Long-term Forecasting of El Niño Events via Dynamic Factor Simulations*

Mengheng Li[(e)] *Siem Jan Koopman*[(a,b,c)†] *Rutger Lit*[(a)]
Desislava Petrova[(d)]

[(a)]Vrije Universiteit Amsterdam, The Netherlands
[(b)]Tinbergen Institute, The Netherlands
[(c)]CREATES, Aarhus University, Denmark
[(d)]Barcelona Institute for Global Health, Climate and Health Programme, Spain
[(e)]University of Technology Sydney, UTS Business School, Australia

October 8, 2018

Abstract

We propose a new forecasting procedure which particularly explores opportunities for improving the precision of medium and long-term forecasts of the Niño3.4 time series that is linked with the well-known El Niño phenomenon. This important climatic time series is subject to an intricate dynamic structure and is interrelated to other climatological variables. The procedure consists of three steps. First, a univariate time series model is considered for producing prediction errors. Second, signal paths of the prediction errors are simulated via a dynamic factor model for the errors and explanatory variables. From these simulated errors, ensemble time series for Niño3.4 are constructed. Third, forecasts are generated from the ensemble time series and their sample average is our final forecast. As part of these dynamic factor simulations, we also obtain the forecast of the El Niño event which is a categorical variable. We present empirical evidence that our procedure can be superior in its forecasting performance when compared to other econometric forecasting methods.

Some key words: Climate econometrics, Dynamic models, Kalman filter, Simulation smoothing, Factor models, Unobserved components, long-term forecast, Multivariate time series.

JEL classification: C32, C42.

*We would like to thank Xavier Rodó, Andreas Pick, Jörg Breitung, Julia Schaumburg, Onno Kleen, Eric Hillebrand, Felix Pretis, two anonymous referees and seminar and workshop participants at Vrije Universiteit Amsterdam, Erasmus University Rotterdam, Tinbergen Institute, the Conference on Climate Econometrics (Aarhus, 2016), and the Rhenish Multivariate Time Series Econometrics (Rotterdam, 2017) for useful comments and helpful suggestions on previous versions of this paper. We are indebted to the Editors and the referees for their supportive suggestions. Any remaining errors are the responsibility of ours alone. All computations are carried out by the object-oriented matrix language `Ox` of Doornik (2007) and the library of state space functions `SsfPack` of Koopman et al. (2008).

†Corresponding author: S.J. Koopman, Department of Econometrics, Vrije Universiteit Amsterdam, SBE, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. `s.j.koopman@vu.nl`

1 Introduction

El Niño is a well-known phenomenon in climate science and is characterised by higher than average sea surface temperatures in the central and eastern equatorial Pacific Ocean. It has a substantial impact on the climate in many parts of the world. Hence, it has been given much coverage in the popular media, and it is the subject of extensive research in the scientific world. El Niño typically causes changes in weather patterns related to temperature, pressure and rainfall. Thus, a warm event may not only have a negative impact on local economies, but can also have negative consequences for public health, as in some regions these changes increase substantially the risk of water-borne and/or vector-borne diseases. Given its huge impact particularly on some developing countries bordering the Pacific Ocean, it is self-evident that a timely forecast of the next El Niño event is important. Much scientific research has been devoted to the development of forecasting methods for El Niño. The oscillation is characterised by an irregular period of between 2 and 7 years. Currently, forecasts are issued regularly for up to three seasons in advance, but the long-term of more than one year ahead forecasts remain a real challenge, and have only been attempted in the domain of theoretical hindcast studies. At the same time, one of the two main theories about the physics underlying El Niño implies that it may be a self-sustaining climatic fluctuation that is quasi-periodic, with several dominant peaks in its spectrum, the main one being at about every 4-5 years, a secondary at about 2 years, and a third one at about 1.5 years. This suggests that it may be predictable at lead times of several years. We propose a new forecasting procedure for El Niño and explore its ability to produce accurate forecasts for medium to long lead times. The forecasting methods are based on state-of-the-art developments in time series econometrics including state space simulation methods.

An overview of the current operational empirical models for the ENSO predictions is presented in Barnston et al. (2012). These ENSO prediction models are based on different approaches including inverse modelling (Penland and Magorian, 1993), multiple polynomial regression (Kravtsov et al., 2005; Kondrashov et al., 2005), multiple regression (Clarke and Van Gorder, 2003; Knaff and Landsea, 1997), multivariate empirical orthogonal functions and Markov modelling (Xue et al., 1994, 2000), constructed analogues (Dool, 1994, 2007), canonical correlation analysis (Barnston and Ropelewski, 1992), and neurological network modelling (Tangang et al., 1997). All these models have been em-

pirically studied and compared with the state-of-the-art dynamical ENSO models. The overall conclusion is that their forecast performances are similar, for both short-term and long-term horizons; see the discussions in Barnston et al. (2012). The exceptions are cases where the target or start period is close to the months of March-May; this period is the season corresponding to the so-called spring predictability barrier; see Duan and Wei (2013). Hence, in an operational framework, the skill of all models is good for the predictions of about 2-3 seasons ahead and gradually decreases for longer lead times than 8-9 months. There are exceptions for some dynamical models.

In more recent studies with a retrospective forecasting focus, several models have demonstrated longer lead forecast capabilities extending to between 1 and 4 years ahead; see Chen et al. (2004), Ludescher et al. (2013), Petrova et al. (2017) and Gonzalez and Goddard (2016). Our study follows this line of research since our proposed method has not been tested operationally and has the limitation of being a study based on ENSO hindcasts, rather than on forecasts performed in real time. Chen et al. (2004) presents dynamic predictions of the strongest El Niño events in recorded history dating back to 1877 and at lead times of up to 24 months. Ludescher et al. (2013) use sea surface temperature anomalies, teleconnections and network methods to issue statistical predictions of the warm events in 1981-2011 at lead times of up to about 20 months. Petrova et al. (2017) use a dynamic components model and a number of specially designed subsurface temperature and wind stress predictors (also used in the present methodology) to predict the major EN events in 1996-2015 at lead times of up to 34 months. Furthermore, Petrova et al. (2018) show a prediction of the whole ENSO time series between 1972-2016 at a lead time of 2 years with an improved version of the model described in Petrova et al. (2017). Finally, Gonzalez and Goddard (2016) use decadal prediction experiments from the Coupled Model Intercomparison Project (CMIP5) to explore the long-term predictability of ENSO with different dynamical schemes. They show consistent skill for lead times of up to 24 months, but some ENSO events are predicted longer in advance (4 years and more).

In our study we address the classical forecasting problem in time series analysis where we need to forecast a single time series of interest that is subject to a possibly intricate serial correlation structure and for which a large set of explanatory variables is available. The standard approach is to consider a linear model that simultaneously treats the

dynamic structures in the time series and the explanatory variables, which may be endogenous or exogenous. Both in econometrics and statistics, much attention is given to the selection of the explanatory variables and their dynamic interactions with the variable of interest. The preferred methodology for model specification depends heavily on whether the aim is in-sample fit or out-of-sample forecasting. In the latter case, it is typically argued that working towards more parsimonious models can be beneficial in forecasting. For ENSO forecasting, Petrova et al. (2017) have adopted this standard approach with the unobserved components time series models of Harvey (1989) as their base model. In this study we explore and develop new multivariate methods to treat the explanatory variables specifically for the purpose of improving the ENSO forecasts from such univariate models.

We denote the variable of interest by the scalar y_t and the explanatory variables by the vector X_t for which we have observations at time points $t = 1, \ldots, T$ but we do not have observations available at "future" time points $T+1, T+2, \ldots$. We assume that the time series y_t and those in X_t are stationary but persistent processes. We abstract ourselves from non-stationary and cointegration issues in the current study although we formally test whether the time series can be treated as stationary. We further notice that X_t may also contain lagged explanatory variables. The aim is to produce an estimate for y_{T+h} where h is the forecast horizon, based on the available observations for y_t and X_t up to time T. For small values of h, we refer to the forecast made at time T, denoted by $\hat{y}_{T+h|T}$, as a *short-term forecast* while for larger values of h we refer to the forecast as a *long-term forecast*. It depends on the time series and the purpose of the study of what forecast horizon is associated with short, medium or long-term forecasts. Our focus is mostly directed towards long-term forecasting but our proposed procedure is also valid for short-term forecasting. To produce a forecast for y_{T+h}, the information in $X_{T+1}, \ldots, X_{T+h}$ may be highly relevant for it but we do not have their observations available at time T. A possible solution to this problem is to produce forecasts for the explanatory vector $X_{T+1}, \ldots, X_{T+h}$ in an ad-hoc way from which the actual forecast $\hat{y}_{T+h|T}$ can be computed. Given that these forecasts for the explanatory variables may be inaccurate, especially for larger values of h, the forecast $\hat{y}_{T+h|T}$ is often less accurate, even when compared to strictly univariate forecasts (no use of explanatory variables); see the discussion in Ashley (1988).

Another solution for the handling of explanatory variables in the forecasting problem

is to jointly analyze y_t and X_t within, for example, a vector autoregressive (VAR) model. As is known, there is a surge in the number of parameters that need to be estimated as the dimensionality of X_t increases. The estimation errors can negatively affect the forecast accuracy for the variable of interest, especially in the case of long-term forecasting; see, for example, the discussion in Litterman (1986). Recently, much attention is given to shrinkage methods applied to large sets of explanatory variables including principal components and empirical Bayes methods. In many studies where shrinkage procedures are adopted, convincing evidence of improved forecast accuracy is presented; see also the discussion in Stock and Watson (2012). In our current study we aim to build on these methods and develop a novel and parsimonious model-based forecasting procedure for ENSO time series and events. Our proposed method puts more weight on the main variable of interest by separating its modelling from that of the set of explanatory variables. This approach is related to the "targeted predictors" of Bai and Ng (2008) and the weighted likelihood method of Blasques et al. (2016). In these approaches, the modelling of the variable of interest is central while the treatment of the many explanatory variables is only aimed at improving the forecast accuracy of the variable of interest.

The remainder is organised as follows. We present and discuss our new three-step forecasting method in Section 2. We also provide an intuition and motivation of the overall forecasting method. A selection of the results of our empirical study is presented in Section 3. It includes the specification of the univariate model for the Niño3.4 time series and the role of the explanatory variables in the multivariate analysis. The main results of our forecasting study is presented in Section 4. Concluding remarks are in Section 5 together with suggestions for further research.

2 The three-step forecasting method

Our proposed forecasting method consists of three modelling stages: (i) analysis, modelling and prediction of the variable of interest, in our case Niño3.4, using a univariate time series model; (ii) a joint analysis of the prediction errors from step (i) and the explanatory variables, using a multivariate time series model; (iii) simulation of the prediction error series from step (i) conditional on the data set of explanatory variables using the multi-

variate model in step (ii); the simulated error series can be transformed to an ensemble time series of the variable of interest which can be forecasted using the univariate model of step (i) with the parameter re-estimated. The sample average of these ensemble forecasts is the final forecast. From this three-step method, confidence intervals and forecasts for the probability of El Niño and La Niña events can be obtained[1]. In this section, we discuss the three steps in detail. In Appendix A we provide a more technical motivation of the three-step method and we present the results of a small but intuitive Monte Carlo study.

The three-step forecasting procedure includes a parsimonious treatment of the many explanatory variables that all have short-term and long-term interrelationships with the variable of interest. In a dynamic linear regression analysis, it requires many coefficients for the explanatory variables and their lags. When applying shrinkage methods such as LASSO, the exclusion of coefficients is based on statistical decision making. Our approach is not concerned with the selection of variables but with the relevant measurement of combinations of signals from explanatory variables that cannot be explained by its own dynamics. These signals reflect the main and persistent dynamics in the explanatory variables that particularly facilitates medium- to long-term forecasting of the variable of interest. We could have opted for a complete dynamic factor analysis which assigns equal importance to the variable of interest and explanatory variables. With a mix of an univariate time series model that is parsimonious and a dynamic factor model that is flexible, the three-step procedure is able to provide accurate forecasts.

