Improving the Long-Lead Predictability of El Niño Using a Novel Forecasting Scheme Based on a Dynamic Components Model

Desislava Petrova · Siem Jan Koopman · Joan Ballester · Xavier Rodó

Received: date / Accepted: date

Desislava Petrova
Climate Dynamics and Impacts Unit, Catalan Institute for Climate Science (IC3)
Carrer del Dr. Trueta, 203, 08005 Barcelona, Catalonia, Spain
Department of Physics, University of Barcelona
Tel.: +34-693515762
Fax: +34-935-679977
E-mail: desislava.petrova@ic3.cat

Siem Jan Koopman
Department of Econometrics, VU University Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
Tel.: +31-20-5986019
Fax: +31-20-5986020
E-mail: s.j.koopman@vu.nl

Joan Ballester
Climate Dynamics and Impacts Unit, Catalan Institute for Climate Science (IC3)
Carrer del Dr. Trueta, 203, 08005 Barcelona, Catalonia, Spain
California Institute of Technology (Caltech), Pasadena, California, United States
Tel.: +34-935-679977
Fax: +34-935-679977
E-mail: joanballester.ic3@gmail.com

Xavier Rodó
Climate Dynamics and Impacts Unit, Catalan Institute for Climate Science (IC3)
Carrer del Dr. Trueta, 203, 08005 Barcelona, Catalonia, Spain
Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
Tel.: +34-935-679977
Fax: +34-935-679977
E-mail: xavier.rodo@ic3.cat
Abstract  El Niño (EN) is a dominant feature of climate variability on inter-annual time scales driving changes in the climate throughout the globe, and having wide-spread natural and socio-economic consequences. In this sense, its forecast is an important task, and predictions are issued on a regular basis by a wide array of prediction schemes and climate centres around the world. This study explores a novel method for EN forecasting. In the state-of-the-art the advantageous statistical technique of Structural (Unobserved Components) Time Series has not been applied. Therefore, we have developed such a model with regression parameters obtained by a State Space approach. Its distinguishing feature is that observations consist of various components - level, seasonality, cycles, disturbance, and regression variables incorporated as explanatory covariates. These components are modeled separately and ultimately combined in a single forecasting scheme.

Customary statistical models for EN prediction essentially use SST and wind stress in the equatorial Pacific. In addition to these, we introduce a new domain of regression variables accounting for the state of the subsurface ocean temperature in the western and central equatorial Pacific, motivated by our analysis, as well as by recent and classical research, showing that subsurface processes and heat accumulation there are fundamental for the genesis of El Niño. An important feature of the scheme is that different regression predictors are used at different lead months, thus capturing the dynamical evolution of the system and rendering more efficient forecasts.

The new model has been tested with the prediction of all warm events that occurred in the period 1996-2015. Retrospective forecasts of these events were made for long lead times of at least two and a half years. Hence, the present study demonstrates that the theoretical limit of ENSO prediction should be sought much longer than the commonly accepted "Spring Barrier". The high correspondence between the forecasts and observations indicates that the proposed model outperforms all current operational statistical models, and behaves comparably to the best dynamical models used for EN prediction. Thus, the novel way in which the modeling scheme has been structured could also be used for improving other statistical and dynamical modeling systems.

Keywords  El Niño Southern Oscillation · prediction · predictability · subsurface dynamics · time series
1 Introduction

The equatorial Pacific region, and especially the western equatorial Pacific (WPAC), is fundamental for the modulation of the weather patterns worldwide through changes in the surface-boundary condition and the thermal heating of the atmosphere, while the signal of El Niño Southern Oscillation (ENSO) variability in this region is the strongest after the seasonal cycle (Sarachik and Cane, 2010). Therefore, it is not surprising that ENSO is a subject of intense scientific research and schemes for its prediction are plentiful with various rates of success (Barnston et al., 2012). At the same time, the limitations of its predictability are debated, with some studies arguing that high-frequency atmospheric noise prevents accurate forecasts to be made at long lead times (Penland, 1996), (Thompson and Battisti, 2000), (Thompson and Battisti, 2001), (Fedorov et al., 2003). Others, instead, imply that EN is to a larger extent the result of an internal self-sustained dynamics, which allows for its prediction well ahead of its typical December peak (Zebiak and Cane, 1987), (Goswami and Shukla, 1991), (Jin et al., 1994), (Tziperman et al., 1994), (Chen et al., 2004).

It is interesting that despite the major improvements of dynamical models, initialization techniques, and the longer and more accurate data sets available for statistical predictions, verification of the real-time ENSO prediction skills during 2002-11 indicates skills somewhat lower than those found for the less advanced models of the 1980s and 1990s (Barnston et al., 2012). In addition, the so-called spring barrier continues to represent a major difficulty for all types of ENSO forecasting models, and their skill for predictions made in the months March, April, May is still non-significant. A potential explanation could lie in the seasonal cycle, because during boreal spring the tropical Pacific is warmest and this seasonal warming is erroneously taken by the models for the growing phase of EN (Sarachik and Cane, 2010). Moreover, the low signal-to-noise ratio in spring also accounts for the lower prediction skill of model forecasts issued in these months (Torrence and Webster, 1998).

Thus, advances in the understanding of the dynamical mechanisms involved in the initiation of El Niño and the transition to La Niña, as well as the introduction of new and improved ENSO prediction models remain necessary.

The majority of the existing statistical forecasting schemes are based on sea surface temperature, sea level pressure, and wind stress data in the equatorial Pacific (Barnston and Ropelewski, 1992), (Penland and Magorian, 1993), (Xue et al., 1994), (Kondrashov et al., 2005). Data sets directly capturing subsurface temperature changes in the tropical Pacific region, however, have not been incorporated in the existing empirical models. As three-dimensional observations of the tropical Pacific (TAO-TRITON data) are only available since the 1990s, and as the statistical models used for operational ENSO forecasting have remained predominantly unchanged over the last 10-15 years, the subsurface indicators that play a fundamental role in ENSO variability have not been included in the existing forecasting skills (Barnston et al., 2012). At the same time, several studies have already highlighted the significance of the accumulation of heat in the subsurface of the western tropical Pacific for the onset of EN (Ramesh and Murtugudde, 2013), (Ballester et al., 2015a).

(Wyrtki, 1975) has proposed a theory, later further developed in the dynamical recharge-discharge oscillatory theory by (Jin, 1997a), (Jin, 1997b), according to which an intensification of the trade winds in the central equatorial Pacific leads to the accumulation of heat in the western tropical Pacific subsurface. Then, with the relaxation of the trade winds and the occurrence of westerly anomalies in the western/central tropical Pacific, this heat is released and allowed to move to the east during the growing phase of EN. What has not been explicitly discussed in this theory, though, is that the propagation of the warm waters/heat towards the east occurs mainly in the subsurface near the thermocline (Ramesh and Murtugudde, 2013), (Ballester et al., 2015a), and warm anomalies appear on the surface in the eastern equatorial Pacific only at a later stage of this growing phase. Moreover, warm anomalies in the subsurface (down to 400-500 metres depth) occur on average as early in the WPAC as 26-30 months prior to the peak of El Niño, and are present there at different levels of depth up to 10-11 months before the peak, when the heat starts to be advected eastwards.

These patterns at longer lags demonstrate that subsurface temperature and surface wind stress indices in the WPAC and the central Pacific (CPAC) could be promising as regression predictor variables for EN. At the same time, sea surface temperature anomalies (SSTAs) there exist at lags 15-17 months, but they are short-lived and not as strong in magnitude as the subsurface anomalies, so their role as predictors could be more limited.

Considering the above, in the present study we propose a statistical unobserved component time series model
for El Niño forecasting based on wind stress, surface and subsurface ocean temperatures in the western and central tropical Pacific. The lead-lag times for these variables are between 30 and 0 months prior to the peak of EN, and as it is customary, predictions are made for the SSTs in the Niño 3.4 region (5° N - 5° S, 170° W - 120° W, (Barnston et al., 1997), black time series in Figure 1(a)). Section 2 of this manuscript describes the data sets and statistical methods used for the analysis; in Section 3 we give a detailed explanation of the proposed ENSO modeling technique, while in Section 4 we clarify how the model is applied; in Section 5 we present some forecasting results, and then we end with a discussion of the results in Section 6, and conclusions in Section 7.

2 Data and Statistical Analysis

Wind stress data has been derived from the NCEP/NCAR reanalysis (Kalnay et al., 1996). Sea surface temperature is the NOAA-ERSST-V3 before 1982 and the NOAA-OI-SST-V2 data afterwards, provided by the NOAA/OAR/ESRL PSD (www.esrl.noaa.gov/psd/). Subsurface temperature data before 2012 is from the Subsurface Temperature And Salinity Analyses by (Ishii et al., 2005), archived at the National Center for Atmospheric Research, Computational and Information Systems Laboratory (www.rda.ucar.edu/datasets/ds285.3/), and from the Hadley Centre EN4.0.2 analyses data (Good et al., 2013) afterwards. The EN events used for the composite anomalies are based on the definition of the Climate Prediction Center for the period 1978-2012: December 1982, 1991, 1997, 2002, 2006, 2009 (CPC, 2015). The 1986–87 event was excluded from the composite calculations because the peak of this event was in September–October rather than in December–January. However, calculations with this event included do not produce significantly different results, and the temperature and zonal wind anomalies at the respective lag times are qualitatively the same (not shown).

A Multi Taper Method (MTM) was used for the spectrum density estimation and signal reconstruction of the N3.4 time series shown in Figure 2. MTM is a nonparametric method that reduces the variance of spectral estimates through the use of a small set of orthogonal tapers, which are multiplied by the data to minimize spectral leakage (Ghil et al., 2002).

A recursive Butterworth procedure was used (Moron and Plaut, 2003), (Ballester et al., 2011) to filter SST and windstress data, so that only frequencies corresponding to periods of 14-18 months (Figure 3), 24-28 months (Figure 4), and 46-63 months (Figure 5) have been kept. A Butterworth filter was also applied to remove high-frequency variability and thus filter out the annual cycle in the composites of sea surface and subsurface temperatures, and windstress anomalies in Figures 6, 7 and 8.

Complex Orthogonal Function (CEOF) analyses was applied to the area-weighted filtered SSTAs and windstress anomalies in the equatorial Pacific (Figures 3, 4 and 5). This technique decomposes variability into real and imaginary spatial patterns that are amplified by real and imaginary time-varying coefficients, respectively. The patterns represent the main modes of variability of the data as a function of the phase (φ) in periodic temporal coefficients and periodic temporal scores (Ballester et al., 2011). The temporal scores were correlated with the filtered spatio-temporal SSTAs and windstress anomalies in the tropical Pacific area.

3 Model

Structural time series models are formulated in terms of components of interest that could be directly interpreted (Harvey, 1989). Moreover, time series models based on unobserved components are particularly effective when complex features are present in the time series, such as mixed frequencies (various modes of variability of the time series), outliers, structural breaks and nonlinear non-Gaussian processes (Tong, 1990). In fact, they could also be easily linked to a regression model. In this way, a structural model provides a basis for making predictions of future observations, as well as a description of the salient features of a given time series (Harvey and Shephard, 1993). These features allow for the great flexibility of the approach, suitable to cope with nonlinearities (Durbin and Koopman, 2001), common in the climate system and in ENSO in particular. The linear univariate unobserved components time series model that we consider is given by
\[ y_t = \mu_t + \gamma_t + \psi_{1t} + \psi_{2t} + \psi_{3t} + x^t \delta + \varepsilon_t \] (1)

where \( y_t \) represents observations on a single variable \( y \), in this case the monthly N3.4 index at time \( t \); \( \mu_t \) is a level component that we specify as a random walk process; \( \gamma_t \) is a seasonal component with seasonal length \( S = 12 \); \( \psi_{1t} \), \( \psi_{2t} \) and \( \psi_{3t} \) are three cycle components with different parameter values for frequency \( \lambda \), persistence \( \varphi_\psi \) and variance \( \sigma^2_\psi \); \( x^t \delta \) represents a predictor regression variable; and \( \varepsilon_t \) is the remainder irregular term. The trend, seasonal, and cycle components are modeled by linear dynamic stochastic processes, which depend on disturbances (Harvey and Koopman, 2000). For more information the reader is referred to (Harvey, 1989) and (Durbin and Koopman, 2001). The components are formulated in a flexible way and they are allowed to change over time. The disturbances driving the components are independent of each other.

