left) and phase-amplitude causality (right) in modeled Nino3.4 time series by the data-based model. Shown is the aggregate of 5 realizations of k-nearest neighbours estimate of mutual information and conditional mutual information. Significance against 500 Fourier transform surrogate data.

As seen from the analysis of modelled data (Fig. 7), the main phase synchronization bands (annual cycle with quasi-biennal cycle and combination frequencies) are also captured by the modelled data, while the phase - amplitude interactions are not very well captured. This might arise from the low complexity of the model, or the absence of some nonlinear interactions in the model design (apart from quadratic).

## 6 Modelling surrogate data with statistical model

Surrogate data (or analogous data) is a method to generate synthetic data set (time series) that preserve some of the statistical properties, while omitting the others. One way

of using them, is to test statistical significance by contradiction. This involves posing a null hypothesis describing some kind of a process and then generating an ensemble of surrogate data according to null hypothesis using Monte Carlo methods. One of the most used technique for generating surrogate data is the Fourier transform surrogate [29] (FT surrogates), which preserve the linear correlations in the data (periodogram or spectrogram, including autocorrelation) of the time series, but omits any other interactions in them.

As an example, consider two intertwined Lorenz systems, where one of them drives the other. Now, using the time series in one dimension, say the x dimension from both Lorenz systems, we can use some method for detecting causality, e.g. conditional mutual information between the two time series of two Lorenz systems. We get the value of conditional mutual information, but this is still not enough to interpret it in the means of whether there is a causal relationship between them or the result arose by chance. For this purpose, we construct an ensemble of Fourier transform surrogate data (which qualitatively preserves properties of the time series, but allows no causal relationship between them) and repeat the analysis using the very same method on this ensemble and finally compare the value for actual data with the histogram of values obtained from the ensemble of surrogate data. When the value from the data exceeds some percentile (e.g. $95^{\text{th}}$) of the surrogate data distribution, we say that the causal relationship is significant in comparison with e.g. 500 FT surrogates.

When studying nonlinear cross-scale interactions in time series using the above method, the statistical test involves creating an ensemble of surrogate, synthetic time series and repeat the analysis for the whole ensemble. Then we computed the percentile, where the observed interactions could not arose by random chance. Of course, one could use Fourier transform method to generate the surrogate time series, effectively posing a null hypothesis of a linear process which has the same spectrum to that of an observed data. On the other hand, one can create a more sophisticated null hypothesis by exploiting the options of a data-based model: when one consider just a linear model, omit the dynamical seasonal dependence in $\mathbf{B}^{(0)}$ and $\mathbf{c}^{(0)}$ terms (as in eq. (6)) and use the simplest noise parametrisation (just consider the spatial covariance structure), the model will omit the nonlinear interactions and could pose as a surrogate data model copying the basic statistical properties of a modelled time series. This way, the analysis would show whether the cross-scale interactions are arising from the seasonal dependent dynamics, or from nonlinear (e.g. quadratic) interactions between subsystems and so on.

When comparing Fig. 6 (testing against 500 Fourier transform surrogates) and Fig. 8 (testing against 500 data-based model surrogates), the significant interactions are virtually the same, expect in the latter, the "fluctuations" (or they might be false positives as well) are attenuated to minimum. This way, we can get better idea of the statistical significance of the interactions between subsystems, in particular the nonlinear ones, since we are testing against the model with just linear interactions.

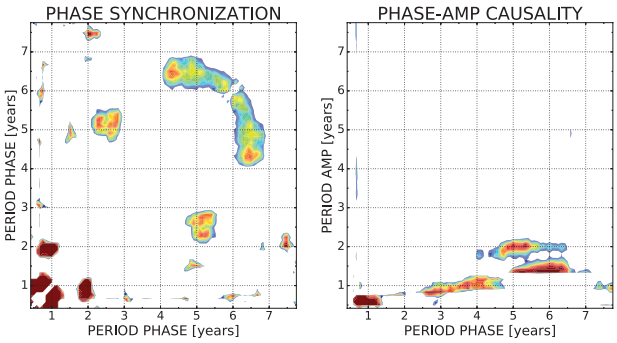


Figure 8: Phase synchronization (left) and phase-amplitude causality (right) in modelled Nino3.4 time series by the data-based model. Shown is the aggregate of 5 realizations of k-nearest neighbours estimate of mutual information and conditional mutual information. Significance against 500 surrogate time series created with data-based model.

## 7 Conclusions

Statistical modelling in climate science is continuously getting more attention, since their usage is not limited to forecast some of the phenomena of interest (like ENSO), but could also be used to infer some of the statistical properties and relationships among different subsystems. Since the statistical models live in phase space of particularly reduced dimensionality, when we could observe the interactions of interest, the identification of their sources will become more feasible.

We showed that the statistical model with the right settings, which were selected based on careful inspection of the modelled system, could generate synthetic time series of interest, copying the desired properties of the system - both linear and nonlinear statistics. Since the stochasticity is the important aspect of the data-based model, various parametrization techniques exist to correctly model the system's external forcing. Finally, the possibility of usage of the low complexity model as surrogate data was discussed, showing advantages of usage of such technique to infer statistical significance.

The outlook for future work combines various different paths which appeared. One direction would be focusing on statistical modelling itself, experimenting with various variable model, with input time series and their preprocessing and so on and so forth. Other direction would be connecting the statistical models with dynamical ones, in the sense, that statistical models could be used for parametrization of e.g. sub-grid phenomena (microphysics of clouds, local convection etc.) in large coupled atmospheric-oceanic models.