2.1 Step (i): univariate model

The time series properties of the variable of interest is first established by means of a univariate linear dynamic model. In our study we consider the class of unobserved components (UC) time series of Harvey (1989) but other linear dynamic models can also be considered, including those from the well-known class of autoregressive moving average (ARMA) models. We consider the monthly time series variable Niño3.4 denoted by y_t,

[1]The El Niño and La Niña events are treated as binary variables. For example, an El Niño event takes place if a consecutive sequence of temperature exceedances is observed. We show later how this complexity can be solved within the simulation procedure.

for time points $t = 1, \ldots, T$. We analyse y_t using the decomposition model as given by

$$y_t = \mu_t + \gamma_t + \psi_t + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{1}$$

where the dynamically evolving stochastic components are the trend μ_t, modelled as a random walk process, the seasonal γ_t, modelled as in Proietti (2000), the sum of p independent cyclical processes $\psi_t = \psi_{1,t} + \ldots \psi_{p,t}$, where the individual cycles $\psi_{j,t}$, for $j = 1, \ldots, p$, are modelled as in (Harvey, 1989, §2.3.3), and the Gaussian white noise disturbance ε_t with mean zero. Each dynamic component has a variance parameter while each cycle process also has parameters for persistence and frequency. Maximum likelihood estimation is done routinely using the Kalman filter; see Durbin and Koopman (2012). This UC model has been adopted for the forecasting of the Niño3.4 time series in Petrova et al. (2017) who extend model (1) by a careful selection of explanatory variables and who show that accurate forecasts can be obtained, especially for lead time of 1.5 to 2.5 years.

We denote the set $\{x_1, \ldots, x_t\}$ by $x_{1:t}$, for any variable x and for $t = 1, \ldots, T$. For a given UC model, the Kalman filter computes the one-step ahead prediction errors given by

$$v_t^{uc} = y_t - \mathrm{E}_{uc}(y_t | y_{1:t-1}), \tag{2}$$

where E_{uc} refers to the expectation under the conditional density $p_{uc}(y_t|y_{1:t-1})$ implied by the linear Gaussian UC model (1). We have $v^{uc} = (v_1^{uc}, \ldots, v_T^{uc})'$ which is defined by $v^{uc} = L\, y$, where $y = (y_1, \ldots, y_T)'$ and $L \in \mathbb{R}^{T \times T}$ is a $T \times T$ lower-triangular matrix with ones on the main diagonal; the lower elements of L are functions of the parameters in the UC model, see Durbin and Koopman (2012, Section 4.13). Since L is invertible, we also have $y = L^{-1} v^{uc}$. Hence y_t from the UC model (1) can be reconstructed as

$$y_t = g_t(v_{1:t}^{uc}), \qquad t = 1, \ldots, T, \tag{3}$$

where $g_t(v_{1:t}^{uc}) = \ell_t\, v^{uc}$ and ℓ_t is the t-th row of the lower-triangular matrix L^{-1}.

2.2 Step (ii): dynamic factor model

In the next step we jointly consider the prediction errors v_t^{uc} obtained from Step (i) and a set of $N-1$ explanatory or predictor variables $X_t \in \mathbb{R}^{N-1}$, for $t = 1, \ldots, T$. We assume that all time series in X_t are stationary, possibly after a transformation. For the joint analysis, we consider the dynamic factor model (DFM) as given by

$$\begin{bmatrix} v_t^{uc} \\ X_t \end{bmatrix} = \Lambda f_t + \xi_t, \qquad f_t = \Phi_1 f_{t-1} + \ldots + \Phi_r f_{t-r} + \nu_t, \tag{4}$$

where $\Lambda \in \mathbb{R}^{N \times p}$ is the factor loading matrix, $f_t \in \mathbb{R}^p$ is the vector of factors, $\xi_t \in \mathbb{R}^N$ is the Gaussian disturbance vector with mean zero and some (diagonal) variance matrix Σ_ξ, $\Phi_i \in \mathbb{R}^{p \times p}$ is the autoregressive coefficient matrix, for $i = 1, \ldots, r$, with r being the order of the vector autoregressive (VAR) process for f_t, that is $f_t \sim \text{VAR}(r)$, and $\nu_t \in \mathbb{R}^p$ is a Gaussian disturbance vector (independent of ξ_t) with mean zero and some variance matrix Σ_ν. The estimation of unknown parameters in the model is carried by the two-step method of Doz et al. (2011) under stationarity conditions; see discussion in Appendix A.

Due to the dynamic specification for factor $f_t \sim \text{VAR}(r)$ in (4), relations between the prediction errors v_t^{uc} and the explanatory variables in X_t are not exclusively contemporaneously but also exist through the lags of X_t. In particular, when considering the conditional density

$$p_{dfm}(v_t^{uc} | X_{1:T}), \qquad t = 1, \ldots, T, \tag{5}$$

we can measure the impact of $X_{1:T}$ on v_t^{uc} based on the dynamic factor model (4). To explore the features of the smoothed density, we will generate a large set of simulations from this density using the simulation smoother of Durbin and Koopman (2002). Each sample of v^{uc} from the smoothing density (5) is a linear function of $X_{1:T}$. This collection of time series sampled from (5) can measure the amount of variation that is implied by (4) and explained by $X_{1:T}$. We denote these simulated prediction error series from (5) by $v_t^{(i)}$ for $i = 1, \ldots, M$ where M is the number of simulated series. The current and lagged relations in $X_{1:T}$ may capture the remaining persistent and possibly cyclical dynamics in the time series variable of interest y_t.

In case the univariate UC model (2) is the correct model specification for the true

data generation process of y_t, the prediction errors in v^{uc} should not be affected by the variables in X_t. The implication for model (4) is that the first row of Λ should consist of only zeroes. Therefore, the inadequacies of the univariate UC model to describe the underlying dynamics of y_t will be brought to light by the smoothing density (5)[2]. Model (4) can be regarded as a parsimonious and simple way of linking X_t, including its current and lagged values, with the part of y_t that cannot be explained by its own past.

2.3 Step (iii): forecasting via simulation and estimation

Based on the set of simulated series $v_{1:T}^{(i)}$, for $i = 1, \ldots, M$, from Step (ii), we can construct a set of artificial time series y_t via the inverse transformation (3). We refer to this sequence of M time series y_t as the set of *ensemble time series* and is specifically generated by

$$y_t^{(i)} = g_t(v_{1:t}^{(i)}), \qquad v_t^{(i)} \sim p_{dfm}(v_t^{uc}|X_{1:T}), \qquad t = 1, \ldots, T, \tag{6}$$

for $i = 1, \ldots, M$. The ensemble time series $y_t^{(i)}$ is the result from a special interaction between the UC model and the DFM. This is a key feature of our forecasting method.

Since the univariate UC model is more parsimonious and solely targeted towards designing an optimal forecasting function for y_t, we generate the forecasts y_{T+h}, for $h = 1, 2, \ldots$, based on the univariate UC model (2), that is

$$\hat{y}_{T+h}^{uc} = \mathrm{E}_{uc}(y_{T+h}|y_{1:T}) = \mathrm{E}_{uc}(y_{T+h}|v_{1:T}^{uc}), \tag{7}$$

where the latter equality holds since $v_{1:T}^{uc}$ is the result of a nonsingular transformation of $y_{1:T}$, that is $v^{uc} = L\,y$. For each ensemble time series (6), we compute the forecasts in (7),

$$\hat{y}_{T+h}^{(i)} = \mathrm{E}_{uc_i}(y_{T+h}|y_{1:T}^{(i)}) = \mathrm{E}_{uc_i}(y_{T+h}|v_{1:T}^{(i)}), \qquad h = 1, 2, \ldots, \tag{8}$$

for $i = 1, \ldots, M$, where E_{uc_i} is expectation with respect to the UC model (1) but with the parameters in the UC model replaced by their ML estimates for the i-th ensemble time

[2] There are many reasons for the inadequacies of an analysis based on a univariate model including model misspecification and parameter estimation errors. Even when the prediction errors v_t^{uc} are white noise, it does not imply that $p_{dfm}(v_t^{uc}|X_{1:T}) = p_{dfm}(v_t^{uc})$.

series. The final forecast from our three-step method involving dynamic factor simulations (DFS) is then simply obtained by the average

$$\hat{y}_{T+h}^{dfs} = \frac{1}{M}\sum_{i=1}^{M} \hat{y}_{T+h}^{(i)}.$$

This final forecast can be regarded as the Monte Carlo estimate of the forecast function

$$\hat{y}_{T+h} = \int \mathrm{E}_{uc}(y_{T+h}|v)p_{dfm}(v|X_{1:T})\mathrm{d}v,$$

where $p_{dfm}()$ is defined above. A key insight is that the DFM from which we construct ensemble time series incorporates the predictive contribution of $X_{1:T}$ for y_t. It leads to the highly convenient property that there is no need to forecast X_t in any way.

2.4 Discussion

This three-step forecasting method is based on a univariate time series model for the variable of interest and on a multivariate time series model for linking the variable of interest with a set of predictor or explanatory variables. Both models are approximations to the true model that has generated the y_t and X_t variables. Hence we treat both models as misspecified while they still provide the basis for an accurate and feasible method of forecasting the variable of interest y_t. In particular, given the empirical evidence that univariate models are often highly effective in producing accurate short-term forecasts, we take step (i) as a solid basis for forecasting. The role of the dynamic factor model is to incorporate the (possibly many) explanatory variables in the dynamic variation of y_t that cannot be explained by its own past. Since the underlying dynamic factors are modelled as vector autoregressive processes and since the focus is on $p_{dfm}(v|X_{1:T})$, current and lagged interrelationships between the variables v and $X_{1:T}$ are also accounted for. All ensemble time series generated from the dynamic factor model are treated separately by first estimating the parameters and second by computing the forecasts, both for the univariate model. The average of these ensemble forecasts is then our final forecast. A technical account of the method is provided in Appendix A.

3 Empirical analysis of Niño3.4 series

3.1 The data

Our data set includes the monthly time series of temperature values which is referred to as the Niño3.4 time series and which is the area-averaged sea surface temperature in the region (5° N - 5° S, 170° W - 120° W). In this area the El Niño events are identified, see also the discussion in Barnston et al. (1997). The National Centers for Environmental Information (NOAA) defines an El Niño or La Niña event as a phenomenon in the equatorial Pacific Ocean characterised by a five consecutive 3-month running mean of sea surface temperature (SST) anomalies in the Niño 3.4 region that is above (below) the threshold of +0.5°C (-0.5°C)[3].

In our empirical study, the Niño3.4 time series denoted by y_t is the variable of key interest which is observed from January 1982 to the end of 2015 with 34 years of data and 408 monthly observations. For this period, observations for 24 predictor variables are available which consist of physical measures of zonal wind stress and sea temperatures at different depths in the ocean and at different locations. Petrova et al. (2017) give a detailed account of the selection of these variables. We do have observations available beyond 2015 but only for a few variables. Graphs, acronyms and references to data sources for all time series are presented in Appendix B.

3.2 The Niño3.4 time series

Step (i) of the forecasting procedure requires a univariate model for the Niño3.4 time series. Petrova et al. (2017) proposes to use the UC model with a stochastic trend, monthly seasonal and three cyclical components. From the sample spectrum, one can easily identify three or four peaks where the first one clearly corresponds to the monthly seasonality. To determine the appropriate specification for the UC model, we firstly examine its stationarity[4] using the well-known augmented Dickey-Fuller (ADF) and KPSS tests, see Phillips and Xiao (1998). The ADF test (with intercept) strongly implies that the null

[3]Details can be found on the website of NOAA, `https://www.ncdc.noaa.gov/`

[4]In our in-sample analysis, we add dummy variables in the UC model to account for short-term cooling effects due to a series of Montserrat Volcano eruptions in 1995 on the global temperature.

Table 1

UC model for Niño3.4 series

no. cycles	Standard deviations							Tests			
	σ_μ	σ_γ	σ_{ψ_1}	σ_{ψ_2}	σ_{ψ_3}	σ_{ψ_4}	σ_ϵ	BIC	JB	LB(36)	EW
1	0.04	0.00	0.13				0.00	-2.45	0.22 (0.89)	51.89 (0.02)	6.19 (<0.01)
2	0.00	0.00	0.18	0.10			0.00	-2.65	2.40 (0.30)	50.71 (0.02)	4.31 (<0.01)
3	0.00	0.00	0.11	0.09	0.13		0.00	-2.68	2.30 (0.32)	40.94 (0.11)	2.08 (0.12)
4	0.00	0.00	0.09	0.07	0.13	0.56	0.02	-2.69	1.82 (0.40)	39.11 (0.13)	1.83 (0.16)

Each row indicates a UC model with a specific number of stochastic cycle components. The estimates of standard deviations of the stochastic component disturbances are reported. BIC denotes the Bayesian information criterion. The residual test statistics are JB for the Jarque-Bera normality test, LB(36) for the Ljung-Box autocorrelation test based on 36 lags, and EW for the exponential Wald statistic of Vogelsang (1997) concerning structural breaks in the trend function, all with their p-values (in brackets). The p-value for the LB(36) test is adjusted for the degrees of freedom but given the discussion in Harvey (1989, §5.3), it needs to be treated as indicative.

hypothesis of a unit root is rejected. Based on the KPSS test, we cannot reject the null hypothesis of trend stationarity which suggests that the Niño3.4 time series is generated from a stationary process around a fixed trend. Within the KPSS test procedure, the estimated trend reduces to a constant. Secondly, we choose the number of cycles based on diagnostic residual statistics and on the structural break tests of Vogelsang (1997). We learn from Table 1 that models with one or two cycles are likely misspecified; for example, the Ljung-Box test statistics imply autocorrelated residuals. The exponential Wald test for a constant intercept suggests the existence of structural breaks[5]. We conclude that a model with two cycles is insufficient for capturing the underlying dynamics in y_t.