The model in state space form (Durbin and Koopman, 2001) is based on the state and disturbance vectors as given by

\[
\alpha_t = (\delta^t, \mu_t, \gamma_t, \gamma_{t-1}, \ldots, \gamma_{t-10}, \psi_{1t}, \psi_{2t}^+, \psi_{3t}, \psi_{3t}^+)' , \quad \epsilon_t = (\varepsilon_t, \eta_t, \omega_t, \kappa_{1t}, \kappa_{1t}^+, \kappa_{2t}, \kappa_{2t}^+, \kappa_{3t}, \kappa_{3t}^+)' ,
\]

where we assume that the dimension of the regression coefficient vector \( \delta \) is \( k \times 1 \), and the corresponding system matrices are given in Appendix A. They are time-invariant except the matrix \( Z_t \). The unknown disturbance variances need to be estimated together with the discount factor \( \varphi_{\psi,j} \) and cycle frequency \( \lambda_{c,j} \) for \( j = 1, 2, 3 \) using the Kalman Filter (Kalman, 1960). We further have five log-variances to estimate. We enforce restrictions \( 0 < \varphi_{\psi,j} < 1 \) and \( 0 < \lambda_{c,j} < \pi \) via transformations, for \( j = 1, 2, 3 \). The unknown parameters are collected in the \( 12 \times 1 \) parameter vector \( \theta \) that is given by

\[
\theta = (a_\varepsilon, a_\eta, a_\omega, a_{\kappa,1}, a_{\kappa,2}, a_{\kappa,3}, b_{\psi,1}, b_{\psi,2}, b_{\psi,3}, d_1, d_2, d_3)' ,
\]

where

\[
\sigma_\varepsilon = \exp(a_\varepsilon), \quad \varphi_{\psi,j} = \exp(b_{\psi,j})/[1 + \exp(b_{\psi,j})], \quad \lambda_{c,j} = 2\pi/[2 + \exp(d_j)],
\]

for \( \ell = \varepsilon, \eta, \omega, \{\kappa,j\} \) and \( j = 1, 2, 3 \). Parameter estimation in this model is achieved through maximum likelihood. Steady state is reached after weak convergence relative to 1e-007.

3.1 Level Component (Long-Term Variability), \( \mu_t \)

The level component is modeled as a random walk process and is given by

\[
\mu_{t+1} = \mu_t + \eta_t , \quad \eta_t \sim \text{NID}(0, \sigma^2_\eta),
\]

where \( \text{NID}(0, \sigma^2) \) refers to a normally independently distributed series with mean zero and variance \( \sigma^2 \). The disturbance series \( \eta_t \) is serially independent and mutually independent of all other disturbance series related to \( y_t \). The initial trend \( \mu_1 \) is for simplicity treated as an unknown coefficient that needs to be estimated together with the unknown variance \( \sigma^2_\eta \). This component is included in the model to account for the long-term fluctuations of the N3.4 time series around its mean, assuming that it is stationary (see Figure 1(a)). A simple interpretation is that the level at the current step is equal to the level in the previous step plus a white noise disturbance (Harvey, 1989).

3.2 Seasonal Component, \( \gamma_t \)

To account for the monthly variation in the annual cycle of the N3.4 time series, the component \( \gamma_t \) is included in the model. More specifically, \( \gamma_t \) represents the seasonal effect at time \( t \) that is associated with season \( s = s(t) \) for \( s = 1, \ldots, S \), where \( S \) is the seasonal length (\( S = 12 \) for monthly data). In particular, we adopt the notation

\[
\gamma_t = \gamma^{[s]}_t ,
\]

if we need to emphasize that \( \gamma_t \) represents the seasonal effect for season \( s \).
We use a seasonal pattern that is fixed over time (see Figure 1(b)), i.e. we have a set of $S$ seasonal effects $\gamma_1, \ldots, \gamma_S$ which are taken as unknown coefficients that need to be estimated together with the other coefficients in the model. The seasonal effects must have the property that they sum to zero over the full year to make sure that they are not confounded with the trend component, that is

$$\gamma_1 + \ldots + \gamma_S = 0, \quad \gamma_t = \gamma_{t-S}, \quad t = S + 1, \ldots, n. \quad (3)$$

The seasonal pattern could also change slowly over time, by relaxing the summing-to-zero requirement with a stochastic equation

$$\gamma_{t+1} = -\gamma_t - \ldots - \gamma_{t-S+2} + \omega_t, \quad \omega_t \sim \text{NID}(0, \sigma^2_\omega), \quad (4)$$

In the present study the disturbance variance $\omega_t = 0$. In this way we have $S-1$ unknown seasonal coefficients that need to be estimated by the Kalman Filter.

There is a marked seasonal cycle in the tropical Pacific, which has a substantial impact on the ENSO cycle and the evolution of its phases (Tziperman et al., 1997), (Krishnamurthy et al., 2015). This effect is known as phase-locking to the annual cycle (Rasmusson and Carpenter, 1982), (An and Choi, 2009), (Stein et al., 2011), as ENSO phases normally grow in the boreal summer and autumn and the peak of the associated anomaly is in the winter months of December, January and February (DJF) (Sarachik and Cane, 2010). Thus, including an annual cycle component in the forecasting model is necessary for the correct seasonal development of the signal and, therefore, for a more accurate prediction. It is in fact a fundamental part of the ENSO dynamics, especially as the atmospheric conditions should be appropriate in order for an initial surface and subsurface ocean temperature perturbation to grow and propagate. The seasonal variations of the atmospheric convergence zones modulate unstable ocean-atmosphere interactions (Philander, 1989), which in turn favour the development of EN or LN.

### 3.3 Cycle Components, $\psi_t$

In order to account for interannual variability present in the ENSO phenomenon (Figure 2), we further include a number of additional cycle components. A stochastic formulation of a cycle component can be based on a time-varying trigonometric process, but with frequency $\lambda_c$ associated with the typical length of a cycle. We have

$$\begin{pmatrix} \psi_{t+1} \\ \psi^+_{t+1} \end{pmatrix} = \varphi_\psi \begin{pmatrix} \cos \lambda_c & \sin \lambda_c \\ -\sin \lambda_c & \cos \lambda_c \end{pmatrix} \begin{pmatrix} \psi_t \\ \psi^+_t \end{pmatrix} + \begin{pmatrix} \kappa_t \\ \kappa^+_t \end{pmatrix}, \quad (5)$$

where the discount factor $0 < \varphi_\psi < 1$ is introduced to enforce a stationary process for the stochastic cycle component. The disturbances and the initial conditions for the cycle variables are given by

$$\begin{pmatrix} \kappa_t \\ \kappa^+_t \end{pmatrix} \sim \text{NID}(0, \sigma^2_\kappa I_2), \quad \begin{pmatrix} \psi_1 \\ \psi^+_1 \end{pmatrix} \sim \text{NID}(0, \frac{\sigma^2_\psi}{1-\varphi_\psi^2} I_2),$$

where the disturbances $\kappa_t$ and $\kappa^+_t$ are serially independent and mutually independent, also with respect to disturbances that are associated with other components. The coefficients $\varphi_\psi, \lambda_c$ and $\sigma^2_\kappa$ are unknown and need to be estimated together with the other parameters. This stochastic cycle specification is further discussed by (Harvey, 1989).

Generally, the ENSO oscillation is said to have a period of between two and seven years. As we already pointed out, seasonality has an important role in the evolution of the overall irregular cycle (Tziperman et al., 1997), (An and Choi, 2009). At the same time, several other inter-annual spectral peaks at different time scales can be noted in Figure 2(a), showing the power density spectrum of the entire N3.4 time series. It will be demonstrated in the next sections that the modes of variability that correspond to these spectral peaks are key not only for understanding ENSO, but also for its simulation.
Previous studies have identified one significant signal at a frequency close to the annual cycle, recognized as a separate coupled mode of variability of the equatorial Pacific ocean-atmosphere system (Zebiak, 1985), (Neelin, 1990), (Mantua and Battisti, 1995), (Jiang et al., 1995), (Fedorov and Philander, 2001), (Jin et al., 2003). Its relevance to the evolution of the ENSO cycle and its prediction has been demonstrated especially by (Jin et al., 2003), where this mode is called a “near-annual” mode and defined to have a period of 12-18 months. It is characterized by a westward propagation of SST and zonal wind stress anomalies from the eastern equatorial Pacific (EPAC) to the WPAC (Figure 3). These anomalies are similar to patterns of the mean annual cycle in the EPAC, but enhanced and propagating all the way to the western boundary (Mantua and Battisti, 1995), (Jin et al., 2003). Westward coupled modes of this kind, driven by near-surface advective processes (mostly anomalous zonal advection of mean temperature gradients), have also been identified in the tropical Pacific in the theoretical work of (Philander et al., 1984), (Gill, 1985), (Neelin, 1991), and some others. In the present model we too found that the period of one of the cycle components estimated with the Kalman Filter is approximately 1.5 years (16-18 months). The graphical representation of this cycle is shown in the time series in Figure 1(c).

This near-annual mode has been shown to be modulated both by the annual cycle and ENSO. It has been hypothesized by (Jiang et al., 1995) that it is produced through the nonlinear resonance of the low-frequency ENSO mode, which is discussed later, with the annual cycle. Then, its interaction with the main ENSO cycle and its phase-locking with the mean annual cycle result in fluctuations in its amplitude, which was low in the late 1980s, for example, and then increased in the second half of the 1990s (Jin et al., 2003) (Figure 1(c)). When its period is close to 12 months (i.e. to the annual cycle), phase- and frequency-locking likely occur and lead to the greater amplitude of the mode, and to near-annual LN-like events (as between 1996-2004, see Figure 1(c) and Figure 3(d), (e), (i) and (j)) (Jin et al., 2003). However, the continued long warm background period taking place in the tropical Pacific during 1985-1991 (due to ENSO variability, Figure 1(e)) has resulted in the lower frequency of this mode, reducing the opportunity for its non-linear phase-locking with the main oceanic Rossby mode (Mantua and Battisti, 1995), and leading to “mini El Niño events” that have occurred between all major EN events in that period (Jin et al., 2003) (Figure 1(c), Figure 3(a), (b), (f) and (g)).

Thus, this fast mode of variability has important implications for the evolution of ENSO, and hence, for the accurate prediction of EN and LN events. Therefore, its inclusion into the proposed prediction scheme could significantly improve the model performance and skill. In fact, it has been demonstrated by (Mantua and Battisti, 1995) that the irregularity of ENSO reproduced by the Zebiak-Cane (ZC) model (Zebiak, 1985) is due to a large extent to the presence of a “mobile mode” with similar characteristic features as the near-annual mode discussed here.

A number of studies have discussed a biannual/quasi-biannual (QB) (24-28 months period) component of ENSO variability (Trenberth, 1976), (Rasmusson and Carpenter, 1982), (Lau and Shen, 1988), (Yasunari, 1989), (Rasmusson et al., 1990), (Jiang et al., 1995), which has been detected in both the equatorial surface zonal wind and SSTs. (Rasmusson et al., 1990) discovered through an SVD analysis of the zonal winds in the tropics a biannual cycle characterized by consistent eastward-propagating anomalies across the WPAC and the CPAC. In the EPAC it was found to have a meridional propagation also affecting local SSTs. Moreover, a strong and close relationship between the biannual components of the zonal winds in the WPAC and the SST time series in the EPAC was demonstrated. From this analysis, the conclusion has been reached that all well-defined warm phases of ENSO coincide with a distinctive warm phase of the biannual mode.