## References


[1] Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor: The WCRP CMIP3 multi-model dataset: A new era in climate change research. *A Bull. Amer. Meteor. Soc.* **88** (2007) 1383–1394

[2] Taylor, K.E., R.J. Stouffer and G.A. Meehl: An Overview of CMIP5 and the experiment design. *A Bull. Amer. Meteor. Soc.* **93** (2012) 485–498

[3] Bjerknes, V.: Dynamic meteorology and hydrology, Part II. Kinematics. *Gibson Bros.*, Carnegie Institute, New York. (1911)

[4] Lorenz, E. N.: Deterministic nonperiodic flow. *J. Atmos. Sci.* **20** (1963) 130–141

[5] Van den Dool, H. M.: Long-range weather forecasts through numerical and empirical methods. *Dyn. Atmos. Oceans* **20** (1994) 247–270

[6] Penland, C.: Random forcing and forecasting using principal oscillation pattern analysis. *Mon. Weat. Rev.* **117** (1989) 2165–2185

[7] Kravtsov, S., D. Kondrashov, and M. Ghil: Multilevel regression modeling of nonlinear processes: Derivation and applications to climate variability. *J. Climate* **18** (2005) 4404—4424

[8] Philander, S. G. H.: El Niño, La Niña, and the Southern Oscillation. *Academic Press* (1990)

[9] Alexander, M. A., I. Bladé, M. Newman, J. R. Lanzante, N.-C. Lau, and J. D. Scott: The atmospheric bridge: The influence of ENSO teleconnections on air–sea interaction over the global oceans. *J. Climate* **15** (2002) 2205–2231

[10] WIKIPEDIA.ORG: El Niño–Southern Oscillation `https://en.wikipedia.org/wiki/El_Ni%C3%B1o%E2%80%93Southern_Oscillation`, downloaded June 27, 2016.

[11] Ghil, M. and A. W. Robertson: Solving problems with GCMs: General circulation models and their role in the climate modeling hierarchy. *Academic Press*, (2000) 285–325

[12] Hannachi, A., D. B. Stephenson, and K. R. Sperber: Probability-based methods for quanfifying nonlinearity in the ENSO. *Clim. Dyn.* **20** (2003) 241–256

[13] Ghil, M. and N. Jiang: Recent forecast skill for the El Niño / Southern Oscillation. *Geophys. Res. Lett.* **25** (1998) 171-–174

[14] Timmermann, A., H. U. Voss, and R. Pasmanter: Empirical dynamical system modeling of ENSO using nonlinear inverse techniques. *J. Phys. Oceanogr.* **31** (2001) 1579–1598

[15] Wetherill, G. B.: Regression Analysis with Applications. *Chapman and Hall* (1986)

[16] Hannachi, A., I. T. Jolliffe and D. B. Stephenson: Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int. J. Climatol.* **27** (2007) 1119–1152

[17] Penland, C. and M. Ghil: Forecasting Northern Hemisphere 700-mb geopotential height anomalies using empirical normal modes. *Mon. Wea. Rev.* **121** (1993) 2355—2372

[18] Kondrashov, D., S. Kravtsov, A. W. Robertson and M. Ghil: A Hierarchy of Data-Based ENSO Models. *J. Climate* **18** (2005) 4425–4444

[19] McCullagh, P., and J. A. Nelder: Generalized Linear Models. *Chapman and Hall* (1989)

[20] Rayner N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent and A. Kaplan: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108** (2003) 4407

[21] Welch, P. D.: The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio* **AU-15** (1967) 70–73

[22] Chekroun, M. D., D. Kondrashov and M. Ghil: Predicting stochastic systems by noise sampling, and application to the El Niño-Southern Oscillation. *P. Natl. Acad. Sci. USA* **108** (2011) 11766–11771

[23] A. A. Tsonis and P. J. Roebber: The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications*, **333** (2004) 497–504

[24] S. Havlin, D. Y. Kenett, E. Ben-Jacob, A. Bunde, R. Cohen, H. Hermann, J. Kantelhardt, J. Kertész, S. Kirkpatrick, J. Kurths, et al.: Challenges in network science: Applications to infrastructures, climate, social systems and economics. *The European Physical Journal Special Topics*, **214** (2012) 273–293

[25] M. Paluš, D. Hartman, J. Hlinka, and M. Vejmelka: Discerning connectivity from dynamics in climate networks. *Nonlinear Processes in Geophysics* **18** (2011) 751–763

[26] Ebert-Uphoff and Y. Deng: Causal discovery for climate research using graphical models. *Journal of Climate* **25** (2012) 5648–5665

[27] Y. Deng and I. Ebert-Uphoff: Weakening of atmospheric information flow in a warming climate in the community climate system model. *Geophysical Research Letters* **41** (2014) 193–200

[28] Paluš, M.: Multiscale atmospheric dynamics: Cross-frequency phase-amplitude coupling in the air temperature. *Phys. Rev. Lett.* **112** (2014) 1–5

[29] Theiler, J., S. Eubank, A. Longtin, B. Galdrikian, and J. Doyne Farmer: Testing for nonlinearity in time series: The method of surrogate data. *Physica D* **58** (1992) 77–94