Models with three and four cycles both produce diagnostic test statistics that support the hypothesis of white noise residuals. We prefer the model with three cycles because the Bayesian information criterion is only slightly larger compared to a model with four cycles. Furthermore, the estimated fourth cycle is a noisier version of the estimated first cycle. In all practical terms, the fourth cycle appears to be redundant in the model. Finally, estimated component standard deviations in Table 1 suggest that the model with

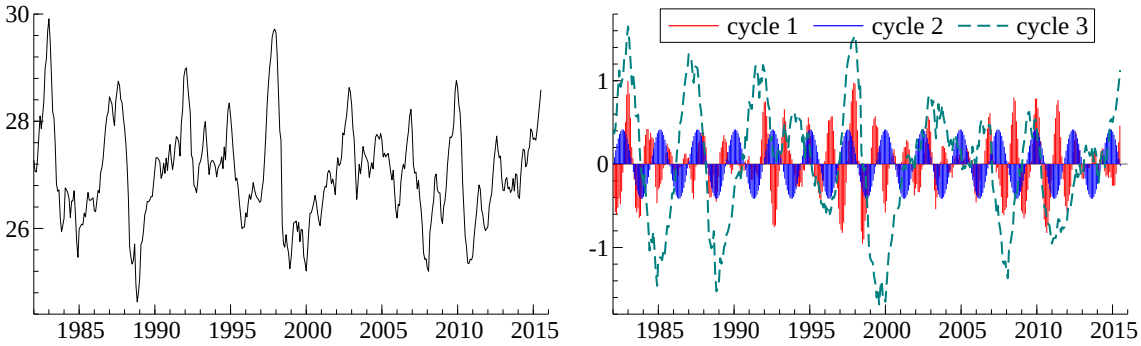
[5]We subtract the smoothed estimates of the seasonal and cyclical components from the Niño3.4 time series y_t, and regress it on constant intercepts; we follow Vogelsang (1997) in 1% trimming and p-value interpolation.

three cycles has all cycles estimated as stochastic while the estimated trend reduces to a fixed value (intercept). Also the estimated seasonal component is non-stochastic (reduces to fixed dummy effects) and the irregular noise components vanishes from the model. However, in our analysis we keep the UC model with stochastic components for trend, seasonal and irregular because it is agnostic to estimation uncertainty when considering rolling-window forecasting in our empirical study.

Figure 1 presents the estimated deseasonalised Niño3.4 series $y_t - \gamma_t$ and the three extracted cycles. The estimated standard deviations for the three stochastic cycles are reported in Table 1. The estimates for the damping factor (that determines the speed of mean reverting of a cycle) are given by 0.96, 0.99 and 0.98 and for the cycle frequency are given by 0.36, 0.21 and 0.12, respectively. These cycle parameter estimates correspond approximately to the three peaks (which are not associated with the seasonal oscillation) in the sample spectrum for the Niño3.4 time series; they indicate cycle periods of 1.5, 2.5 and 4 years. The first cycle corresponds to the short-term cyclical dynamics that accounts for more variation in y_t than the seasonal component. The second cycle accounts for bi-annual systematic variation in the time series but its amplitude is relatively small compared to the seasonal and other cyclical components. The third cycle component captures most of the variation in y_t since it has the highest amplitude.

Figure 1

Decomposition of Niño3.4 time series



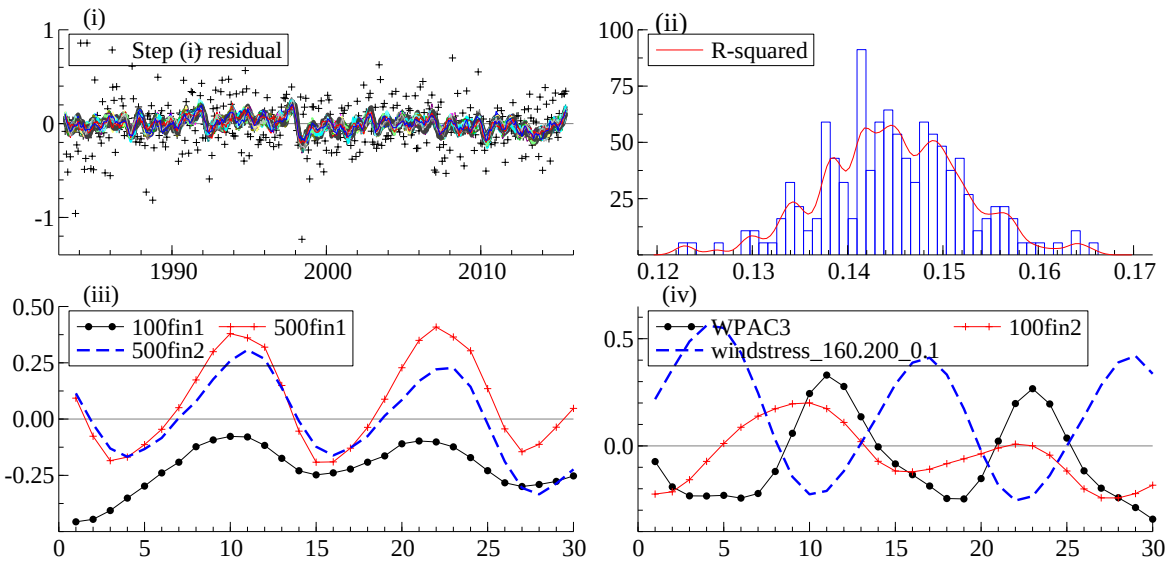
The left panel presents the deseasonalised Niño3.4 time series which is calculated by $y_t - \gamma_t$ with γ_t being the extracted seasonal component. The right panel shows the three extracted stochastic cycles with period 1.5, 2.5 and 4 years.

3.3 The ensemble residuals

To provide some more insights into the role of the ensemble time series in Step (iii), we illustrate some features of the ensemble residuals which are simulated from the smoothing density $p_{dfm}(v_t^{uc}|X_1,\ldots,X_T)$ in Step (ii). In Figure 2 we present a sequence of $M = 200$ ensemble prediction errors $v_t^{(i)}$, for $i = 1,\ldots,M$, together with the actual error v_t^{uc}. It is interesting to observe that all $v_t^{(i)}$s form a band that goes through the v_t^{uc}, despite the fact that unconditionally the residuals can be regarded as Gaussian white noise as is evidenced by the residual diagnostic statistics reported in Table 1. From graph (ii), we learn that the variation in a single $v_t^{(i)}$ series explains on average around 14% of the variation in v_t^{uc}. This highlights the amount of information on the underlying dynamics of y_t that a univariate model fails to capture but is recovered by the use of predictor variables through a dynamic factor simulation procedure.

Figure 2

Prediction and Ensemble residuals for Niño3.4 time series



The figure presents the one-step ahead prediction errors v_t^{uc} in Step (i). It also shows that variation of predictor variables in X_t are able to explain the variation of v_t^{uc}. (i): v_t^{uc} and the ensemble residuals $v^{(i)}$ obtained through the DFM in Step (ii). (ii): The histogram of R-squared of M regressions $v_t^{uc} = c+\beta v_t^{(i)}+\epsilon_t$. (iii) and (iv): Correlations between $v_t^{(i)}$ and a selection of variables in X_t, contemporaneously and for its lags from 1 to 30.

Figure 2 further illustrates the rich dynamic structures that the ensemble residuals capture from X_t. In order to investigate the individual contribution of a variable in X_t to $v_t^{(i)}$, we present the average correlation of $v_t^{(i)}$ with a selection of variables in X_t and their

lagged values, up to lag 30, in the graphs (*iii*) and (*iv*). The correlations are computed for each $v_t^{(i)}$ and an average is taken over the M ensemble residual series. This procedure is similar in design and has the same aim as the R^2 plots of Stock and Watson (2002, see their Figure 1) between principal components and individual explanatory variables. For example, we observe that "100fin2" continuously explains variability for the lags 10-30, which is expected as typically before an El Niño event takes place there is a subsurface anomalous warming of the ocean in the area where this variable is defined, which warming later propagates and plays an important role in the generation of El Niño in the eastern Pacific. Similarly, there is a high correlation for "500fin2" at 16-30 lags, again corresponding to this early subsurface warming of the ocean at greater depths. These variables were originally constructed with the purpose to capture important dynamical information about El Niño at its early developing stages and corresponding to long lead times. With our procedure we aim to extract this information, and use it efficiently and parsimoniously in our forecasting of the El Niño events.

4 Forecasting Niño3.4 time series and El Niño events

4.1 Design of the forecasting study

In the main part of our empirical forecasting study we consider the Niño3.4 time series as the variable of interest y_t that we want to forecast h-step ahead for $h = 1, 2, \ldots, 30$. Hence the maximum forecast window is 2.5 years. For producing the forecasts of y_t we adopt different forecast modelling approaches and methods, including our three-step forecasting procedure which we indicate by UCDFS, the unobserved components model with dynamic factor simulations. To compare the forecasting performance of the different methods, we carry out a rolling-window forecasting study with each window covering 275 monthly observations. The first window starts in May 1982 and we use the observations in the estimation window to estimate the parameters and use these to compute the forecasts $\hat{y}_{T+h|T}$ for $h = 1, 2, \ldots, 30$, using different methods. We repeat this for the 100 windows and obtain 100 out-of-sample forecast errors for each horizon h. We then compute measures for the forecast precision for each forecast length h.

The results for our UCDFS method are the focus of our discussions. We compare different measures of its predictive accuracy with those obtained from alternative model-based forecasting methods. Apart from two autoregressive moving average models, an AR(6) and an ARMA$(4,3)$, and the local level model (LL), the following models are considered: (i) the seasonal autoregressive moving average (SARMA) model with seasonal AR and MA orders both equal to 12 and with regular AR order 2 and MA order 1 (these orders are determined by the in-sample Bayesian information criterion, BIC), denoted by SARMA$(12,2,1)$; (ii) the unobserved components time series model (UCM) that is used in Step (i) of our three-step forecasting procedure; (iii) the unrestricted vector autoregression of order 2, or VAR(2) for the observation vector $(y_t, X_t')'$ (VAR order is determined by in-sample BIC); (iv) the forecasting procedure of Stock and Watson (2002), referred to as S&W, which is based on 5 to 7 principal components that summarise 95% of the variation in $X_{1:T}$, across all rolling windows; (v) the standard dynamic factor model (DFM) with observation vector $(y_t, X_t')'$ and with five dynamic factors modelled as stochastic cycles[6]; (vi) the least absolute shrinkage and selection operator (LASSO) selecting predictors from the collection $(X_t', X_{t-1}', \ldots, X_{t-36}')'$, i.e. 888 series with up to 36 lags as suggested by Petrova et al. (2017)[7]; (vii) the collapsed dynamic factor model (CDFM) of Bräuning and Koopman (2014) that collapses the 889-dimensional vector $(y_t, X_t', X_{t-1}', \ldots, X_{t-36}')'$ via five principal components and uses a DFM based on y_t and the principal components. We apply the DM test for equal predictive accuracy between UCDFS and each of the model-based forecasting methods described above.

In the last part of our empirical forecasting study, we carry out a forecast study for the probability of El Niño events. These forecasts are constructed via simulation and are based on the same methods described above.

4.2 Forecasting loss functions and precision criteria

We measure the precisions of the h-step ahead out-of-sample forecasts using different loss functions. We adopt index i to indicate a particular rolling window. To measure the

[6]The model is cast directly into state space form and the loading, persistence and covariance matrices are estimated by maximum likelihood using the Kalman filter.