Similarly to the fast-frequency component, the biannual oscillation also tends to be phase-locked to the annual cycle (Rasmusson et al., 1990). Typically associated with it are westerly wind anomalies in the WPAC (between 130° - 160°E) and north of the Equator in March-April of the first year of the cycle (Figure 4(f) and (g)). These patterns progress eastward and peak by September-October near the date line, moving south of the Equator by that time, testifying for a meridional displacement (Figure 4(h) and (i)). After October they propagate to...
the EPAC and begin to disappear. Then, by November easterlies start to form west of 130°E and north of the Equator. These develop in DJF and move towards the east (Figure 4(j)). By March-April of the second year a reversal of the wind pattern is at place.

The evolution of SSTs in the biannual mode, on the other hand, develops from near neutral anomalies in March-April (Figure 4(a)) to a typical ENSO-like structure in September-October of the first year of the cycle (Figure 4(b)). In the second year, the SSTA pattern is reversed and the opposite phase of the ENSO-like structure develops (Figure 4(d) and (e)). In this way, equatorial negative (positive) SSTAs in the CPAC and EPAC are preceded by easterly (westerly) surface zonal wind anomalies in the WPAC and CPAC (Figure 4).

Thus, the estimation of a period close to two years (between 26-30 months; Figure 1(d)) for the second cycle component of the proposed model corresponds to this mode of variability of the equatorial Pacific ocean-atmosphere system. Importantly, the phase evolution and amplitude of this cycle appear to be approximately regular (see Figure 1(d)). On the other hand, (Rasmusson et al., 1990) has noted that even though the cycle is regular, the amplitude and phase of the mode are subject to low-frequency changes, which will be discussed later in this manuscript. In any case, the QB mode has been shown to be a fundamental element of the ENSO dynamics, essential for its prediction (Rasmusson et al., 1990), (Jiang et al., 1993), (Jiang et al., 1995).

3.3.3 Low-Frequency Cycle Component, \( \psi_3 t \)

The main period of ENSO has been established to be between 4 and 5 years, which, as expected, corresponds to the most pronounced spike in the power spectral density of the N3.4 time series (Figure 2(a)). Moreover, as shown by (Jiang et al., 1995), the dominant mode of variability of SSTs and surface zonal winds in the tropical Pacific is a low-frequency oscillation with a period of approximately 46-63 months. It has been referred to in the literature as the low-frequency ENSO mode (Rasmusson et al., 1990), or as the quasi-quadrennial (QQ) mode of variability (Jiang et al., 1995).

For SST the QQ component represents an eastward propagating oscillation with maximum anomalies close to 120°W (Figure 5(b) and (e)). For the zonal wind it represents an oscillation with an eastward propagation in the WPAC and with maximum anomalies close to 170°E (Figure 5(f) and (i)). Actually, its spatial characteristics for both SST and zonal wind closely resemble the main spatial characteristics of the biannual mode (Figures 4 and 5). Most El Niño events have been found to correspond to the warm phases of these two modes (Jiang et al., 1995), a few correspond to only that of the QQ mode, while warm years in the time series that were not termed as warm events officially have been shown to correspond to a warm phase of the biennial cycle alone, but out-of-phase low frequency component (Rasmusson et al., 1990).

Thus, the third cycle included in the model coincides with this main oscillatory mode, and its estimated period is about 52-56 months. Unlike the biannual cycle, its phase and amplitude are clearly irregular and time-varying (Figure 1(e)).

The robustness of the analysis about the significance of these modes of variability holds both in the case of using raw data or anomalies (Jiang et al., 1995). Our results confirm this, as the estimated frequencies of the three cycle components were the same regardless of filtering the mean annual cycle (not shown). As can be seen in Figure 2, they are associated with the main frequency signals in the spectrum of the N3.4 time series (also see Figure 1). We then assume that these are fundamental time scales of the ENSO phenomenon, and that its irregularity could be to a large degree explained in terms of the relative phase positions and amplitudes of the three cycle components described here.

3.4 Regression Variables, \( x'_t \delta \)

Additionally, we have included a set of explanatory variables in the model for capturing specific dynamic variations in the time series, not explained by the components discussed above. Such variables could also be used to allow for outliers and breaks in the model, or to account for part of the seasonality of the dependent variable in case it is not fully captured by the seasonal component. In fact, it has been proposed that there is a partial 2:1 phase synchronization of ENSO to the annual cycle, and that El Niño is typically associated with a weaker annual cycle (Stein et al., 2011). Since the seasonal component here has been estimated as fixed, i.e.
Improving the Long-Lead Predictability of El Niño

with constant amplitude and periodicity (Section 3.2, Figure 1(b)), we incorporate regression variables that not only aim to explain more variability in the N3.4 time series, but also to model more subtle seasonal fluctuations.

Thus, we extend the decomposition with a multiple regression effect \( x_t' \delta \) for \( t = 1, \ldots, n \) (grey time series in Figure 1(a)), where \( x_t \) is a \( K \times 1 \) vector of predetermined covariates and \( \delta \) is a \( K \times 1 \) vector of regression coefficients. Elements of \( \delta \) could also be allowed to change over time. However, as we want to establish stable relationships between the dependent variable and a set of explanatory variables, \( \delta \) is kept constant for the full sample length or, for at least, a large part of the sample. Therefore, for the present study we have kept the coefficients \( \delta \) fixed in time, and we will explore the time-varying case in a future study.

In very general terms, ENSO represents the alternation of heat buildup and release from the equatorial region (Wyrtki, 1985), (Zebiak, 1989), (Jin, 1997a). The forcing of the zonal wind causes fluctuations in this heat content through Kelvin and Rossby wave dynamics, and SSTs are affected by these waves through downwelling and upwelling processes (Ballester et al., 2015a). SST anomalies, on the other hand, influence the atmosphere and result in anomalies in the zonal wind patterns - the so-called Bjerknes feedback (Bjerknes, 1969). Therefore, we can say that ENSO could be deemed predictable due to the intrinsic ocean variability, as well as the deterministic wind-driven ocean dynamics and the resulting equatorial heat content redistribution (McPhaden, 2004). Based on this assumption, we use zonal equatorial wind stress, and surface and subsurface ocean temperature as regression variables in our model.

### 3.4.1 Equatorial wind stress

Tropical wind stress is among the main drivers of surface and subsurface temperature anomalies in the equatorial Pacific. (Wyrtki, 1975) was the first to hypothesize that El Niño is preceded by abnormally strong southeasterly trade winds in the CPAC, more precisely between 180°E and 220°E, starting approximately two years before the event (Figure 6(a)). Moreover, such anomalies in the zonal wind component represent one of the factors preceding the occurrence of an EN that has an essential role in the triggering of its whole mechanism (Wyrtki, 1975), (Jin, 1997a), (Ballester et al., 2015a). Thus, following our analysis (Figure 6), as well as previous studies (Wyrtki, 1975), (McPhaden, 2004), (Sarachik and Cane, 2010), we have chosen two regions in the CPAC to calculate zonal wind stress indices and obtain time series to be used as regression variables in the proposed model (Figure 6(a), (b), Regions I and II in Table 1). It is interesting to note that one of these regions (Region II) is located entirely in the Southern Hemisphere, as this is the area of maximum easterly zonal wind anomalies at about 11-14 months before the boreal peak of EN (Figure 6(b)). This is also in agreement with the results of (McGregor et al., 2012), where it is shown that the zonal wind anomalies suddenly shift southward at the end of the calendar year.

The strengthening of the westward surface trade winds, on the other hand, increases the intensity of the westward circulation of the South Equatorial Current (SEC) (Yu and McPhaden, 1999), which deepens the thermocline in the WPAC (Sarachik and Cane, 2010). This process piles up warm water at those latitudes and accumulates heat, which often leads to the spreading of the warm pool toward the CPAC (both on the surface and in the subsurface). The warming induces an area of anomalous low surface pressure, over which westerly wind anomalies tend to appear at the Equator near the date line (Figure 6(c), (McPhaden, 2004)). In fact, SST-modulated westerly wind anomalies are critical for the ocean-atmosphere coupling before a warm event (Eisenman et al., 2005), and have occurred prior to every significant El Niño in the last several decades (Figure 6(c), (McPhaden, 2004)). As mentioned earlier (Section 3.3.2) and shown by other studies (Tziperman and Yu, 2007), (Gebbie and Tziperman, 2009), there is a deterministic portion of these anomalies, which is in interplay with the biennial ENSO cycle of SSTs, and could be used as a precursor for it (Gebbie and Tziperman, 2009). Therefore, another critical region for the model is located over the easternmost part of the WPAC and over the CPAC (see Figure 6(c), Region III in Table 1). Over this region we average zonal wind stress to obtain a time series used as a regression variable at short lead times (5-8 months before the typical December peak).

### 3.4.2 Equatorial Pacific surface and subsurface temperatures

The western Pacific plays a fundamental role in the oscillatory behaviour of ENSO and in the generation of warm and cold phases. It is where the warm pool develops - an upper-ocean area of very warm and well-mixed...
waters, surface horizontal current convergence and subsurface divergence, and therefore downwelling motion (Brown and Fedorov, 2010), (Ballester et al., 2015a). As mentioned earlier and shown by (Wyrtki, 1975), (Jin, 1997a), (Jin, 1997b), the abnormally strong trade winds increase the east-west slope of sea level in the CPAC, intensify the westward SEC, and accumulate warm water (heat buildup) in the surface and subsurface WPAC region (Ballester et al., 2015b). Right after a relaxation of the trades, this stored warm water would be allowed to move eastward through the Equatorial Undercurrent (EUC) and the North Equatorial Countercurrent (NECC) in a dynamically consistent way (Jin, 1997a), (Ballester et al., 2015b), and to lead to an EN event in the EPAC (Wyrtki, 1985), (Ramesh and Murtugudde, 2013).

Substantial heat content increase along the equator has preceded all El Niños since 1980 by at least two seasons, and it is a necessary condition for its occurrence (McPhaden, 2004). In addition, it has been shown by (Ramesh and Murtugudde, 2013) that regardless of the regime shifts in the tropics, the buildup of warm water in the WPAC (Figure 7) and the onset of its eastward displacement has remained unchanged, always starting at a particular stage of the cycle - about 18-20 months before the peak (Figure 8), when the heat buildup in the WPAC has grown substantially (Figure 7). This is also the time when the subsurface anomalies begin to appear on the surface in the far western Pacific (Figure 8(a)), and start to weaken the trades (see previous Section 3.4.1).

In this way, a variable that accounts for this heat buildup and the onset of its eastward migration could be used as a precursor for EN. Based on this mechanism we defined two sets of regions over which to average surface and subsurface temperatures in the WPAC/CPAC and obtain time series indices for regression variables (Table 2). The first set of regions corresponds to the early subsurface warming in the WPAC (between latitudes 12°S - 8°N) (Figure 7, Table 2). As can be seen in Figure 7(a)-(c) anomalous warming of the subsurface exists on average as early as 25-30 months prior to the typical peak of EN in boreal winter, at depths between 150-400 metres.

The second set of regions corresponds to the onset of the advection of warm water eastward along and below the thermocline, and the subsidence of the trades on the surface (Figures 6 and 8). During this stage the warm pool extends towards the CPAC in the subsurface (100-300 metres depth) through the strengthened EUC (Ballester et al., 2015b), therefore, this set of regions is located closer to the CPAC (see Figure 8, Table 2). Warm anomalies start to appear in these regions at approximately 25 months or even before (Figure 8(b),(c)), intensify and stay there until about 8-9 months prior to the peak of EN, when they are further advected eastwards to enter the EPAC (Figure 8(a)). At the same time, cold anomalies start to develop in the WPAC region instead, corresponding to the onset and rapid development of an EN event (Figures 7 and 8). This coincides with the occurrence of westerly wind anomalies as discussed earlier, which trigger upwelling Rossby waves and decrease the subsurface temperature in the WPAC due to shoaling of the thermocline there (McPhaden, 2004). Thus, the temperature indices averaged over the first set of regions are also used as regression variables during the mature phase, when there are strong cold anomalies in the WPAC (lead/lag times 0-9, Figure 7(a)-(c), (Yu and Mechoso, 2001)).

Equatorward subsurface (100 metres depth and below) advection from the northern off-equatorial region at this time (5-9 months before the EN peak) due to anomalous upwelling between 5°N - 10°N and downwelling between 3°S - 1°S, also results into an area of strong cold anomalies in the box [140°E - 210°E]x[5°N - 10°N] (Supplementary Figure 1). This inter-hemispheric asymmetry has also been found by (McGregor et al., 2012), where it is revealed in the EOF2 of the integrated upper ocean heat content and shown to lead to the typical EN pattern (EOF1) by about 8-9 months. These vertical velocity anomalies and inter-hemispheric asymmetry result from the latitudinal distribution of zonal (Yu and Mechoso, 2001) and meridional (McGregor et al., 2012) wind stress anomalies. Hence, we calculate a temperature time series from this box and use it as a regression variable in the model at the corresponding significant lead-lag times to further account for the dynamical processes happening before an El Niño, and improve the overall forecast skill of the model.

4 Model Configuration

Based on the recharge-discharge oscillatory theory (Jin, 1997a), and as discussed previously, the buildup of heat at the equatorial Pacific is a prerequisite for El Niño, and the event itself discharges this excess heat poleward.
In addition, the time between two consecutive events could generally be determined by the time needed for the tropical ocean to accumulate enough heat, and normally the amplitude of the event is in a direct proportion to the magnitude of the excessive heat content (McPhaden, 2004). The component configuration of our structural model is then synchronized with these assumptions about the warm buildup in the WPAC and its migration eastwards.

It has been shown that the correct representation of the feedback between SST and westerly winds has a high impact on the dynamical regime of ENSO (Gebbie and Tziperman, 2009). The correct inclusion of the SSTs, subsurface temperature, and zonal wind stress time series at the respective lead/lag times of importance of these variables could lead to a more accurate forecast. In this way, using the dynamics of EN described earlier (see Sections 3.4.1, 3.4.2, Figures 6, 7 and 8), as well as rigorous statistical indicators for goodness of fit (see (Harvey, 1989) for details), such as the prediction error variance ($\text{pev}$), the prediction error mean deviation ($\text{pemd}$), the information criteria Akaike ($\text{AIC}$) and Bayesian Schwartz ($\text{BIC}$), and the coefficient of determination ($R^2$) (see Tables 3, 4, 5 and 6 in Appendix B), we have configured the model in such a way that the different regression variables are only used at their respective monthly lead/lag times of dynamical relevance (see Tables 7 and 8 in Appendix B). These lead/lag times have also been determined based on the estimated coefficients and $t$ values of the explanatory covariates (Tables 3, 4, 5 and 6 in Appendix B, (Harvey, 1989)). The core of the model described in Section 3 is kept constant, but the regression variables given by $x_t' \delta$ are varied depending on the time before the peak when a forecast is started.

As seen in Tables 7 and 8, for every lead/lag time the most statistically relevant predictors (zonal wind stress or temperature at specific depths in Regions I, II or III) are added to the model. These variables are associated with high $t$ values when fitted at these specific lead/lag times, and the overall coefficient of determination of the model $R^2$, as well as the coefficient of determination based on differences around seasonal means $Rs^2$, increase when these particular variables are included (see Tables 3, 4, 5 and 6 in Appendix B). Additionally, the main anomalies of the covariates at the respective lead/lag times are also taken into consideration (the composite anomalies shown in Figures 6, 7 and 8) when determining the model configuration. In summary, depending on the relative time before the potential peak of EN that a forecast is started, different regression variables are added to the model, in order to enhance its forecasting performance.

5 Forecasting results

In order to test and cross-validate the proposed prediction scheme we attempted to forecast all El Niño events that occurred in the period 1996-2015 at long (20-34 months), medium (10-19 months), and short (3-9 months) lead times (Figures 9 and 10). EN events were chosen according to the classification of the Climate Prediction Center for the period 1996-2015: December 1997, 2002, 2006, 2009, 2014 (CPC 2015). As a final test, we proceeded with an experiment, in which we predicted the whole ENSO time series in the period 1982-2014 (Figure 11).

5.1 Forecasts of individual El Niño events

Long-lead (20-34 months in advance; magenta, light blue, dark green and beige curves in Figure 9) forecasts of the 5 events are shown in Figure 9(a)-(e). Some shifts and mismatches in the exact onset and magnitude of the events can be noticed, but their occurrences are properly forecast, and an indication for a developing El Niño is clearly present even at these very long lead times. The 21 months ahead predictions (beige curves in Figure 9(a)-(e)) already represent the events very well in terms of timing and amplitude. It is important to note that at 26 months lead the peak of the 1997/98 EN is predicted to reach an anomaly of $+2 \, ^\circ\text{C}$ (dark green curve in Figure 9(a) and blue curve in Figure 10, Table 9), and that an event of extreme magnitude is foreseen more than two years in advance, long before the appearance of a series of westerly wind bursts in early 1997, assumed by some studies to be the prime reason for the strength of this El Niño (McPhaden and Yu, 1999). This is in support of the theory proposed by (Chen et al., 2004) that the evolution especially of the big events is to a greater extent determined by the initial condition of the ocean, and the atmospheric forcing is rather a secondary modulator or consequence. In addition, the 2014/15 weak EN is forecast with high accuracy 24
months ahead (dark green curve in Figure 9(e), black curve in Figure 10, Table 9), while a large number of the operational models failed and predicted an extreme event similar in size to the 1997/98 one (McPhaden et al., 2014), (Glantz, 2015).

Medium-range (10-19 months in advance; red, blue and green curves in Figure 9) forecasts of the same events are depicted in Figure 9(f)-(j). It is evident that for some of the events the forecast improves as the lead time becomes shorter, as in the case of the 2002/03 and 2009/10 El Niños (Figure 9(g) and (i)). For the 1997/98, the 2006/07 and the 2014/15 events the skill of the model remains similar (Figure 9(f), (h) and (e)). Again, we want to highlight the accurate prediction of the 2014 EN at 16-17 months lead time (blue curve in Figure 9(j), Table 9), and the consistency of the forecast indicating a minor event (black curve in Figure 10) instead of a major one.

As expected, the overall performance of the model is improved at short lead times (3-6 months; dark blue and dark yellow curves in Figure 9(k)-(o)). The peak of the temperature anomaly is already accurately forecast in all 5 cases (Table 9, Figure 10). An interesting feature is that most events (the 1997/98, the 2002/03, and the 2009/10) are better forecast 12-14 months in advance than 9-11 months ahead (velvet curves in Figure 9(k)-(o), blue, red and yellow curves in Figure 10, Table 9). We relate this issue to the fact that 8-9 months before a typical boreal winter peak is the time of the spring barrier, when the signal to noise-ratio is lower and when most models tend to lose their predictability skills (Barnston et al., 2012). This result confirms that some El Niños are more reliably forecast at medium- and long-lead times (Izumo et al., 2014), rather than at shorter ones, something that could seem counterintuitive and calls for further investigation.

Table 9 and Figure 10 summarize the results for the prediction of the target month of January for the five EN cases. Roughly one year and a half to two years in advance the forecasts successfully indicate whether the respective event is expected to be strong, moderate or weak. Moreover, for some lead months the temperature anomaly value of the prediction is almost exactly the same as the value in the observations - for instance 12-14 and 21-23 months ahead of the 2002/03 event, 27 months ahead of the 2006/07 event, and 6-8 and 12-14 months ahead of the 2009/10 event.

Finally, a feature to be noted is that even at short lead times the the evolution of the temperature anomaly during some of the events, for instance the 2002/03 and the 2006/07, appears slightly lagged by 1-2 months in the model than in the observations (Figure 9 panels (b), (g), (l) and (c), (h), (m), respectively). This delay could be linked to the fact that in the current configuration of the model the seasonal cycle has been estimated as a non-varying fixed cycle with a seasonal peak in the months of April - May (Figure 1(b)). Even though in general terms this representation is correct, we know that in reality the seasonality varies especially in EN years. In the present scheme we attempted to model this additional variability through the regression explanatory covariates and the near-annual cycle component, however, this issue should be better addressed in the future. Thus, one avenue for improvement would be to test the model with a time-varying seasonal component, which we expect to result in an even more accurate evolution and timing of the predicted events.

So far it has been shown how the proposed model is highly skillful in forecasting the warm phase of ENSO, of those El Niños that occurred in the last 20 years. In the next subsection we demonstrate its potential to predict the overall phenomenon, including neutral years.

5.2 ENSO time series forecast

The proposed forecasting scheme has been especially targeted at the prediction of both the timing and magnitude of El Niño events, and in the previous subsection its skill in that respect was shown. In the present study the regression covariates used in the model and described earlier (Section 3.4) have been incorporated with the aim to improve the forecast of the warm phase of the oscillation, without considering the cold LN phase. Indeed, the rest of the components of the model are formulated in such a way as to address the prediction of the entire ENSO phenomenon. Therefore, it is expected that the model is also skillful at predicting neutral and LN years, albeit with less precision.

In Figure 11(a) a forecast across the whole interval spanning 1983 - 2014 at six months lead time can be seen. The data between 1952 - 1982 was used for the estimation of the model components, and then prediction
was attempted for the remaining period. This forecast is more complex than the ones presented in the previous subsection when components were estimated on the basis of the available observations prior to every warm event. Instead here, general conditions for the fitting period are applied, and not specific ones related to each ENSO event. In addition, most of this fitting interval covers a time before the ENSO regime shift in 1977 (Hare and Mantua, 2000), while the forecast is made for the years after this shift. As expected the model is capable of predicting all EN events and neutral years in the period, but some of the LN events are missed, for example the cold ones occurring in the years 2000 and 2008.

From the scatterplots of the observed Niño 3.4 index against forecasts at lead times of 3, 6 and 18 months (Figure 11(b), (c) and (d)) it becomes clear that the model has correctly predicted the LN events of extreme amplitude at a lead time of 3 months (Figure 11(b)). With its current formulation, however, the cold anomalies exceeding -2 °C cannot be captured at a lead time of 6 months (Figure 11(c)), while at the longer lead time of 18 months anomalies exceeding -1 °C are missed (Figure 11(d)). Furthermore, the positive skewness coefficients of 0.43 and 0.45 for the forecast timeseries in Figure 11(c) and (d), respectively, indicate an asymmetry of the probability distribution functions with longer tail toward the warm ENSO events. In reality, the rapid termination of El Niño after boreal winter is on average followed by the development of La Niña in the following year (Figure 1(a), (Ballester et al., 2015a)). In these cases the ocean dynamics follows a different pattern than a symmetric pattern (with opposite sign) to the one that precedes El Niño. The eastward propagation of the subsurface cold anomalies and their surface appearance in the EPAC occur faster, for example. More aspects of the asymmetry between the two ENSO phases have been discussed in detail by (Kang and Kug, 2002), (An and Jin, 2004), (An and Choi, 2009), (Yu et al., 2011), and (Ballester et al., 2015a). Therefore, the current predictor regression variables of the model cannot properly signal the development of some of the LN events. Hence, specific regression covariates related to the dynamics and playing a role during and before La Niña will be added to the modeling scheme in the future with the objective to improve the representation of the cold ENSO phase.

6 Discussion

The results presented in the previous section suggest that the structural time series model (Section 3) described in this study outperforms the statistical ENSO models included in the cross-validation exercise, and it is comparable to some of the most accurate dynamical models. Our prediction scheme successfully goes beyond the spring barrier, and also through the one after that, which is an implication for the much longer predictability limit of ENSO than generally accepted. This result, to our knowledge, has only been accomplished by (Chen et al., 2004) so far. Still, we report forecasts started two and a half years before the boreal EN peak, which is even longer than the lead time of the predictions shown therein. Also, in (Chen et al., 2004) a complete account of the reasons underlying this performance is not provided, while here a dynamical understanding of the EN-associated processes accompanies the definition of our model structure and regional covariates.

Additionally, we want to highlight the simplicity of the proposed prediction scheme, as it uses readily available data directly as input - monthly values of SSTs, subsurface ocean temperature in the vicinity of the equatorial Pacific, and zonal wind stress; and it does not require extensive computations. Therefore, our model appears to have a clear advantage over the more complicated statistical schemes, as well as over the computationally more expensive dynamical models. We would also like to discuss one important advantage that is especially due to the state space approach applied here.