[7]The LASSO threshold or tuning parameter is chosen by BIC and the typical number of nonzero coefficients is 25.

predictive accuracy, we consider the loss differential function

$$d_{i,h} = L(e_{i,h}^{(j)}) - L(e_{i,h}^{(k)}),$$

for some loss function $L(\cdot)$ and h-step ahead forecast errors $e_{i,h}^{(j)}$ and $e_{i,h}^{(k)}$ obtained from model j and k, respectively. We compute the DM test statistic of Diebold and Mariano (1995) using the loss differentials and corrected for small sample size as proposed by Harvey et al. (1997); that is

$$DM_h^* = \sqrt{\frac{M + 1 + h^2 - 3h}{M}} DM_h \ \sim \ \text{Student's } t_{M-1},$$

where DM_h is the standard DM test statistic for the h-step ahead forecast loss differential with the heteroskedastic and autocorrelation-consistent estimator for the asymptotic variance of DM_h, see Giacomini and White (2006) for a discussion.

In Table 2 we present the results of our El Niño forecasting study. The upper panel presents the h-step root mean squared forecast error (RMSE) of model j as given by

$$RMSE_h^{(j)} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} L(e_{i,h}^{(j)})}, \qquad h = 1, \ldots, 30,$$

with loss function $L(e_{i,h}^{(j)}) = {e_{i,h}^{(j)}}^2$. The lower panel presents the h-step ahead mean linex forecast error $MLFE_h^{(j)}$ of model j as given by

$$MLFE_h^{(j)} = \frac{1}{M} \sum_{i=1}^{M} L(e_{i,h}^{(j)}), \qquad h = 1, \ldots, 30, \tag{9}$$

with the linex loss function $L(e_{i,h}^{(j)}) = \exp(\beta e_{i,h}^{(j)}) - \beta e_{i,h}^{(j)} - 1$. This loss function is originally proposed by Varian (1975), which is a combined measure of loss in point forecast and forecast direction of change. The parameter β measures the aversion towards either negative ($\beta > 0$) or positive ($\beta < 0$) forecast errors. We choose $\beta = 1$ since an underestimation of the Niño3.4 time series increases the probability of missing an El Niño event.

Table 3 presents the mean ranked probability score (MRPS). It summarises the loss

resulted from mistakenly forecasting the events of El Niño and La Niña. The rank probability score (RPS) statistic of Epstein (1969) is given by

$$RPS_{i,h}^{(j)} = \frac{1}{2}\sum_{k=1}^{3}\left(\sum_{J=1}^{k}(\hat{p}_{J,i,h}^{(j)} - e_{J,i,h})\right)^2,$$

where $J = 1$ indicates an El Niño event, $J = 3$ a La Niña event, and $J = 2$ corresponds to no event. The probability $\hat{p}_{J,i,h}^{(j)}$ is model j's h-step ahead forecast probability for each of the three categories $J = 1, 2, 3$ for estimation window i. The binary variable $e_{J,i,h}$ indicates the actual outcome $(0, 1)$ for each of the three categories. This scoring rule is sensitive to distance by taking into account the probability mass assigned to all categories and not only the category of the observed outcome. $MRPS_h^{(j)}$ is given by $\frac{1}{M}\sum_{i=1}^{M} RPS_{i,h}^{(j)}$. We regard it as a precise indicator of the quality of event forecasts. We extend each of the ensemble series $y_t^{(m)}$, for $m = 1, \ldots, M$, obtained in Step (iii) of UCDFS with 30 missing values. Afterwards, the simulation smoother is used to draw N series which gives us a cloud of $h = 1, \ldots, 30$ step ahead forecasts of size MN. Once we have simulated the M ensemble forecasts, we can count the number of times an El Niño event, La Niña event, or no event occurs. The forecast probability $\hat{p}_{j,i,h} = \frac{1}{M}\sum_{m=1}^{M}\hat{p}_{j,i,h}^{(m)}$ is simply the number of times each event occurs divided by the total number of draws with $\hat{p}_{j,i,h}^{(m)}$ being the forecast probability from the ensemble series $y_t^{(m)}$. The confidence interval for forecast probability is thus given by $\hat{p}_{j,i,h} \pm Z_\alpha\sqrt{\mathrm{Var}(\hat{p}_{j,i,h})}$, where Z_α is the α-th quantile of standard normal distribution and the variance $\mathrm{Var}(\hat{p}_{j,i,h})$ is computed according to the *law of total variance* as

$$\mathrm{Var}(\hat{p}_{j,i,h}) = \frac{1}{M}\sum_{m=1}^{M}\hat{p}_{j,i,h}^{(m)}\left(1 - \hat{p}_{j,i,h}^{(m)}\right) + \frac{1}{M}\sum_{m=1}^{M}\left(\hat{p}_{j,i,h}^{(m)} - \hat{p}_{j,i,h}\right)^2,$$

In Figures 6 and 7 we visualise forecast probabilities for three selected El Niño events.

4.3 Point forecast and forecast direction of change

The main results of our forecasting study are presented in Table 2; its structure and contents are as follows. A cell with a value but without a symbol (except for column with header UCDFS) indicates that our proposed UCDFS forecasting method outperforms the corresponding forecasts from other model-based methods at a 5% level. A shaded cell

indicates that UCDFS underperforms another model at the 5% level. A single asterisk "$*$" indicates that UCDFS outperforms another model at 10% level, whereas double asterisks "$**$" indicate no significant difference in predictive accuracy between UCDFS and another model.

In terms of squared forecast errors, which is a measure for point forecast accuracy, UCDFS significantly outperforms all other multivariate models from 5-step ahead onwards, except for LASSO and VAR which are inferior to UCDFS at the 10% level for $h = 7$, 8, 9 and 29 (not reported). The VAR model produces accurate forecasts for small h, surpassing the UCDFS method at $h = 1$ and 3; however its forecast precision quickly deteriorates as h increases. The accurate forecasts of UCDFS for long horizons are of crucial importance in the context of El Niño forecasting as it can facilitate a timely warning system. For $h = 1$, the null of equal predictive accuracy in terms of squared forecast errors between CDFM and UCDFS cannot be rejected which may underline the role of parsimony in forecasting since both methods rely on dimension-reduction methods and principal components. The automatic variable selection method LASSO also achieves parsimony, but its forecasting performance is close to the CDFM except for some minor improvements when $h > 20$. For the selected tuning parameter in LASSO, the number of nonzero coefficients is similar to the number of factors we extract in CDFM, where both models start with 888 explanatory variables. The S&W method also relies on principal components but is the worst-performing method in our study in terms of $RMSE_h$; we see values that double those of UCDFS. Figure 3 presents the ratios of RMSE between the different models and the UCDFS method; the relative RMSE of S&W increases rapidly, indicating that the method becomes less reliable as h increases. The UCDFS method does not outperform other models for small h; this is due to the maximum likelihood estimation of the parameters in other models (estimation is done via the prediction error decomposition using the Kalman filter); hence the adopted parameters for forecasting are optimised for minimising the 1-step ahead forecast errors. This is not the case for the UCDFS method, as is seen in Step (ii) of the UCDFS method.

We learn from Table 2 that generally univariate models deliver better point forecasts than multivariate models. As pointed out, 1-step ahead forecast error minimisation when maximising the Gaussian likelihood function leads simple models such as AR(6)

Table 2

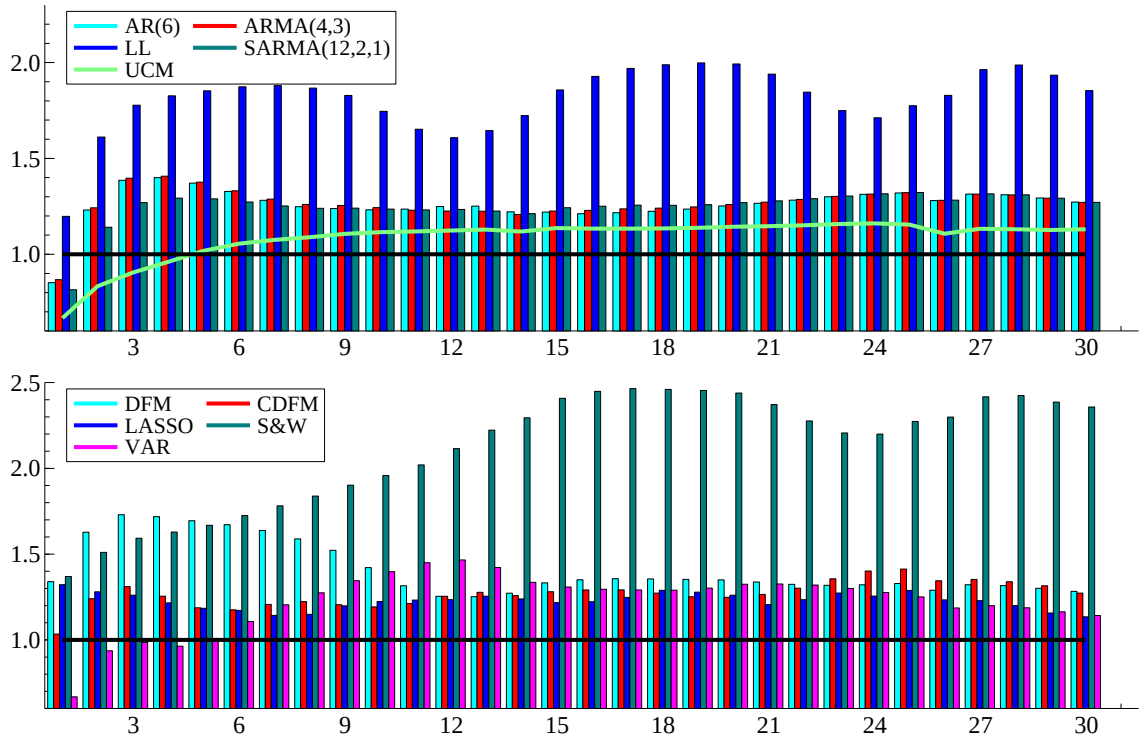
Forecast Accuracy Statistics

h	AR	ARMA	LL	SARMA	UCM	DFM	CDFM	LASSO	S&W	VAR	UCDFS
1	0.3089	0.3145	0.4343	0.2954	0.2349	0.4857*	0.3749**	0.4792	0.4966	0.2422	0.3626
2	0.6014**	0.6071*	0.7874	0.5574**	0.3956	0.7953*	0.6059*	0.6255*	0.7378	0.4576**	0.4886
3	0.8281	0.8347	1.0624	0.7584	0.5251**	1.0339	0.7834	0.7534	0.9518	0.5890	0.5976
5	1.0291	1.0328	1.3902	0.9673	0.7417**	1.2718	0.8909	0.8887**	1.2519	0.7567**	0.7505
10	1.0197	1.0290	1.4448	1.0226	0.8961	1.1765	0.9870	1.0130	1.6205	1.1566	0.8277
15	0.9921	0.9964	1.5100	1.0106	0.8973	1.0833	1.0414	0.9897	1.9583	1.0640	0.8131
20	0.9872	0.9932	1.5715	1.0009	0.8756	1.0644	0.9845	0.9939	1.9235	1.0446	0.7886
25	0.9922	0.9934	1.3342	0.9942	0.8430*	0.9990	1.0621	0.9684	1.7092	0.9403	0.7518
30	1.0565	1.0552	1.5392	1.0553	0.9114	1.0653	1.0568	0.9419	1.9577	0.9484	0.8304
1	0.0580	0.0606	0.0932**	0.0441	0.0383	0.0819**	0.0709**	0.1917	0.0713**	0.0274	0.0712
2	0.2330**	0.2383**	0.3204*	0.1699**	0.1174	0.3428	0.2160	0.3085	0.1970	0.0940	0.1353
3	0.4799*	0.4952**	0.6336	0.3462**	0.2176**	0.3929*	0.3976	0.4154	0.4223	0.1521	0.2170
5	0.9492**	0.9811**	1.4089	0.7300*	0.5101*	0.8587*	0.5985	0.5409	1.0239	0.2705*	0.3827
10	1.0804	1.1438	2.1526	1.0091	0.8741*	0.7926	0.7880	0.5497**	1.4534	1.2208	0.5095
15	1.0085	1.0189	1.9107	0.8892*	0.8951	0.9840	0.8416*	0.5079**	2.9457	1.0798	0.5103
20	0.9824	1.0017*	2.0235	0.8516	0.8765*	0.7568	0.8260*	0.6662	2.7628	0.9976	0.5121
25	1.0095	1.0129	1.0968	0.8448**	0.8934	1.1381	0.9428	0.6287**	2.4299	0.8016	0.4501
30	1.0726	1.0687	1.8031	0.8905**	0.9253	0.7196	0.8846*	0.6794	1.5882	0.6683	0.4790

The table reports the root mean squared forecast error $RMSE_h$ (upper panels) and the mean linex forecast error $MLFE_h$ (lower panels) for univariate models (left hand panels) and multivariate models (right hand panels) introduced in Section 4.1. A cell without a symbol (except for the column with header UCDFS) means the DFS method outperforms another model at the 5% level. A shaded cell indicates that UCDFS underperforms compared to another model at the 5% level. Single asterisk "$*$" indicates that UCDFS outperforms another model only at the 10% level. Double asterisks "$**$" indicate no significant difference in predictive accuracy between UCDFS and another model. AR is AR(6), ARMA is ARMA$(4,3)$ and SARMA is SARMA$(12,2,1)$.