Two of the forecast events in Figure 9 were unusual in the respect that the frequencies of the three cycle components estimated with the Kalman Filter were different. In the case of the 2006/07 event (Supplementary Figure 2), the quasi-biennial and quasi-quadrennial cycles were preserved, but the third cycle was estimated to have a frequency associated with decadal scales of variability (Supplementary Figure 2(c)). Actually, when the model was fixed as described in Section 3 and shown in Figure 1, the event was shifted and forecast for the next year 2007/08 (not shown). This confirms the important role of the decadal tropical Pacific variability in the modulation of the ENSO amplitude, which was reported by previous studies (Kleeman et al., 1999), (Yeh and Kirtman, 2004), (Yeh and Kirtman, 2005). As the model in its current configuration has only three cycle...
components, we cannot deduce whether the near-annual cycle component still has a vital contribution to the evolution of this event.

In Supplementary Figure 2(a), however, it can be seen that the estimated two-year oscillation is not regular as in Figure 1(d). We have plotted together the cycles from Figure 1(c) and (d) (the near-annual and the biannual oscillations), and we superimposed the irregular quasi-biannual component from Supplementary Figure 2(a) (see Supplementary Figure 3). The irregular two-year cycle appears to result from the convolution of the other two cycles. This could explain the occasional shift in the amplitude and phase of the otherwise regular biannual cycle, which was mentioned earlier (see Section 3.3.2). In fact, it has been difficult to clearly isolate the peaks in the spectrum of both zonal wind and SSTs associated with the oscillations at periods of 16-18 and 24-28 months in the analysis done by (Rasmussen et al., 1990) and (Jiang et al., 1995), favouring this kind of conclusion. In any case, a model configuration in which both cycles are present in an explicit way would most probably result into an even more precise forecast of the 2006/07 El Niño, but this is not in the scope of the present article.

The other event, for which the frequencies of the cycle components were estimated differently was the prominent 1997/08 event. In the case of this El Niño, which is also the most extreme on record, the quasi-biennial cycle was calculated as for the 2006/07 event (see Supplementary Figures 2(a) and 4(a)), while the other two cycle frequencies corresponded to periods of 43 – 45 and 62 – 65 months (Supplementary Figure 4(b),(c)). In this way, the event is characterized by a "double" or "split" quasi-quadrennial cycle. In order to shed more light onto this distinction, we examined these two cycles in more detail. In Supplementary Figure 4 we can clearly see that the cycle with 65 months period has its peak in 1992/93 (Supplementary Figure 4(c)), while the one with 45 months peaks in 1991 (Supplementary Figure 4(b)). On the one hand, in 1991/92 there was an official El Niño event. On the other hand, the year 1993 was indeed exceptionally warm (Figure 1(a)). These two cycles then peak coincidentally in 1998, which preconditions an event of higher magnitude. It should also be taken into consideration that the prediction of the 1997/08 event is made by fitting the data between 1982-1995, which is a short period of time when only one event of similar, but still lower magnitude, occurred. Thus, the model has performed very well by forecasting this exceptional event.

These preliminary results lead to the conclusion that at least some of the bigger and more extreme EN events occur as a result of the superposition of these two cycles in combination with the quasi-biennial cycle, the near-annual cycle and/or a cycle on decadal timescales. A recent study (Chen et al., 2015) classified EN events into three different categories - extremely strong ones, weak warm events reminiscent of the "warm pool El Niño", and the canonical ones with moderate warming in the central/eastern equatorial Pacific. Thus, taking into account this classification, we presume that on average for the canonical events, the cycles that have an important role in EN development are the ones discussed in Section 3 and shown in Figure 1. On the other hand, for the more specific smaller events, and for the very extreme ones, the dynamical mechanisms seem to be complicated by the effect of additional cycles superimposed over these basic ones. Considering the results presented here, we believe that the parameter re-estimation of the various model components, and especially of the three cycle frequencies, based on the observational data prior to each event, is crucial for the accurate forecasting of El Niño. Hence, the long-lead time prediction capabilities of our scheme come from the correct representation of the ocean-atmosphere system at each stage of its evolution. The fact that the state space form and the Kalman Filter allow for the unknown parameter re-estimation whenever an updated information about the coupled system becomes available is a fundamental advantage of our model.

7 Concluding Remarks and Future Directions

A new ENSO forecasting scheme has been developed, based on structural unobserved components time series modeling with a state space approach. The unknown parameters of the unobserved components are re-estimated before every event, thus rendering the unique flexibility of the model. Regression variables especially designed to account for ocean surface and subsurface anomalous processes at the equatorial Pacific prior to individual El Niño events are incorporated at their specific time of relevancy. Since all unknown parameters are calculated together in a dynamic way, the addition of these covariates also affects the estimation of all the other model components in an efficient way, helping for the correct update of the whole system.

The model has been tested and cross-validated through the retrospective forecasts of the El Niños events in
the recent period 1996-2015, and it has successfully predicted all the events that occurred at both long lead
times (two years and a half in advance), as well as at shorter lead times before and after the spring predictability
barrier. Essentially, the model has accurately forecast a moderate EN for the end of 2014/beginning of 2015,
unlike the majority of the operational models that warned of an event of substantial magnitude similar to that
of the 1997/98 one.

A direct conclusion of high relevance is that the predictability limit of ENSO could be extended to at least
two and a half years before the El Niño peak. Our study also clearly demonstrates that there is still room for
improvement in the prediction of ENSO, not only in the forecasting of the cold phase of the oscillation, but also
in that of EN itself. One immediate avenue for future work should be the inclusion of explanatory covariates
especially aimed at improving the prediction of LN events, by taking into account the La Niña specific dynamics.
Another area for improvement would be the transformation of the modeling scheme, so that all the important
cycles discussed here - the additional quasi-quadrennial cycle, the decadal cycle, as well as independent near-
annual and biannual cycles, could be separately incorporated in it. Finally, special attention should be paid to
the incorporation of seasonality as discussed in Section 5, by changing the seasonal component of the model
from being fixed to a time-varying one.
| Table  | Description                                                                 | Page |
|--------|-----------------------------------------------------------------------------|------|
| 1      | Regions over which zonal wind stress is averaged to obtain time series used as predictor regression variables in the model. | 18   |
| 2      | Regions over which surface and subsurface temperatures are averaged to obtain time series used as predictor regression variables in the model. | 18   |
| 3      | Diagnostics of wind stress predictor regression variables.                   | 20   |
| 4      | Diagnostics of surface and subsurface temperature predictor regression variables. | 21   |
| 5      | Diagnostics of subsurface temperature predictor regression variables.         | 22   |
| 6      | Diagnostics of subsurface temperature predictor regression variables.         | 23   |
| 7      | Predictor regression variables added to the model at lead times between 17 and 36 months, based on the criteria and results shown in Tables 3-6. | 24   |
| 8      | Predictor regression variables added to the model at lead times between 0 and 15 months, based on the criteria and results shown in Tables 3-6. | 25   |
| 9      | Predictions of the January target month for all EN events shown in Figure 9. Given in brackets is the probability that the respective value would occur based on a kernel normal probability density estimation of the N3.4 index with bandwith of the kernel smoothing window h=0.2704. Events are categorized as: strong/extreme (0-9%), moderate (10-24%), weak (25-35%). Coloured cells indicate if the category of the event is predicted (blue), or if an adjacent category is predicted instead (yellow). The observation values are in bold. | 26   |
List of Figures

1 Components graphics of the model. Shown are temperature (°C) time series of the (a) level and regression components together with the N3.4 index, (b) the seasonal component, and the three cycle components of periods: (c) 1.5, (d) 2.5, (e) 4.5 years. 27

2 (a) Multi Taper Method (MTM) power spectra for the observed N3.4 time series. The solid line indicates the power density and dashed lines the respective confidence level (CL) based on a red noise null hypothesis. The red indicators correspond to the near-annual, biannual, and quasi-quadrennial ENSO modes of variability. Reconstructed components from the multitaper decomposition in (a), corresponding to the (b) seasonal, (c) near-annual, (d) biannual, (e) quasi-quadrennial modes. 28

3 Pearson correlations between the temporal scores of the first CEOF modes of filtered SSTs and surface wind stress anomalies, and filtered spatio-temporal SST anomalies and wind stress anomalies in the equatorial Pacific region. A Butterworth filter has been applied to the SST and wind stress data sets, so that only frequencies corresponding to periods between 14 and 18 months (associated with the near-annual mode of variability) have been kept. Panels correspond to the respective phases of the CEOF shown on the figure. Shaded areas indicate significant anomalies. 29

4 Same as Figure 3, but the Butterworth filter has been applied so that only frequencies corresponding to periods between 24 and 28 months (associated with the biannual mode of variability) have been kept. 30

5 Same as Figure 3, but the Butterworth filter has been applied so that only frequencies corresponding to periods between 46 and 63 months (associated with the low-frequency mode of variability) have been kept. 31

6 Composites of surface zonal wind stress (Nm⁻², arrows) anomalies with respect to all EN events in the period 1978-2012. Shown are anomalies (a) 24, (b) 13, (c) 7 months before the winter peak of EN. The red boxes indicate the three zonal wind stress regions from Table 1 - (a) Region I, (b) Region II, (c) Region III. 32

7 Composites of subsurface temperature (°C, shading) anomalies at (a) 150, (b) 250, (c) 400 metres depth with respect to all EN events in the period 1978-2012 in Region I (see Table 2). Data is filtered using a low-pass Butterworth filter (cut-off frequency 18, order 10). 33

8 Same as Figure 7, but at (a) surface, (b) 100, (c) 200 metres depth in Region II (see Table 2). 34

9 Time series of area-averaged sea surface temperature (°C) anomalies in the Niño 3.4 region. Shown are forecasts of the (a), (f), (k) 1997/98, (b), (g), (l) 2002/03, (c), (h), (m) 2006/07, (d), (i), (n) 2009/10, and (e), (j), (o) 2014/15 EN events, starting 29-34 (magenta), 27-28 (light blue), 24-26 (dark green), 21-22 (beige), 17-19 (red), 13-16 (blue), 11-12 (green), 8-9 (velvet), 6 (dark blue), and 3-5 (dark yellow) months before the peak of El Niño, respectively. Observations are in black. 35

10 Predictions of the January target month for all EN events shown in Figure 9: 1997/98 (blue), 2002/03 (red), 2006/07 (green), 2009/10 (yellow), 2014/15 (black). The lead months for the respective events are as in Table 9, and lead month 0 corresponds to the observations. 36

11 (a) Retrospective forecast of the EN3.4 time series in the period 1983-2014. The EN3.4 observation is in red and the model prediction at 6 months lead time is in blue. Scatterplots of the EN3.4 time series observation against forecast at (b) 3, (c) 6, (d) 18 months lead time. The respective regression coefficients are 0.70, 0.45 and 0.30. 37
Table 1  Regions over which zonal wind stress is averaged to obtain time series used as predictor regression variables in the model.

| Regression variable | Region I                  | Region II                  | Region III                  |
|---------------------|---------------------------|---------------------------|-----------------------------|
| Zonal Wind stress   | $[180° - 220°]x[4s - 4n]$| $[180° - 210°]x[10s - 0]$ | $[160° - 200°]x[0 - 10n]$  |

Table 2  Regions over which surface and subsurface temperatures are averaged to obtain time series used as predictor regression variables in the model.

| Regression variable | Region I                  | Region II                  |
|---------------------|---------------------------|---------------------------|
| SST WPAC            | $[140° - 160°]x[5s - 5n]$|                           |
| SST WPAC2           | $[140° - 180°]x[10s - 5n]$|                           |
| SST WPAC3           | $[120° - 170°]x[10s - 5n]$|                           |
| SST WPAC4           | $[140° - 160°]x[10s - 0]$  |                           |
| Temperature at 50m. depth | $[120° - 170°]x[10s - 5n]$|                           |
| Temperature at 100m. depth | "cold" | $[140° - 210°]x[5n - 10n]$               |
| Temperature at 150m. depth | $[120° - 140°]x[10s - 5n]$ | $[150° - 180°]x[7s - 7n]$               |
| Temperature at 200m. depth | $[120° - 140°]x[10s - 7n]$ | $[150° - 180°]x[7s - 7n]$               |
| Temperature at 250m. depth | $[120° - 140°]x[7s - 7n]$ | $[140° - 170°]x[7s - 7n]$               |
| Temperature at 300m. depth | $[120° - 140°]x[7s - 7n]$ | $[160° - 200°]x[10s - 3n]$               |
| Temperature at 400m. depth | $[120° - 140°]x[5s - 5n]$ | $[150° - 170°]x[10s - 0]$               |
| Temperature at 500m. depth | $[120° - 140°]x[5s - 5n]$ | $[150° - 170°]x[10s - 0]$               |
Appendix A

The linear Gaussian state space model form that we have used is as in (de Jong, 1991), that is:

\[ y_t = Z_t \alpha_t + G_t \epsilon_t, \quad \alpha_{t+1} = T_t \alpha_t + H_t \epsilon_t, \quad \epsilon_t \sim \text{NID}(0, I), \]  

(A.1)

for \( t = 1, \ldots, n \), and where \( \epsilon_t \) is a vector of serially independent disturbance series. The \( m \times 1 \) state vector \( \alpha_t \) contains the unobserved components and their associated variables. The measurement equation is the first equation in (A.1) and it relates the observation \( y_t \) to the state vector \( \alpha_t \) through the signal \( Z_t \alpha_t \). The transition equation is the second equation in (A.1) and it is used to formulate the dynamic processes of the unobserved components in a companion form. The deterministic matrices \( T_t, Z_t, H_t \) and \( G_t \), are referred to as system matrices and they are sparse and known:

\[ T_t = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \]

\[ H_t = \begin{bmatrix} 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k & 0_k \end{bmatrix}, \]

\[ Z_t = \begin{bmatrix} 0 \end{bmatrix}, \]

\[ G_t = \begin{bmatrix} 0 \end{bmatrix}, \]

where \( 0_k \) is a \( k \times 1 \) vector of zeros, \( I_k \) is a \( k \times k \) identity matrix, \( p(i) = \cos \lambda_i \) and \( q(i) = \sin \lambda_i \) for \( \lambda_i = \frac{2\pi i}{5}, i=1,2,\ldots,\lfloor \frac{5}{2} \rfloor \); \( c(j) = \varphi_{\psi,j} \cos \lambda_{c,j} \) and \( s(j) = \varphi_{\psi,j} \sin \lambda_{c,j} \) for \( j = 1,2,3 \).
Every explanatory regression variable that has been used in the analysis (surface temperature, subsurface temperature at different depth levels and regions, and zonal wind stress at different regions) has been tested separately with the model described in Section 3 during the fitting procedure. In this way each variable was fitted at every lag time between 0-35 months, and based on the in-sample estimations for each of these fittings (the sample spanned the data from January 1982 to December 2012), the $t$ value (an indicator of statistical significance, showing the mean effect on the predicted variable for one unit of change of the predictor variable), and the $R^2$ value (modified coefficient of determination based on seasonal means - the ratio between the variance explained by the model and the variance of the seasonally differenced time series), were used to determine significance of the regression variable at the respective lag time. The significant values obtained in this way are summarized in the following Tables 3, 4, 5 and 6.

### Table 3  Diagnostics of wind stress predictor regression variables.

| Regression variable | Time Series Region | Lag (months) | $t$ value | $R^2$  |
|---------------------|--------------------|--------------|-----------|--------|
| Zonal Wind stress   | Region I           | 7            | −1.7992   | 0.3220 |
|                     |                    | 8            | −2.4897   | 0.3110 |
|                     |                    | 9            | 2.4706    | 0.3125 |
|                     |                    | 16           | 2.3206    | 0.3048 |
|                     |                    | 20           | −2.6512   | 0.3167 |
|                     |                    | 21           | 2.5316    | 0.3141 |
|                     |                    | 25           | 1.6419    | 0.3119 |
|                     |                    | 27           | 2.8902    | 0.3229 |
|                     |                    | 28           | −2.2488   | 0.3038 |
|                     |                    | 29           | 1.7566    | 0.3007 |
|                     |                    | 30           | −1.5120   | 0.3042 |
| Region II           |                    | 13           | 2.0347    | 0.3036 |
|                     |                    | 25           | 2.0687    | 0.3151 |
|                     |                    | 27           | 3.1003    | 0.3254 |
|                     |                    | 28           | −2.6615   | 0.3080 |
|                     |                    | 30           | −1.4777   | 0.3122 |
| Region III          |                    | 6            | −1.6367   | 0.3045 |
|                     |                    | 8            | −2.4165   | 0.3105 |
|                     |                    | 15           | −3.1531   | 0.3455 |
|                     |                    | 16           | 2.5574    | 0.3072 |
|                     |                    | 19           | −2.2799   | 0.3304 |
|                     |                    | 20           | −2.9359   | 0.3402 |
|                     |                    | 28           | −1.5839   | 0.3025 |
## Table 4
Diagnostics of surface and subsurface temperature predictor regression variables.

| Regression variable | Time Series Region | Lag (months) | t value | $R^2$ |
|---------------------|--------------------|--------------|---------|------|
| SST WPAC            |                    | 2            | -2.2425 | 0.2981 |
|                     |                    | 14           | -2.3408 | 0.2898 |
|                     |                    | 15           | 2.6407  | 0.2888 |
|                     |                    | 17           | 1.8385  | 0.2716 |
| WPAC2               |                    | 2            | -2.5059 | 0.3004 |
|                     |                    | 14           | -2.1784 | 0.2882 |
|                     |                    | 15           | 2.5313  | 0.2878 |
|                     |                    | 17           | 2.0083  | 0.2759 |
| WPAC3               |                    | 2            | -3.5643 | 0.3118 |
|                     |                    | 15           | 2.6477  | 0.2889 |
|                     |                    | 17           | 1.5224  | 0.2725 |
| WPAC4               |                    | 2            | -4.2786 | 0.3218 |
|                     |                    | 5            | 2.6127  | 0.2922 |
|                     |                    | 15           | 2.7964  | 0.2940 |
|                     |                    | 17           | 1.7389  | 0.2739 |
| Temperature at 50m. depth | Region I | 2            | -3.3747 | 0.3466 |
|                     |                    | 3            | -2.0442 | 0.3352 |
|                     |                    | 14           | -1.5026 | 0.3232 |
|                     |                    | 17           | 1.8814  | 0.3079 |
|                     |                    | 19           | -1.4214 | 0.3092 |
| Temperature at 100m. depth | Region I | 2            | -4.6209 | 0.3512 |
|                     |                    | 3            | -3.7341 | 0.3435 |
|                     |                    | 6            | -1.8102 | 0.3341 |
|                     |                    | 7            | -1.9743 | 0.3349 |
|                     |                    | 8            | -2.3376 | 0.3331 |
|                     |                    | 10           | -1.9429 | 0.3239 |
|                     |                    | 19           | -2.6290 | 0.3181 |
|                     |                    | 20           | -1.7858 | 0.3108 |
|                     |                    | 35           | -2.3930 | 0.3275 |
|                     | Region II          | 2            | -2.1028 | 0.3069 |
|                     |                    | 4            | -1.9415 | 0.2981 |
|                     |                    | 7            | 1.9118  | 0.3061 |
|                     |                    | 9            | 2.2558  | 0.2680 |
|                     |                    | 13           | 2.2515  | 0.2886 |
|                     |                    | 22           | -3.1696 | 0.2953 |
|                     |                    | 25           | 3.2334  | 0.2879 |
|                     |                    | 26           | 2.2032  | 0.2878 |
|                     |                    | 35           | -2.2919 | 0.2956 |
|                     |                    | 36           | 1.6189  | 0.2953 |
| Temperature at 100m. depth "cold" | Region I | 6            | 1.7370  | 0.2907 |
|                     |                    | 8            | -2.3226 | 0.2930 |
|                     |                    | 9            | -1.7627 | 0.2827 |
|                     |                    | 26           | 1.6915  | 0.2798 |
Table 5  Diagnostics of subsurface temperature predictor regression variables.

| Temperature variable | Time Series Region | Lag (months) | t value  | $R^2$ |
|----------------------|--------------------|--------------|----------|-------|
| Temperature at 150m. depth | Region I | 2 | −3.1557 | 0.3135 |
|                       |         | 3 | −2.8267 | 0.3102 |
|                       |         | 8 | −3.1077 | 0.3121 |
|                       |         | 9 | −2.1767 | 0.2968 |
|                       |         | 10 | −2.5713 | 0.2994 |
|                       |         | 19 | −1.9338 | 0.2861 |
|                       |         | 26 | −1.8794 | 0.2957 |
|                       | Region II | 4 | 1.8673 | 0.3252 |
|                       |         | 5 | 2.5595 | 0.3294 |
|                       |         | 10 | 1.4741 | 0.3196 |
|                       |         | 14 | 1.7662 | 0.3225 |
|                       |         | 22 | −1.7018 | 0.2815 |
|                       |         | 25 | 1.9408 | 0.3142 |
|                       |         | 35 | −1.6431 | 0.3211 |
|                       |         | 36 | −1.8571 | 0.3267 |
| Temperature at 200m. depth | Region I | 8 | −2.0524 | 0.3315 |
|                       |         | 9 | −2.5669 | 0.3296 |
|                       |         | 10 | −1.8334 | 0.3234 |
|                       |         | 19 | −1.9453 | 0.3127 |
|                       |         | 20 | −1.6745 | 0.3104 |
|                       | Region II | 3 | 2.6488 | 0.3392 |
|                       |         | 4 | 3.3578 | 0.3363 |
|                       |         | 5 | 3.0530 | 0.3344 |
|                       |         | 13 | 1.4313 | 0.3197 |
|                       |         | 14 | 2.1888 | 0.3256 |
|                       |         | 35 | −1.5980 | 0.3209 |
|                       |         | 36 | −2.0702 | 0.3284 |
| Temperature at 250m. depth | Region I | 5 | 2.0590 | 0.3272 |
|                       |         | 6 | 1.6471 | 0.3343 |
|                       |         | 8 | −3.1977 | 0.3428 |
|                       |         | 9 | −2.1869 | 0.3269 |
|                       |         | 12 | 1.9994 | 0.3217 |
|                       |         | 24 | 1.6067 | 0.3079 |
|                       |         | 28 | −2.0947 | 0.3492 |
|                       |         | 29 | 1.6823 | 0.2818 |
|                       |         | 35 | −1.6184 | 0.3212 |
|                       | Region II | 5 | 3.2892 | 0.3380 |
|                       |         | 11 | −1.8307 | 0.3227 |
|                       |         | 13 | 1.5958 | 0.3210 |
|                       |         | 21 | −2.1908 | 0.2859 |
|                       |         | 27 | −2.2017 | 0.3239 |
|                       |         | 34 | −2.2762 | 0.3234 |
| Regression variable | Time Series Region | Lag (months) | t value | $R_s^2$ |
|---------------------|--------------------|-------------|---------|--------|
| Temperature at 300m. depth | Region I | 9 | -1.7998 | 0.3240 |
|                      |         | 24 | 1.6264 | 0.3078 |
|                      |         | 27 | -1.6121 | 0.3200 |
|                      |         | 28 | -2.2777 | 0.3170 |
|                      |         | 29 | 1.7923 | 0.3163 |
|                      |         | 34 | -1.5290 | 0.3179 |
|                      |         | 35 | -2.8543 | 0.3321 |
|                      | Region II | 2 | 1.5273 | 0.3185 |
|                      |         | 4 | 2.6804 | 0.3182 |
|                      |         | 6 | 2.1884 | 0.3025 |
|                      |         | 17 | 2.2280 | 0.3023 |
|                      |         | 32 | -1.6696 | 0.3211 |
| Temperature at 400m. depth | Region I | 17 | -1.5188 | 0.3077 |
|                      |         | 24 | 1.4976 | 0.2850 |
|                      |         | 28 | -1.4708 | 0.2843 |
|                      |         | 29 | 1.7926 | 0.3038 |
|                      | Region II | 10 | -1.6361 | 0.3215 |
|                      |         | 12 | -1.7666 | 0.3240 |
|                      |         | 15 | 1.6449 | 0.3088 |
|                      |         | 29 | -2.0702 | 0.3033 |
| Temperature at 500m. depth | Region I | 2 | -1.8180 | 0.3192 |
|                      |         | 3 | -1.753 | 0.3202 |
|                      |         | 6 | 2.2588 | 0.3080 |
|                      |         | 10 | -2.0625 | 0.3109 |
|                      |         | 11 | -2.2651 | 0.3124 |
|                      |         | 36 | -1.4756 | 0.3156 |
|                      | Region II | 4 | 2.5037 | 0.3164 |
|                      |         | 5 | 4.1929 | 0.2648 |
|                      |         | 6 | 2.9560 | 0.3432 |
|                      |         | 10 | -2.6661 | 0.3298 |
|                      |         | 11 | -4.0363 | 0.3416 |
|                      |         | 12 | -1.6022 | 0.3222 |
|                      |         | 15 | 2.0388 | 0.3244 |
|                      |         | 17 | 1.9125 | 0.3122 |
Table 7 Predictor regression variables added to the model at lead times between 17 and 36 months, based on the criteria and results shown in Tables 3-6.