Figure 3

Relative RMSE of different models against UCDFS



The panels show the RMSE forecast accuracy statistics in ratios of those for UCDFS, for selected models (upper panel univariate models and lower panel multivariate models) and for different lead times h. A ratio higher than 1 means underperformance compared to UCDFS.

and ARMA$(4, 3)$ to outperform UCDFS, for small h. Although univariate models have simple dynamic structures, the ARMA structures are flexible enough to pick up hidden signals in the Niño3.4 series. From the upper panel of Figure 3, we can observe that the SARMA$(12, 2, 1)$ model produces more accurate forecasts than AR(6) and ARMA$(4, 3)$ models for the short horizons, while their long-term forecasts have the same precision as the forecasts converge to their corresponding unconditional means, as h increases. Our main motivation to opt for the UCM as the univariate model in Step (i) of the UCDFS procedure is its delivery of the most accurate point forecasts for all h, among all univariate models considered. Comparing multivariate models for $h > 6$ in Figure 3, we find that persistently significant improvements are achieved by the proposed UCDFS procedure. These findings are encouraging since we compare the UCDFS results with competitive methods from the multivariate literature. Furthermore, the reported forecast precisions are generally lower than those reported in studies using operational empirical models for

ENSO predictions; see Barnston et al. (2012) and Ludescher et al. (2013).

To verify whether the forecast rankings, as implied by Figure 3, are constant over time, we consider the fluctuation test of Rossi and Sekhposyan (2016). Specifically, we compare the forecast performance between our preferred UCDFS method and the two best-performing competitors UCM and LASSO, in the out-of-sample period. The test is based on the simple regression

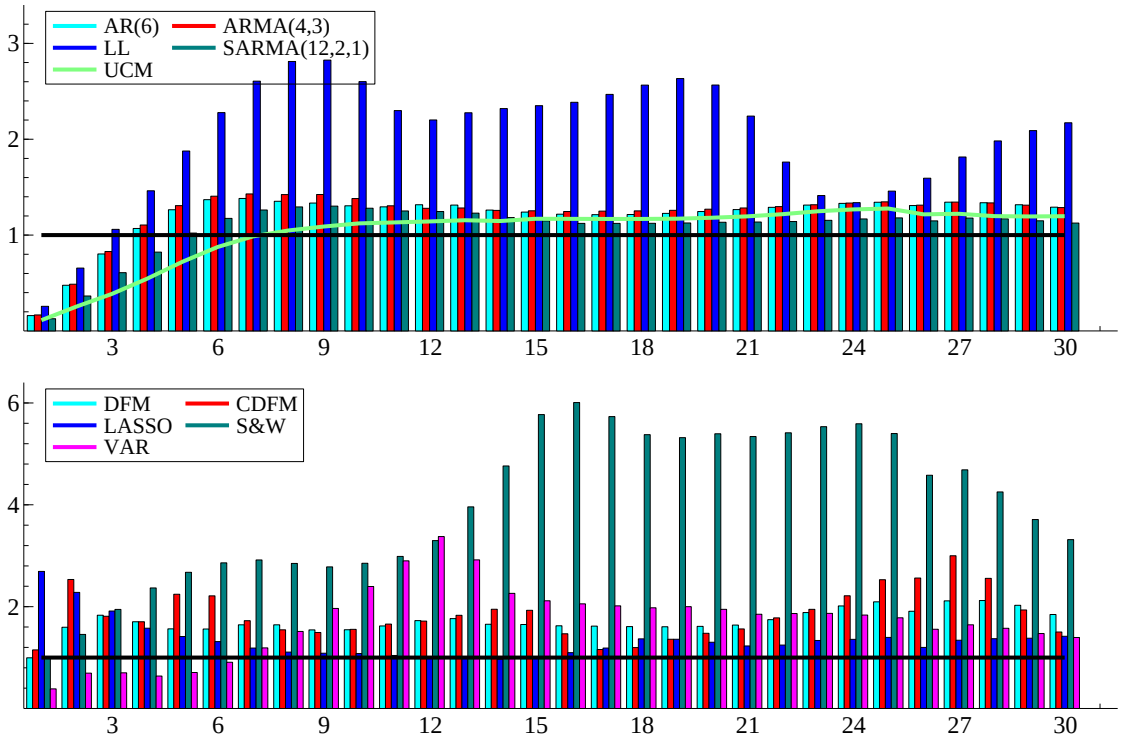
$$e_{i,h}^{(j)} = a + b \cdot \hat{y}_{i,h}^{(\text{UCDFS})} + c \cdot \hat{y}_{i,h}^{(j)} + \text{error}, \qquad j \in \{\text{UCM, LASSO}\}, \tag{10}$$

where $\hat{y}_{i,h}^{(k)}$ is the forecast from method k = UCDFS,UCM,LASSO, and a (intercept), b and c are regression coefficients. The regression is carried out for each forecast horizon h. The significance of the regression coefficient c indicates the usefulness of UCDFS on top of the other model j. We have 100 forecast errors for each method. By taking a regression sample length of 30 (here we have $i = 1, \ldots, 30$), we obtain 70 moving windows. The Wald statistic is considered for testing the significance of c. The supremum of the 70 Wald statistics is taken as the fluctuation test. We present the results in Appendix C. The results reveal that the UCDFS forecasts persistently outperform those of LASSO, with the exception of $h = 5$. Also UCDFS outperforms UCM for forecast horizons $h > 5$, especially for the moving windows before 2015. This finding highlights the long-term information advantage when using our dynamic factor simulation method.

The lower panels in Table 2 present the forecasting performances in terms of the linex loss criterion. From the comparisons among multivariate models, it follows immediately that the VAR model performs well in forecasting the direction of change in the short-term. UCDFS outperforms the other models for most multi-step ahead forecasts. The bottom panel of Figure 4 presents MLFE ratios for considered multivariate models. We notice that LASSO is quite competitive, especially for forecast horizons $h > 6$. For medium-term forecasts, the DM test for the MLFE fails to find a statistical difference between the accuracies of the LASSO and UCDFS forecasts. The VAR forecasts are poor for values of h around 12 while the S&W forecasts are overall of poor quality for $h > 6$. For the univariate models, MLFE leads to the same conclusions as those for RMSE. The three ARMA and UCM models perform well in terms of MLFE, for all h. The forecast accuracy

Figure 4

Relative MLFE of different models against UCDFS



The panels show the MLFE forecast accuracy statistics in ratios of those for UCDFS, for selected models (upper panel univariate models and lower panel multivariate models) and for different lead times h. A ratio higher than 1 means underperformance compared to UCDFS.

of SARMA$(12, 2, 1)$ surpasses the accuracy of the UCDFS method for $h = 1, 2, 3, 4$ and is even slightly higher than the UCM accuracy for $h > 15$.

Finally, the three steps of the UCDFS method can be implemented in different ways. To investigate the impact of different implementations on forecast accuracy, we also have considered alternative UCDFS implementations for the Niño3.4 time series. We can regard this as a robustness check for our three-step forecasting method. The results of this extended study are reported in Appendix D. We can conclude that other implementations of the UCDFS method only lead to small differences in our forecasting results.

4.4 El Niño event forecast

While it is an important task to forecast the Niño3.4 temperature accurately, a major challenge is to forecast future El Niño events (as well as La Niña and neutral events) accurately. To compute forecasts for these event probabilities, we rely on simulations of

Table 3

MRPS for a selection of Models

h	VAR	DFM	ARMA	SARMA	UCM	UCDFS
1	0.0884	0.0965	0.1748	0.1126	0.0977	0.1004
2	0.1079	0.1184**	0.1856	0.1386	0.1180**	0.1175
3	0.1239**	0.1403*	0.1968	0.1613	0.1357**	0.1333
5	0.1564**	0.1813	0.2172	0.1985	0.1683**	0.1655
10	0.2003	0.1935	0.1694	0.2071	0.1795	0.1533
15	0.2007	0.1960	0.1713	0.2115	0.1759	0.1582
20	0.1951	0.1911	0.1714	0.2094	0.1676	0.1518
25	0.1757	0.1850	0.1686	0.1921	0.1608	0.1467
30	0.1195*	0.1222	0.1172*	0.1331	0.1222	0.1092

We report the mean ranked probability score loss $MRPS_h$ for a selection of models introduced in Section 4.1, the acronyms used are in Table 2. ARMA is ARMA$(4,3)$ and SARMA is SARMA$(12,2,1)$.

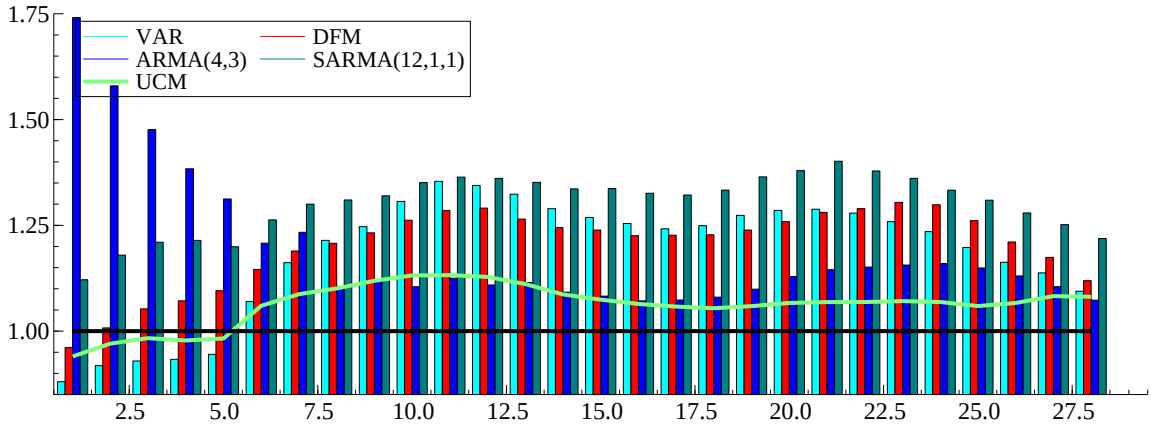
future Niño3.4 sample paths from which the probability of a future El Niño event can be calculated. All considered models are thus cast into state space form and h-step ahead forecast sample paths are drawn simultaneously with the ensemble time series using the simulation smoother that treats future values as missing. Table 3 summarises the MRPS for a selection of models. We do not include the results for CDFM, LASSO and S&W; for these models the computation of MRPS is prohibitively time-consuming because for each different h a different model needs to be considered.

From Table 3 we learn that the multivariate models produce more accurate short-term forecasts for an El Niño event in terms of MPRS. Interestingly, as concluded previously, univariate models are preferred for the short-term forecasting of the Niño3.4 time series in terms of RMSE and MLFE. For example, when $h \leq 3$, the VAR model provides the smallest MRPS with values 0.09, 0.11 and 0.12, while those from ARMA(4,3) are nearly doubled. For the medium and long-term forecasting of event probabilities, the UCDFS method is convincingly the most accurate method. Finally, the relative MRPS values, with respect to UCDFS, are shown in Figure 5. It reveals that the dynamic factor simulations in Step (ii) of UCDFS and the re-estimation for each ensemble series in Step (iii) improve the forecasting performance of UCM from $h = 6$ onwards. The ARMA and SARMA event forecasts are clearly less accurate for all values of h.