| Lag/Lead | Regression variables with high $t$ (over 1.4) and $R^2$ (over 0.26) value for this lag |
|----------|--------------------------------------------------------------------------------------|
| 36       | Temperature at 500m. depth Region I, Temperature at 100m., 150m., 200m. depth Region II |
| 35       | Temperature at 100m., 250m., 300m. depth Region I, Temperature at 100m., 150m., 200m. depth Region II |
| 34       | Temperature at 250m. depth Region II                                                |
| 32       | Temperature at 300m. depth Region II                                                |
| 30       | Zonal wind stress Region I and II                                                   |
| 29       | Temperature at 250m., 300m., 400m. depth Region I                                   |
| 28       | Temperature at 250m., 300m., 400m. depth Region I, Zonal wind stress Region I and III |
| 27       | Temperature at 300m. depth Region I, Temperature at 250m. depth Region II, Zonal wind stress Region I and II |
| 26       | Temperature at 100m. depth “cold”, Temperature at 150m. depth Region I, Temperature at 100m. depth Region II |
| 25       | Temperature at 100m., 150m. depth Region II, Zonal wind stress Region I and II       |
| 24       | Temperature at 250m., 300m., 400m. depth Region I                                   |
| 22       | Temperature at 100m., 150m. depth Region II                                        |
| 21       | Temperature at 250m. depth Region II, Zonal wind stress Region I                    |
| 20       | Temperature at 100m., 200m. depth Region I, Zonal wind stress Region I and III       |
| 19       | Temperature at 50m., 100m., 150m., 200m. depth Region I, Zonal wind stress Region III |
| 17       | SST WPAC, WPAC2, WPAC3, WPAC4, Temperature at 50m., 400m. depth Region I, Temperature at 300m., 500m. depth Region II |
Table 8  Predictor regression variables added to the model at lead times between 0 and 15 months, based on the criteria and results shown in Tables 3-6.

| Lag/Lead | Regression variables with high $t$ (over 1.4) and $R^2$ (over 0.26) value for this lag |
|----------|--------------------------------------------------------------------------------------|
| 15       | SST WPAC, WPAC2, WPAC3, WPAC4, Temperature at 250m., 400m., 500m. depth Region II, Zonal wind stress Region III |
| 14       | SST WPAC, WPAC2, Temperature at 50m. depth, Temperature at 150m., 200m. depth Region II |
| 13       | Temperature at 100m., 200m., 250m. depth Region II |
| 12       | Temperature at 250m. depth Region I, Temperature at 400m., 500m. depth Region II |
| 11       | Temperature at 500m. depth Region I, Temperature at 250m., 500m. depth Region II, Zonal wind stress Region II |
| 10       | Temperature at 100m., 150m., 200m., 500m. depth Region I, Temperature at 150m., 400m., 500m. depth Region II |
| 9        | Temperature at 150m., 200m., 250m., 300m. depth Region I, Temperature 100m. depth Region II, Temperature at 100m. depth Region “cold” |
| 8        | Temperature at 100m. depth Region “cold”, Temperature at 100m., 150m., 200m., 250m., depth Region I, Zonal wind stress Region I and III |
| 7        | Temperature at 100m. depth Region I and II, Zonal wind stress Region I and III |
| 6        | Temperature at 100m. depth Region “cold”, Temperature at 100m., 250m., 500m. depth Region I, Temperature at 300m., 500m. depth Region II |
| 5        | SST WPAC4, Temperature at 250m., depth Region I, Temperature at 150m., 200m., 250m., 500m. depth Region II |
| 4        | Temperature at 100m., 150m., 200m., 300m., 500m. depth Region II |
| 3        | Temperature at 50m. depth, Temperature at 100m., 150m., 500m. depth Region I, Temperature at 200m. depth Region II |
| 2        | SST WPAC, WPAC2, WPAC3, WPAC4, Temperature at 50m. depth, Temperature at 100m. depth Region I and II, Temperature at 150m., 500m. depth Region I, Temperature at 300m. depth Region II |
| 1        | SST WPAC, WPAC3, WPAC4, Temperature at 50m. depth, Temperature at 100m., 150m., 200m., 250m., 300m., 400m. depth Region I and II, 500m. depth Region II |
| 0        | SST WPAC2, Temperature at 50m. depth, Temperature at 100m., 150m., 200m., 250m. depth Region I and II, Temperature at 300m., 400m., 500m. depth Region II |
Table 9 Predictions of the January target month for all EN events shown in Figure 9. Given in brackets is the probability that the respective value would occur based on a kernel normal probability density estimation of the N3.4 index with bandwidth of the kernel smoothing window $h=0.2704$. Events are categorized as: strong/extreme (0-9%), moderate (10-24%), weak (25-35%). Coloured cells indicate if the category of the event is predicted (blue), or if an adjacent category is predicted instead (yellow). The observation values are in bold.

| Lead month | 1997-8 | 2002-3 | 2006-7 | 2009-10 | 2014-15 |
|------------|--------|--------|--------|---------|---------|
| 27-29      | 1.420(10%) | 0.644(28%) | 0.645(28%) | 0.950(19%) | 0.223(41%) |
| 24-26      | 2.000(4%) | 0.784(24%) | 0.602(29%) | 1.162(15%) | 0.549(32%) |
| 21-23      | 1.401(11%) | 1.187(14%) | 0.628(29%) | 1.180(14%) | 0.506(34%) |
| 18-20      | 1.792(6%) | 0.971(19%) | 0.538(32%) | 1.165(15%) | 0.354(38%) |
| 15-17      | 1.544(9%) | 0.758(24%) | 0.639(28%) | 1.262(13%) | 0.565(31%) |
| 12-14      | 1.377(11%) | 1.110(15%) | 0.682(27%) | 1.457(10%) | 0.351(38%) |
| 9-11       | 1.305(12%) | 0.969(19%) | 0.642(28%) | 1.257(13%) | 0.519(33%) |
| 6-8        | 2.006(4%) | 1.138(15%) | 0.718(26%) | 1.446(10%) | 0.500(34%) |
| 3-5        | 2.503(3%) | 1.394(11%) | 0.727(25%) | 1.564(9%) | 0.413(30%) |
| Obs.       | 2.494(3%) | 1.146(15%) | 0.645(28%) | 1.463(10%) | 0.622(30%) |
Fig. 1 Components graphics of the model. Shown are temperature (°C) time series of the (a) level and regression components together with the N3.4 index, (b) the seasonal component, and the three cycle components of periods: (c) 1.5, (d) 2.5, (e) 4.5 years.
Fig. 2 (a) Multi Taper Method (MTM) power spectra for the observed N3.4 time series. The solid line indicates the power density and dashed lines the respective confidence level (CL) based on a red noise null hypothesis. The red indicators correspond to the near-annual, biannual, and quasi-quarrennial ENSO modes of variability. Reconstructed components from the multitaper decomposition in (a), corresponding to the (b) seasonal, (c) near-annual, (d) biannual, (e) quasi-quarrennial modes.
Fig. 3 Pearson correlations between the temporal scores of the first CEOF modes of filtered SSTs and surface wind stress anomalies, and filtered spatio-temporal SST anomalies and wind stress anomalies in the equatorial Pacific region. A Butterworth filter has been applied to the SST and wind stress data sets, so that only frequencies corresponding to periods between 14 and 18 months (associated with the near-annual mode of variability) have been kept. Panels correspond to the respective phases of the CEOF shown on the figure. Shaded areas indicate significant anomalies.
Fig. 4 Same as Figure 3, but the Butterworth filter has been applied so that only frequencies corresponding to periods between 24 and 28 months (associated with the biannual mode of variability) have been kept.
Fig. 5 Same as Figure 3, but the Butterworth filter has been applied so that only frequencies corresponding to periods between 46 and 63 months (associated with the low-frequency mode of variability) have been kept.
Fig. 6 Composites of surface zonal wind stress (Nm$^{-2}$, arrows) anomalies with respect to all EN events in the period 1978-2012. Shown are anomalies (a) 24, (b) 13, (c) 7 months before the winter peak of EN. The red boxes indicate the three zonal wind stress regions from Table 1 - (a) Region I, (b) Region II, (c) Region III.
Fig. 7 Composites of subsurface temperature (°C, shading) anomalies at (a) 150, (b) 250, (c) 400 metres depth with respect to all EN events in the period 1978-2012 in Region I (see Table 2). Data is filtered using a low-pass Butterworth filter (cut-off frequency 18, order 10).
Region II

![Composites for 0m depth](image)

![Composites for 100m depth](image)

![Composites for 200m depth](image)

Fig. 8 Same as Figure 7, but at (a) surface, (b) 100, (c) 200 metres depth in Region II (see Table 2).
Fig. 9  Time series of area-averaged sea surface temperature (°C) anomalies in the Niño 3.4 region. Shown are forecasts of the (a), (f), (k) 1997/98, (b), (g), (l) 2002/03, (c), (h), (m) 2006/07, (d), (i), (n) 2009/10, and (e), (j), (o) 2014/15 EN events, starting 29-34 (magenta), 27-28 (light blue), 24-26 (dark green), 21-22 (beige), 17-19 (red), 13-16 (blue), 11-12 (green), 8-9 (velvet), 6 (dark blue), and 3-5 (dark yellow) months before the peak of El Niño, respectively. Observations are in black.
Fig. 10 Predictions of the January target month for all EN events shown in Figure 9: 1997/98 (blue), 2002/03 (red), 2006/07 (green), 2009/10 (yellow), 2014/15 (black). The lead months for the respective events are as in Table 9, and lead month 0 corresponds to the observations.
Fig. 11 (a) Retrospective forecast of the EN3.4 time series in the period 1983-2014. The EN3.4 observation is in red and the model prediction at 6 months lead time is in blue. Scatterplots of the EN3.4 time series observation against forecast at (b) 3, (c) 6, (d) 18 months lead time. The respective regression coefficients are 0.70, 0.45 and 0.30.
Acknowledgements  JB gratefully acknowledges funding from the European Commission through a Marie Curie International
Outgoing Fellowship (project MEMENTO from the FP7-PEOPLE-2011-IOF call), and from the European Commission and the
Catalan Government through a Marie Curie - Beatriu de Pinós Fellowship (project 00068 from the BP-DGR-2014-B call).
XR gratefully acknowledges funding from the Ministry of Science and Innovation, Spain (project PANDORA CGL 2007-63053).
References

An, S. I. and Choi, J. (2009). Seasonal locking of the ENSO asymmetry and its influence on the seasonal cycle of the tropical eastern Pacific sea surface temperature. *Atmospheric Research*, 94:3–9.

An, S. I. and Jin, F. F. (2004). Nonlinearity and asymmetry of ENSO. *Journal of Climate*, 17:2399–2412.

Ballester, J., Bordoni, S., Petrova, D., and Rodo, X. (2015a). A) On the dynamical mechanism explaining the western pacific subsurface temperature buildup leading to ENSO events. *Geophysical Research Letter*, 42:2961–2967.