The event probability forecasts can be used in a real-time assessment of possible El

Figure 5

Relative MRPS of different models against UCDFS



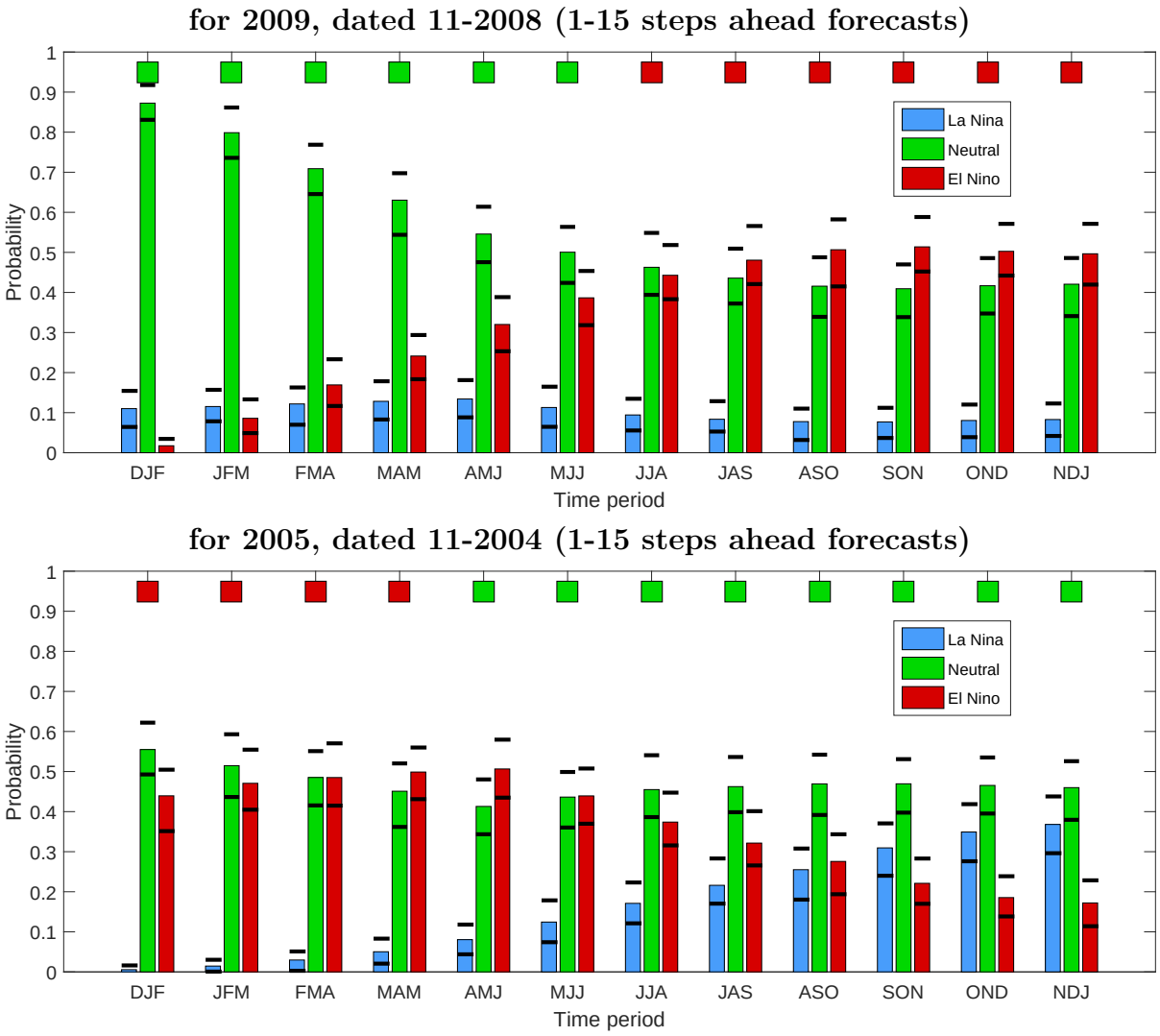
This figure presents the ratio of mean rank probability scores (MRPS) between selected models and UCDFS for different lead times h. A ratio higher than 1 means underperformance compared with UCDFS.

Niño events in the future. It is common practice to consider forecasts of an El Niño event taking place within a three-month period. In our first real-time illustration we compute forecasts in November 2008 for twelve moving three month-periods in 2009. The forecast for the first three-month period (December-February, as indicated by DJF) is constructed from the forecast paths for $h = 1, 2, 3$, January-March (JFM) is for $h = 2, 3, 4$, etc. The first half of 2009 has shown neutral events and the UCDFS forecasts imply "no event" as most probable for the first six three-month periods in 2009. In the second half of 2009, El Niño events take place (more in 2010) and the forecasts show increasing probabilities of an El Niño event. We have repeated this exercise for 2005 (for which the forecasts are computed in November 2005). The first 4 periods in 2005 experience El Niño events with UCDFS forecasting an El Niño event or "no event". The remaining 8 periods of 2005 have "no events" which are mostly in line with what the UCDFS forecasts have indicated.

Another real-time assessment of the UCDFS method is to track the sequence of forecasts of an El Niño event probability throughout a longer period. The three-month period of December-February 2010 has witnessed an El Niño event. In Figure 7 we present the UCDFS forecasts that are computed from 26 months (October 2007) through to 5 months (July 2009) ahead from the actual El Niño event. It is remarkable to see that the DJF 2010 El Niño event has been forecasted already accurately in December 2007 and in the following four three-month periods. The evidence in 2008 became somewhat more subtle

Figure 6

Three-Months Period Probability Forecasts

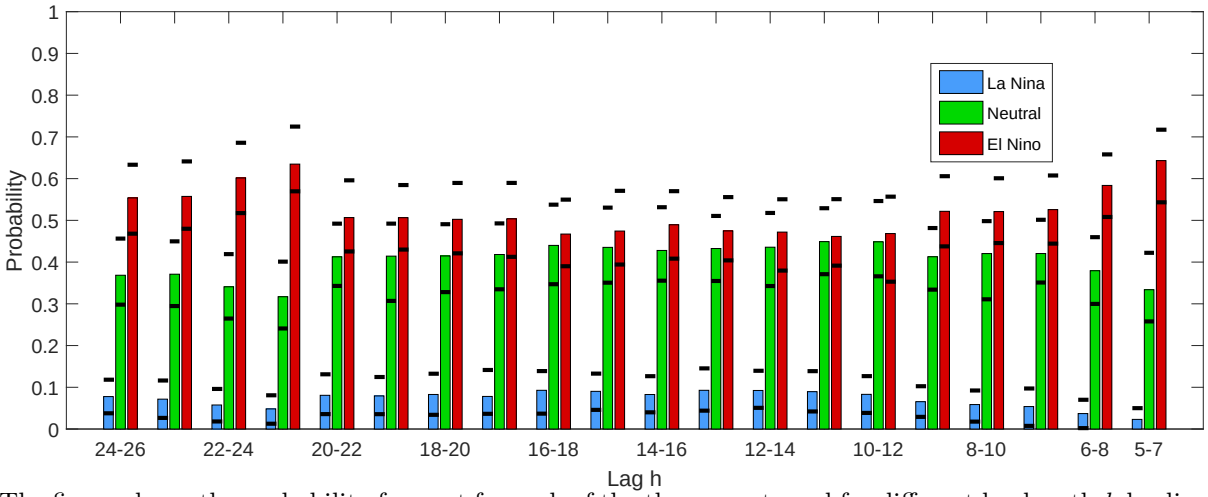


The figures present probability forecasts of November 2008 for the "El Niño", "La Niña" and "No" events in the moving three-months periods of 2009 (upper panel) and those of November 2004 for the events in 2005 (lower panel) with black lines indicating the 95% confidence interval. The boxes at the top of each panel indicate the actual event in a period.

although the forecasted probability of an El Niño event in DJF 2010 prevails for each forecast. The information becoming available from the beginning of 2009 has made the UCDFS method to produce forecasts that strongly suggest an El Niño event in DJF 2010. These results illustrate that the UCDFS method is able to forecast an event consistently over a long period of time.

Figure 7

Probability forecasts for DJF 2010 (an El Niño event)
from December 2007 ($h = 24$-26) to July 2009 ($h = 5$-7)



The figure shows the probability forecast for each of the three events and for different lag length h leading up to the El Niño event of DJF (2010).

5 Conclusions

We have proposed a novel procedure for the long-term forecasting of the Niño3.4 time series and whether El Niño events take place in pre-defined periods. The Niño3.4 time series has a complex dynamic structure and is interconnected with many climatological variables, and including their lagged dependencies. In our forecasting procedure, we do not need an elaborate selection procedure to determine which predictor variables and which lag structures are required to generate forecasts. In the three-step forecasting method, the dynamic factor model is adopted to produce simulated paths of prediction errors associated with a univariate time series model. Each simulated path of prediction errors is drawn conditionally on all predictor variables and transformed into an ensemble time series of Niño3.4. Hence all predictors (current and lagged observations) make a contribution to recover information relevant to the original series. It turns out that the reconstructed ensemble series contain rich dynamical information that facilitates the computation of accurate long lead forecasts. We provide empirical evidence with various robustness verifications for the forecasting of El Niño events in line with standard practice in climate research. We have given both intuitive and technical accounts of the proposed three-step forecasting method. In future research, we can explore other implementations of the method and other forecasting applications in fields such as economics and finance.

References

Ashley, R. (1988). On the relative worth of recent macroeconomic forecasts. *International Journal of Forecasting 4*, 363–376.

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 304–317.

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

Barnston, A. and C. Ropelewski (1992). Prediction of ENSO episodes using canonical correlation analysis. *Journal of Climate 5*, 1316–1345.

Barnston, A., M. Tippett, M. L'Heureux, S. Li, and D. DeWitt (2012). Skill of real-time seasonal ENSO model predictions during 2002-11: Is our capability increasing? *Bulletin of the American Meteorological Society 93(5)*, 631–651.

Blasques, F., S. J. Koopman, M. Mallee, and Z. Zhang (2016). Weighted maximum likelihood for dynamic factor analysis and forecasting with mixed frequency data. *Journal of Econometrics 193*(2), 405–417.

Bräuning, F. and S. J. Koopman (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting 30*(3), 572–584.

Chen, D., M. Cane, A. Kaplan, S. Zebiak, and D. Huang (2004). Predictability of El Niño over the past 148 years. *Nature 428(6984)*, 733–736.

Clarke, A. and S. Van Gorder (2003). Improving El Niño prediction using a space-time integration of Indo-Pacific winds and equatorial Pacific winds and equatorial Pacific upper ocean heat content. *Geophysical Research Letters 30*, 1399.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*, 253–265.

Dool, H. v. d. (1994). Searching for analogues, how long must we wait? *Tellus 46A*, 314–324.

Dool, H. v. d. (2007). *Empirical methods in short-term climate prediction.* Oxford University Press.

Doornik, J. A. (2007). *Object-Oriented Matrix Programming Using Ox, 3rd ed.* London: Timberlake Consultants Press.

Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics 164*, 188–205.

Duan, W. and C. Wei (2013). The 'spring predictability barrier' for ENSO predictions and its possible mechanism: Results from a fully coupled model. *International Journal of Climatology 33*, 1280–1292.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89(3)*, 603–615.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods* (2nd ed.). Oxford: Oxford University Press.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology 8*, 985–987.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

Gonzalez, P. and L. Goddard (2016). Long-lead ENSO predictability from CMIP5 decadal hindcasts. *Climate Dynamics 46(9–10)*, 3127–3147.

Good, S. A., M. J. Martin, and N. A. Rayner (2013). En4: quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans 118*, 6704–6716.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association 55*(292), 708–713.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A. C. and N. Shephard (1993). 10 structural time series models. *Handbook of statistics 11*, 261–302.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting 13*(2), 281–291.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica 46*, 1251–1271.

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

Knaff, J. and C. Landsea (1997). An El Niño-Southern Oscillation climatology and persistence (CLIPER) forecasting scheme. *Weather Forecasting 12*, 633–652.

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

Koopman, S. J., N. Shephard, and J. A. Doornik (2008). *SsfPack 3.0: Statistical algorithms for models in state space form*. London: Timberlake Consultants Press.

Kravtsov, S., D. Kondrashov, and M. Ghil (2005). Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability. *Journal of Climate 18*, 4404–4424.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions – five years of experience. *Journal of Business & Economic Statistics 4*, 25–38.

Ludescher, J., A. Gozolchiani, M. Bogachev, A. Bunde, H. Shlomo, and H. Schellnhuber (2013). Improved El Niño forecasting by cooperativity detection. *PNAS 110(29)*, 11749–11745.

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

Petrova, D., J. Ballester, S. J. Koopman, and X. Rodó (2018). Multi-year statistical prediction of ENSO enhanced by the tropical Pacific observing system. *Working paper*.

Petrova, D., S. J. Koopman, J. Ballester, and X. Rodó (2017). Improving the long-lead predictability of El Niño using a novel forecasting scheme based on a dynamic components model. *Climate Dynamics 48(3–4)*, 1249–1276.

Phillips, P. and Z. Xiao (1998). A primer on unit root testing. *Journal of Economic Surveys 12*, 423–470.

Proietti, T. (2000). Comparing seasonal components for structural time series models. *International Journal of Forecasting 16*, 247–260.

Rossi, B. and T. Sekhposyan (2016). Forecast rationality tests in the presence of instabilities, with applications to Federal Reserve and survey forecasts. *Journal of Applied Econometrics 31*(3), 507–532.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*(3), 653–667.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Stock, J. H. and M. W. Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics 30*, 481–493.

Tangang, F., W. Hsieh, and B. Tang (1997). Forecasting the equatorial Pacific sea surface temperatures by neural network models. *Climate Dynamics 13*, 135–147.

Varian, H. R. (1975). *A Bayesian approach to real estate assessment.* North Holland: Amsterdam.

Vogelsang, T. J. (1997). Wald-type tests for detecting breaks in the trend function of a dynamic time series. *Econometric Theory 13*(6), 818–848.

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

Xue, Y., A. Leetmaa, and M. Ji (2000). ENSO prediction with Markov models: The impact of sea level. *Journal of Climate 13*, 849–871.

Appendices

A Technical motivation of the three-step method

Our proposed procedure circumvents the need to forecast X_t completely through the construction of ensemble series. We emphasise that this device renders the simulated prediction errors $v_1^{(i)}, \ldots, v_T^{(i)}$ from the DFM linear functions of all predictors.

The procedure is rather general and can be implemented using different choices of models and estimation methods. For example, a simple autoregressive model can be considered in Step (i) while Step (ii) can be based on a few predictor variables and one principal component (Bräuning and Koopman, 2014). Other methods can be considered, but the minimum requirement is that they allow for efficient simulation from $p(v_t^{uc}|X_{1:T})$. Our forecasting method can also be motivated from a model misspecification argument since it provides the ingredients to carry out a Durbin-Wu-Hausman test (Hausman, 1978) with respect to $\hat{y}_{T+h}$. Given the DFM in Step (ii), one can also test whether the X_t's are endogenous[8]. In the following, we provide some justifications for each step of the forecasting procedure with a small but intuitive Monte Carlo example.

[8]That is to test if entries in the first row of Λ are jointly significant.

A.1 Step (i): reduced-form univariate model

Linear multivariate time series models can often be written in state space form, for example the DFM (4) (Harvey and Shephard, 1993; Durbin and Koopman, 2012). DFM achieves dimensionality reduction while efficiently modelling auto- and cross-correlation when variables are driven by a few system forces. To facilitate our discussion, we assume the following DGP,

$$\begin{aligned} \begin{bmatrix} y_t \\ X_t \end{bmatrix} &= \Lambda f_t + \xi_t, \quad \xi_t \sim N(0, \Sigma_\xi), \\ f_{t+1} &= \Phi f_t + \nu_t, \quad \nu_t \sim N(0, \Sigma_\nu), \end{aligned} \tag{11}$$

where $y_t \in \mathbb{R}$ is the variable to forecast and where $X_t \in \mathbb{R}^{N-1}$. $f_t \in \mathbb{R}^p$ with $p < N$ is the vector of dynamic factors with the autoregressive matrix Φ, and also $f_t \perp \xi_t$. Loading matrix $\Lambda \in \mathbb{R}^{N \times p}$ and covariance matrices $\Sigma_\xi \in \mathbb{R}^{N \times N}$ and $\Sigma_\nu \in \mathbb{R}^{p \times p}$ are time-invariant.

For notational simplicity we assume a diagonal Σ_ξ with finite positive entries σ_i^2, for $i = 1, \ldots, N$. Furthermore, we assume conditions given in Doz et al. (2011) regarding weak stationarity and uniformly bounded auto-covariance matrix of the vector process $(y_t, X_t')'$.

Since y_t is the variable of interest, in practice one often chooses to fit a univariate model for y_t. To see the consequence of choosing a (linear) univariate model, let us partition $\Lambda = (\Lambda_1', \Lambda_2')'$ such that Λ_1 is the 1-st row of Λ and Λ_2 is Λ without its 1-st row. We firstly define a random weighting matrix κ belonging to a restricted Stiefel manifold

$$\kappa \in \{k \in \mathbb{R}^{p \times q} : k'k = I_q, E(kk') = I_p\}, \tag{12}$$

where $q \leq p$. $\kappa\kappa'$ is not of full rank, but its expectation with respect to the probability measure induced by κ is I_p, a p-dimensional identity matrix. With κ, one can choose a univariate model for y_t via

$$\begin{aligned} y_t &= \Lambda_1 \kappa\kappa' f_t + \Lambda_1 (I_p - \kappa\kappa') f_t + \xi_{1,t}, \quad \xi_{1,t} \sim N(0, \sigma_1^2), \\ \kappa'\kappa\kappa' f_{t+1} &= \kappa' T \kappa\kappa' f_t + \kappa' T (I_p - \kappa\kappa') f_t + \kappa' \nu_t. \end{aligned}$$

Denoting $l = \Lambda_1\kappa$, $\Phi^* = \kappa'\Phi\kappa$, $g_t = \kappa' f_t$ and the disturbance terms

$$\xi_t^* = \Lambda_1(I_p - \kappa\kappa')f_t + \xi_{1,t}, \qquad \nu_t^* = \kappa'\Phi(I_p - \kappa\kappa')f_t + \kappa'\nu_t, \tag{13}$$

one has the following univariate model

$$y_t = lg_t + \xi_t^*, \qquad g_{t+1} = T^* g_t + \nu_t^*. \tag{14}$$

This means that the DFM (11) implies a reduced-form univariate model (14) for y_t. For example, any autoregressive integrated moving average and UC models can be cast into (14) with l containing zeros and ones (Durbin and Koopman, 2012). One should notice that in (14) Λ_1 is under-identified because κ is unknown and not unique. Additionally, the measurement equation resembles an endogeneity problem as $g_t \not\perp \xi_t^*$.

To illustrate, consider the following bivariate model with a single factor,

$$\begin{aligned} \begin{bmatrix} y_t \\ x_t \end{bmatrix} &= \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix} f_t + \begin{bmatrix} \xi_{1,t} \\ \xi_{2,t} \end{bmatrix}, \\ f_{t+1} &= \phi f_t + \nu_t, \end{aligned} \tag{15}$$

where $|\phi| < 1$, and where $\xi_{1,t}$, $\xi_{2,t}$ and η_t are i.i.d zero-mean Gaussian with variance σ_1^2, σ_2^2 and σ_η^2, respectively. This gives a reduced-form model for y_t that is an ARMA(1,1) model because

$$y_t = \phi y_{t-1} + \Lambda_1\nu_{t-1} + \xi_{1,t} - \phi\xi_{1,t-1}. \tag{16}$$

So we can write $y_t = \rho y_{t-1} + \theta\epsilon_{t-1} + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. It can be shown $\rho = \phi$ while θ and σ_ϵ^2 can be calculated by equalising the first two autocovariance functions of y_t implied by (16) and by the ARMA(1,1) model.

Step (i) of the proposed method implies a recursive filter of y_t. In the case of Kalman filter, $E(g_t|y_{1:t-1}) = \sum_{j=1}^{t-1} w_{jt}y_j$ with filtering weights w_{jt} being functions of $y_{1:t-1}$. This applies to any linear univariate models which give $\hat{y}_t = E(y_t|y_{1:t-1}) = \sum_{j=1}^{t-1} w_{jt}y_j$. The resulted prediction errors $v_t = y_t - \hat{y}_t$ for $t = 1, .., T$ are orthogonal to $\hat{y}_t$, which implies that based on a chosen univariate model, prediction errors do not play any role in delivering

point forecast and what matters is the estimated signal $\hat{g}_t = E(g_t|y_{1:t-1})$.

A.2 Step (ii): information recovery

From (13), it can be shown v_t may still contain information on dynamics of y_t, because the part related to $\Lambda_1(I_p - \kappa\kappa')f_t$ equals zero only in expectation. For a chosen univariate model this can be retrieved with Step (ii) of our method, i.e. fitting a DFM to $(v_t, X_t')'$.

Our proposed method uses one-step prediction error $v_t = y_t - l\hat{g}_t$ instead of the smoothing error $y_t - \sum_{j=1}^{T} W_{jt}y_j$ with the j-th smoothing weight W_{jt} being a function of all available data $y_{1:T}$ (Durbin and Koopman, 2012). This is crucial because v_t is only a function of $y_{1:t-1}$ without y_t. It is then only a function of $f_{1:t-1}$, leaving information about f_t unexplained. It follows that

$$\begin{aligned} v_t &= \Lambda_1\kappa\kappa' f_t + \Lambda_1(I_p - \kappa\kappa')f_t + \xi_{1,t} - l\sum_{j=1}^{t-1} w_{jt}y_j \\ &= \Lambda_1(I_p - \kappa\kappa')f_t + \hat{\xi}_{1,t}, \end{aligned}$$

where $\hat{\xi}_{1,t} = \xi_{1,t} + l(g_t - \sum_{j=1}^{t-1} w_{jt}y_j)$. One can observe that the first part of the prediction error $\Lambda_1(I_p - \kappa\kappa')f_t$ may lead to poor forecasting performance because part of the information related to the dynamics of f_t is not used in the univariate model. Additionally, the second part $\hat{\xi}_{1,t}$ results from estimating the univariate model (14) for y_t.

As shown, to choose a univariate model is to choose κ from a subspace of the Stiefel manifold $\{k \in \mathbb{R}^{p\times q} : k'k = I_q\}$. The subspace comes from the requirement $E(v_t) = 0$ for any univariate model, which translates into $E(\kappa\kappa') = I_p$, making κ stochastic. In general, one effectively needs $\kappa \in \{k \in \mathbb{R}^{p\times q} : k'k = I_q, E(kk') = I_p\}$. Yet κ is still undetermined as the realisation of κ also depends on $f_{1:T}$, i.e. the data $y_{1:T}$.

In Step (ii), the following DFM is formed

$$\begin{aligned} \begin{bmatrix} v_t \\ X_t \end{bmatrix} &= \begin{bmatrix} \Lambda_1(I_p - \kappa\kappa') \\ \Lambda_2 \end{bmatrix} f_t + \begin{bmatrix} \hat{\xi}_{1,t} \\ \xi_{2,t} \end{bmatrix}, \\ f_{t+1} &= \Phi f_t + \nu_t, \quad \nu_t \sim N(0, \Sigma_\nu). \end{aligned} \tag{17}$$

To illustrate, if we choose a misspecified AR(1) model as the univariate model for y_t in

model (15) and assume no estimation error, Step (ii) gives

$$\begin{aligned}\begin{bmatrix}\Lambda_1\nu_{t-1}+\xi_{1,t}-\phi\xi_{1,t-1}\\ x_t\end{bmatrix} &= \begin{bmatrix}\Lambda_1\\ \Lambda_2\end{bmatrix}\nu_{t-1}+\begin{bmatrix}\xi_{1,t}-\phi\xi_{1,t-1}\\ \Lambda_2\phi f_{t-1}+\xi_{2,t}\end{bmatrix},\\ &= \begin{bmatrix}\Lambda_1\\ \Lambda_2\end{bmatrix}\nu_{t-1}+\hat{\xi}_t.\end{aligned} \tag{18}$$

Using the simulation smoother and assuming no estimation error, we generate simulated prediction error from (17). That is for $i=1,\ldots,M$,

$$v_t^{(i)} = \Lambda_1(I_p-\kappa\kappa')f_t^{(i)},$$

where $f_t^{(i)}$ is simulated from the Gaussian smoothing density $p_{dfm}(f_t|y_{1:T})$ implied by model (17). Importantly, we see that the ensemble time series remove $\hat{\xi}_{1,t}$ and is given by

$$y_t^{(i)} = l\hat{g}_t + v_t^{(i)} = l\sum_{j=1}^{t-1} w_{jt}y_j + \Lambda_1(I_p-\kappa\kappa')f_t^{(i)}. \tag{19}$$

It can be seen from (13) that a chosen univariate model fails to account for the part of information related to $\Lambda_1(I_p-\kappa\kappa')f_t$ in v_t and it gives $l\hat{g}_t \perp v_t$. The second stage DFM and the third stage simulation however breaks the orthogonality to recover the part of information about $Z_1(I_p-\kappa\kappa')f_t$ that is lost in the first stage univariate model. It is easy to see that the non-zero covariance between $l\hat{g}_t$ and $\tilde{v}_t^{(i)}$ comes from the fact that $\tilde{f}_t^{(i)}$ from the simulation smoother (Durbin and Koopman 2012, Shephard and Pitt 1997) is a function of $v_{1:T}$ and $X_{1:T}$ which are thus correlated with $y_{1:t-1}$.