Ballester, J., Bordoni, S., Petrova, D., and Rodo, X. (2015b). B) Heat advection processes leading to El Niño events as depicted by an ensemble of ocean assimilation products. *Submitted for review*.

Ballester, J., Rodriguez-Arias, M. A., and Rodo, X. (2011). A new extratropical tracer describing the role of the western pacific in the onset of El Niño: Implications for ENSO understanding and forecasting. *Journal of Climate*, 24:1425–1437.

Barnston, A., Chelliah, M., and Goldenberg, S. (1997). Documentation of a highly ENSO-related SST region in the equatorial Pacific. *Atmos.-Ocean*, 35:367–383.

Barnston, A. and Ropelewski, C. (1992). Prediction of ENSO episodes using Canonical Correlation Analysis. *Journal of Climate*, 5:1316–1345.

Barnston, A., Tippett, M., L'Heureux, M., Li, S., and DeWitt, D. (2012). Skill of real-time seasonal ENSO model predictions during 2002-11. Is our capability increasing? *American Meteorological Society*, 93:631–651.

Bjerknes, J. (1969). Atmospheric teleconnections from the equatorial Pacific. *Monthly Weather Review*, 97:163–172.

Brown, J. and Fedorov, A. (2010). How much energy is transferred from the winds to the thermocline on ENSO time scales? *Journal of Climate*, 23:1563–1580.

Chen, D., Cane, M., Kaplan, A., Zebiak, S., and Huang, D. (2004). Predictability of El Niño over the past 148 years. *Nature*, 428:15.

Chen, D., Lian, T., Fu, C., Cane, M., Tang, Y., Murtugudde, R., Song, X., Wu, Q., and Zhou, L. (2015). Strong influence of westerly wind bursts on El Niño diversity. *Nature Geoscience*, 8:339–345.

CPC (2015). Cold and warm episodes by season. [http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml](http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml).

de Jong, P. (1991). The diffuse Kalman Filter. *Annals of Statistics*, 19:1073–1083.

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Eisenman, I., Yu, L., and Tziperman, E. (2005). Westerly wind bursts: ENSO’s tail rather than the dog? *Journal of Climate*, 18:5224–5238.

Fedorov, A., Harper, S., Philander, S. G., Winter, B., and Wittenberg, A. (2003). How predictable is El Niño? *Bull. Am. Meteorol. Soc.*, 84:911–919.

Fedorov, A. and Philander, S. G. (2001). A stability analysis of tropical ocean-atmosphere interactions: Bridging measurements and theory for El Niño. *Journal of Climate*, 14:3086–3101.

Gebbie, G. and Tziperman, E. (2009). Predictability of SST-modulated westerly wind bursts. *Journal of Climate*, 22:3894–3909.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., and Yiou, P. (2002). Advanced spectral methods for climatic time series. *Review of Geophysics*, 40:3.1–3.41.

Gill, A. (1985). Elements of coupled ocean-atmosphere models for the tropics. *Coupled Ocean-Atmosphere Models, Elsevier Oceanography Series*, 40:303–327.

Glantz, M. H. (2015). Shades of chaos: lessons learned about lessons learned about forecasting El Niño and its impacts. *International Journal of Disaster Risk Science*, 6:94–103.

Good, S. A., Martin, M. J., and Rayner, N. A. (2013). EN4: quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, 118:6704–6716.

Goswami, B. N. and Shukla, J. (1991). Predictability of a coupled ocean-atmosphere model. *Journal of Climate*, 4:3–22.
Hare, S. and Mantua, N. (2000). Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Progress in Oceanography*, 47:103–145.

Harvey, A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Harvey, A. and Koopman, S. J. (2000). Signal extraction and the formulation of unobserved components models. *The Econometrics Journal*, 3:84–107.

Harvey, A. and Shephard, N. (1993). Structural time series models. *Elsevier Science Publishers, Handbook of statistics*, 11.

Ishii, M., Shouji, A., Sugimoto, S., and Matsumoto, T. (2005). Objective analyses of SST and marine meteorological variables for the 20th century using COADS and the Kobe Collection. *International Journal of Climatology*, 25:865–879.

Izumo, T., Lengaigne, M., Vialard, J., Luo, J., Yamagata, T., and Madec, G. (2014). Influence of Indian Ocean Dipole and Pacific recharge on following year’s El Niño: interdecadal robustness. *Journal of Climate Dynamics*, 42:291–310.

Jiang, N., Neelin, D., and Ghil, M. (1993). Quasi-quadrennial and quasi-biennial variability in COADS equatorial Pacific sea surface temperature and zonal wind. *Proc 17th Clim Diagn Workshop. Climate Analysis Center, NOAA*, pages 348–353.

Jiang, N., Neelin, D., and Ghil, M. (1995). Quasi-quadrennial and quasi-biennial variability in the equatorial Pacific. *Journal of Climate Dynamics*, 12:291–310.

Jin, F. F. (1997a). An equatorial ocean recharge paradigm for ENSO. Part I: conceptual model. *Journal of Atmospheric Sciences*, 54:811–829.

Jin, F. F. (1997b). An equatorial ocean recharge paradigm for ENSO. Part II: a stripped down coupled model. *Journal of Atmospheric Sciences*, 54:830–847.

Jin, F. F., Kug, J. S., An, S., and Kang, I. S. (2003). A near-annual coupled ocean-atmosphere mode in the equatorial Pacific. *Geophysical Research Letters*, 30.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions, ASMA, Series D*, 82:35–45.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77:437–471.

Kang, I. S. and Kug, J. S. (2002). El Niño and La Niña sea surface temperature anomalies: Asymmetry characteristics associated with their wind stress anomalies. *Journal of Geophysical Research*, 107:4372.

Kleeman, R., McCreary, J., and Klinger, B. (1999). A mechanism for generating ENSO decadal variability. *Geophysical Research Letters*, 26:1743–1746.

Kondrashov, D., Kravtsov, S., Robertson, A., and Ghil, M. (2005). A hierarchy of data-based ENSO models. *Journal of Climate*, 18:4425–4444.

Krishnamurthy, L., Vecchi, G., Msadek, R., Wittenberg, A., Delworth, T., and Zeng, F. (2015). The seasonality of the great plains low-level jet and ENSO relationship. *Journal of Climate*, 28:4525–4544.

Lau, K. and Shen, P. (1988). Annual cycle, quasi-biennial oscillation and Southern Oscillation in global precipitation. *Journal of Geophysical Research*, 93:10975–10988.

Mantua, N. and Battisti, D. (1995). Aperiodic variability in the Zebiak-Cane coupled ocean-atmosphere model: air-sea interactions in the western equatorial Pacific. *Journal of Climate*, 8:2897–2927.

McGregor, S., Timmermann, A., Shneider, N., Stuecker, M., and England, M. (2012). The effect of the South Pacific Convergence Zone on the termination of El Niño events and the meridional asymmetry of ENSO. *Journal of Climate*, 25:5566–5586.

McPhaden, M. (2004). Evolution of the 2002/2003 El Niño. *American Meteorological Society*, 85:677–695.

McPhaden, M., Timmermann, A., Widlansky, M., Balmaseda, M., and Stockdale, T. (2014). The curious case of the El Niño that never happened: A perspective from 40 years of progress in climate research and forecasting.
McPhaden, M. and Yu, X. (1999). Equatorial waves and the 1997/98 El Niño. Geophysical Research Letters, 26:2961–2964.

Moron, V. and Plaut, G. (2003). The impact of El Niño-Southern Oscillation upon weather regimes over Europe and the North Atlantic during boreal winter. International Journal of Climatology, 23:363–379.

Neelin, J. (1990). A hybrid coupled general circulation model for El Niño studies. Journal of Atmospheric Sciences, 47:677–695.

Neelin, J. (1991). The slow surface temperature mode and the fast-wave limit: Analytic theory for tropical interannual oscillations and experiments in a hybrid coupled model. Journal of Atmospheric Sciences, 48:584–606.

Penland, C. (1996). A stochastic model of Indo-Pacific sea surface temperature anomalies. Physica D, 98:534–558.

Penland, C. and Magorian, T. (1993). Prediction of Niño3 sea surface temperatures using linear inverse modeling. Journal of Climate, 6:1067–1076.

Philander, S. (1989). El Niño and La Niña. Journal of Atmospheric Sciences, 47:451–459.

Philander, S., Yamagata, T., and Pacanowski, R. (1984). Unstable air-sea interactions in the tropics. Journal of Atmospheric Sciences, 41:604–613.

Ramesh, N. and Murtugudde, R. (2013). All flavours of El Niño have similar early subsurface origins. Nature Climate Change, 3:42–46.

Rasmussen, E. and Carpenter, T. (1982). Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/El Niño. Mon. Weather Rev., 110:354–384.

Rasmussen, E., Wang, X., and Ropelewski, C. (1990). The biennial component of ENSO variability. Journal of Marine Systems, 1:71–96.

Sarachik, E. and Cane, M. (2010). The El Niño Southern Oscillation Phenomenon. Cambridge University Press.

Stein, K., Timmermann, A., and Shneider, N. (2011). Phase synchronization of El Niño - Southern Oscillation with the annual cycle. Physical Review Letters, 107.

Thompson, C. J. and Battisti, D. S. (2000). A linear stochastic dynamical model of ENSO. Part I: Model development. Journal of Climate, 13:2818–2832.

Thompson, C. J. and Battisti, D. S. (2001). A linear stochastic dynamical model of ENSO. Part II: Analysis. Journal of Climate, 14:445–466.

Tong, H. (1990). Non-linear time series. Oxford University Press.

Torrence, C. and Webster, P. J. (1998). The annual cycle of persistence in the El Niño - Southern Oscillation. Quart. J. R. Meteorol. Soc., 124:1985–2004.

Trenberth, K. (1976). Spatial and temporal variations in the Southern Oscillation. Quart. J. R. Meteorol. Soc., 102:639–653.

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

Tziperman, E., Yu, L. (2007). Quantifying the dependence of westerly wind bursts on the large-scale tropical Pacific SST. Journal of Climate, 20:2760–2768.

Tziperman, E., Zebiak, S., and Cane, M. (1997). Mechanisms of seasonal - ENSO interaction. Journal of Atmospheric Sciences, 54:61–71.

Wyrtki, K. (1975). El Niño - the dynamic response of the equatorial Pacific Ocean to atmospheric forcing. Journal of Physical Oceanography, 5:572–584.

Wyrtki, K. (1985). Water displacements in the Pacific and the genesis of El Niño cycles. Journal of Geophysical Research, 90:7129–7132.

Xue, Y., Cane, M., Zebiak, S., and Blumenthal, M. (1994). On the prediction of ENSO: a study with a low-order Markov model. Tellus A, 46:512–528.

Yasunari, T. (1989). A possible link of the QBO’s between the stratosphere, troposphere and the surface temperature in the tropics. J. Meteorol. Soc. Jpn, 67.

Yeh, S. and Kirtman, B. (2004). The decadal ENSO variability in a hybrid coupled. Journal of American Meteorological Society, 17:1225–1238.
Yeh, S. and Kirtman, B. (2005). Tropical Pacific decadal variability and ENSO amplitude modulation. *Geophysical Research Letters*, 32.

Yu, J., Kao, H., and Lee, T. (2011). Subsurface ocean temperature indices for central-Pacific and eastern-Pacific types of El Niño and La Niña events. *Theoretical Applied Climatology*, 103:337–344.

Yu, J. and Mechoso, C. (2001). A coupled atmosphere-ocean GCM study of the ENSO cycle. *Journal of Climate*, 14:2329–2350.

Yu, X. and McPhaden, M. (1999). Seasonal variability in the equatorial Pacific. *Journal of Physical Oceanography*, 29:925–947.

Zebiak, S. (1985). Tropical atmosphere-ocean interaction and the El Niño/Southern Oscillation phenomenon. *PhD Thesis, Massachusetts Institute of Technology*.

Zebiak, S. (1989). Oceanic heat content variability and El Niño cycles. *Journal of Physical Oceanography*, 19:475–486.

Zebiak, S. E. and Cane, M. A. (1987). A model El Niño-Southern Oscillation. *Monthly Weather Review*, 115:2262–2278.