As argued previously, $\kappa \in \{k\in\mathbb{R}^{p\times q} : k'k = I_q, E(kk') = I_p\}$ is still undetermined. The following shows that dynamic factor simulations help extract information on the dynamics of f_t. Firstly, we define a diagonal matrix

$$F_t^{(i)} = diag(\frac{f_{1,t}^{(i)}}{\bar{f}_{1,t}}, \frac{f_{2,t}^{(i)}}{\bar{f}_{2,t}}, \ldots, \frac{f_{p,t}^{(i)}}{\bar{f}_{p,t}}),$$

where $\bar{f}_t = E(f_t|v_{1:T}, X_{1:T})$, i.e. the smoothed mean of the factors implied by the DFM

(11). Immediately we have $E(F_t^{(i)}|v_{1:T}, X_{1:T}) = I_p$. We can thus write

$$v_t^{(i)} = \Lambda_1(I_p - \kappa\kappa')f_t^{(i)} = \Lambda_1(I_p - \kappa\kappa')F_t^{(i)}\bar{f}_t = \Lambda_1(F_t^{(i)} - \dot{\kappa}^{(i)}(\dot{\kappa}^{(i)})')\bar{f}_t,$$

with $\dot{\kappa}^{(i)}$ being any solution to $\dot{\kappa}^{(i)}(\dot{\kappa}^{(i)})' = \kappa\kappa' F_t^{(i)}$. This means that dynamic factor simulation is equivalent to simulating $\dot{\kappa}$, thus the undetermined κ, from an "informative" restricted Stiefel manifold

$$\dot{\kappa} \in \{\dot{\kappa} \in \mathbb{C}^{p\times q} : \dot{\kappa}'\dot{\kappa} = I_q, E(\dot{\kappa}^{(i)}(\dot{\kappa}^{(i)})'|y_{1:T}, X_{1:T}) = I_p\}^{9}. \tag{20}$$

This mechanism makes clear that $v_t^{(i)}$ from Step (ii) recovers information regarding the dynamics of factor f_t via $X_{1:T}$ for $t = 1, \ldots, T$.

A.3 Step (iii): noise reduction

Dynamic factor simulations give M ensemble time series. For $i = 1, \ldots, M$, we have

$$y_t^{(i)} = l\hat{g}_t + v_t^{(i)} = l\hat{g}_t + \Lambda_1(I_p - \dot{\kappa}^{(i)}(\dot{\kappa}^{(i)})')\bar{f}_t, \tag{21}$$

and we compare it with the univariate model in Step (i)

$$y_t = l\hat{g}_t + v_t = l\hat{g}_t + \Lambda_1(I_p - \kappa\kappa')f_t + \xi_{1,t} + \hat{\nu}_t,$$

where $\hat{\nu}_t = l(g_t - \sum_{j=1}^{t-1} w_{jt}y_j)$ results from estimating the univariate model.

A certain κ from space (12) chosen by univariate model is usually suboptimal, while with Kalman filter and smoother our method makes use of information in $(y_{1:T}, X_{1:T})$ optimally in a linear sense. Furthermore, when estimating the ensemble time series (21) in Step (iii), we get rid of the noise from $\xi_{1,t} + \hat{\nu}_t$ and importantly

$$\mathrm{Var}\big(\Lambda_1(I_p - \dot{\kappa}^{(i)}(\dot{\kappa}^{(i)})')\bar{f}_t\big) < \mathrm{Var}\big(\Lambda_1(I_p - \kappa\kappa')f_t\big). \tag{22}$$

To see this, we notice that both $\mathrm{Var}\big(vec(I_p - \dot{\kappa}^{(i)}(\dot{\kappa}^{(i)})')\big) < \mathrm{Var}\big(vec(I_p - \kappa\kappa')\big)$ and

[9] The conditioning set is $\{y_{1:T}, X_{1:T}\}$ which generates the same information set as $\{v_{1:T}, X_{1:T}\}$.

$\text{Var}(\bar{f}_t) < \text{Var}(f_t)$ in the matrix sense, because both left-hand sides operate on a tighter information set. For example, the latter inequality holds due to the law of total variance

$$\begin{aligned}\text{Var}(f_t) - \text{Var}(\bar{f}_t) &= \text{Var}\big(E(f_t|y_{1:T}, X_{1:T})\big) + E\big(\text{Var}(f_t|y_{1:T}, X_{1:T})\big) - \text{Var}(\bar{f}_t) \\ &= E\big(\text{Var}(f_t|y_{1:T}, X_{1:T})\big),\end{aligned}$$

which is a positive definite matrix. Following the formula of variance for product of random variables derived by Goodman (1960), inequality (22) can be shown[10].

Therefore Step (*iii*) achieves noise reduction. A direct consequence is the increased signal-to-noise ratio from a chosen univariate model (14). Equivalently, the underlying dynamics is strengthened by dynamic factor simulations which use $X_{1:T}$ in an optimal linear sense. For example, the i-th ensemble from the illustrative model (15) is

$$y_t^{(i)} = \phi y_{t-1} + \Lambda_1 \tilde{\nu}_{t-1}^{(i)},$$

using dynamic factor simulations based on the DFM (18). So with the help of x_t, we get rid of the MA(1) part $\xi_{1,t} - \phi\xi_{1,t-1}$ in (16). As a result, fitting an AR(1) in Step (*iii*) mitigates misspecification for the univariate model.

The last step of our method averages the forecasts obtained from all ensemble time series. Jensen's inequality shows that averaging over different realisations of $v_t^{(i)}$, or equivalently of $\dot{\kappa}(i)$ from space (20), leads to improved accuracy. It follows that

$$E\big([E_{v_i}(y_t^{(i)}) - y_t]^2\big) \leq E_{v_i}\big(E([y_t^{(i)} - y_t]^2)\big), \tag{23}$$

where $E_{v_i}(.)$ denotes the expectation with respect to the random variables occurring in the dynamic factor simulations. Additionally, we conjecture that averaging also smooths out accumulated estimation errors when estimating the univariate model and the DFM.

[10] One can derive explicit expression for both sides of the inequality if a joint normal distribution for $(vec(\kappa^{(i)})', v_t, X_t')'$ is assumed.

A.4 Small Monte Carlo study

We simulate 100 data replications from (15) with length of 1000 and parameters

$$\Lambda_1 = 1.2, \quad \Lambda_2 = -0.8, \quad \phi = 0.95, \quad \sigma_{\epsilon_1} = 0.16, \quad \sigma_{\epsilon_2} = 0.18, \quad \sigma_\eta = 0.14.$$

To provide Monte Carlo evidence, we conduct a rolling window exercise with window size $T = 480$ so for each replication we can collect 500 h-step ahead forecasts for $h = 1, \ldots, 20$. We base the comparison on the average root mean squared forecast error $ARMSE_h$ for the h-step ahead forecast given by

$$ARMSE_h = \frac{1}{M} \sum_{m=1}^{M} \sqrt{\frac{1}{J} \sum_{j=1}^{J} (y_{T+h}^{m,j} - \hat{y}_{T+h}^{m,j})^2},$$

where $y_{T+h}^{m,j}$ is the realised value and $\hat{y}_{T+h}^{m,j}$ is the predicted value in the j-th moving window of the m-th replication for $j = 1, \ldots, 500$ and $m = 1, \ldots, 100$.

Six models are considered: a misspecified AR(1) model; AR(1)DFS which is AR(1) with dynamic factor simulations; the correctly specified ARMA(1,1) for the reduced-form of y_t; ARMA(1,1)DFS which is ARMA(1,1) with dynamic factor simulations; DFM(estPar) which is the true model (15) with estimated parameters (for identification we set $\Lambda_1 = 1$); DFM(truePar) is the true model with true DGP parameter values. Table 4 shows the $ARMSE_h$ for each model.

The table is visualised in the left and middle panel of Figure 8. Not surprisingly, DFM(truePar) is the best-performing model for almost all h considered, as it is correctly specified without estimation errors. The forecasting accuracy worsens if one estimates the model. From $h = 3$, DFM(estPar) gives $ARMSE_h$ that is 1.02 times bigger than DFM(truePar) and increases to more than 1.2 times when $h = 20$. It is well-known that forecasts delivered by DFMs are destined to be contaminated by estimation errors especially for increasing h and dimensionality. In our simple model, only four parameters need to be estimated but this contamination effect is evident. For $h \geq 16$, $ARMSE_h$ given by DFM(estPar) is even larger than 0.58 which is never surpassed by the misspecified AR(1) model. This supports the claim that more parsimonious univariate models, though

Table 4

Forecasting Evaluation of Different Models

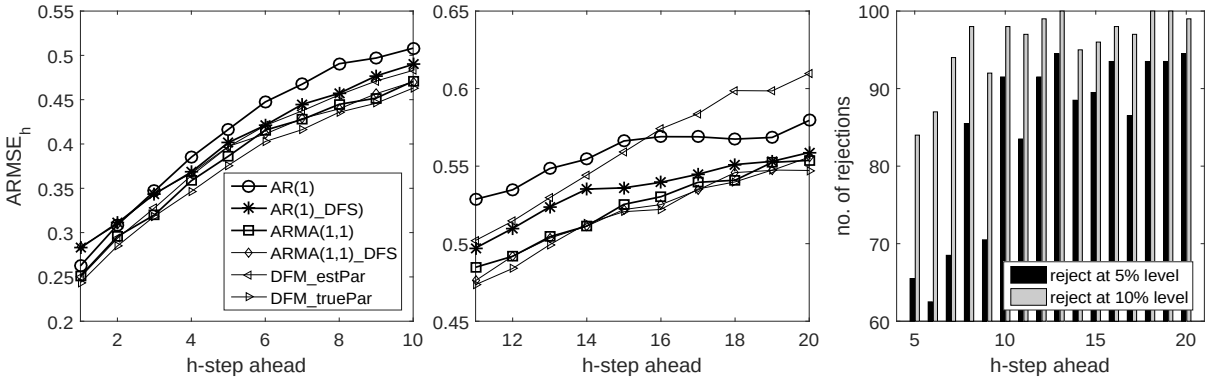
h-step	AR(1)	AR(1)DFS	ARMA(1,1)	ARMA(1,1)DFS	DFM(estPar)	DFM(truePar)
1	0.2598	0.2835	0.2492	0.2832	0.2503	0.2431
2	0.3115	0.3074	0.2907	0.3094	0.2996	0.2845
4	0.3840	0.3702	0.3582	0.3646	0.3660	0.3502
6	0.4499	0.4237	0.4092	0.4124	0.4200	0.4013
8	0.4898	0.4553	0.4408	0.4461	0.4545	0.4391
10	0.5156	0.4893	0.4716	0.4674	0.4868	0.4621
12	0.5343	0.5069	0.4933	0.4906	0.5155	0.4860
14	0.5551	0.5346	0.5136	0.5135	0.5500	0.5085
16	0.5656	0.5460	0.5355	0.5273	0.5709	0.5277
18	0.5746	0.5483	0.5380	0.5424	0.5936	0.5401
20	0.5709	0.5581	0.5585	0.5513	0.6122	0.5493

Reported is the average root mean squared forecast error $ARMSE_h$ for different methods in a rolling window exercise with 100 replications.

misspecified, might help with forecasting (Ashley, 1988).

Figure 8

$ARMSE_h$ for different models



This figure shows h-step ahead forecasting performances among different models based on $ARMSE_h$. Left: $h = 1, ..., 10$; Middle: $h = 11, ..., 20$; Right: Number of rejections from DM test for equal predictive accuracy between AR(1)DFS and AR(1).

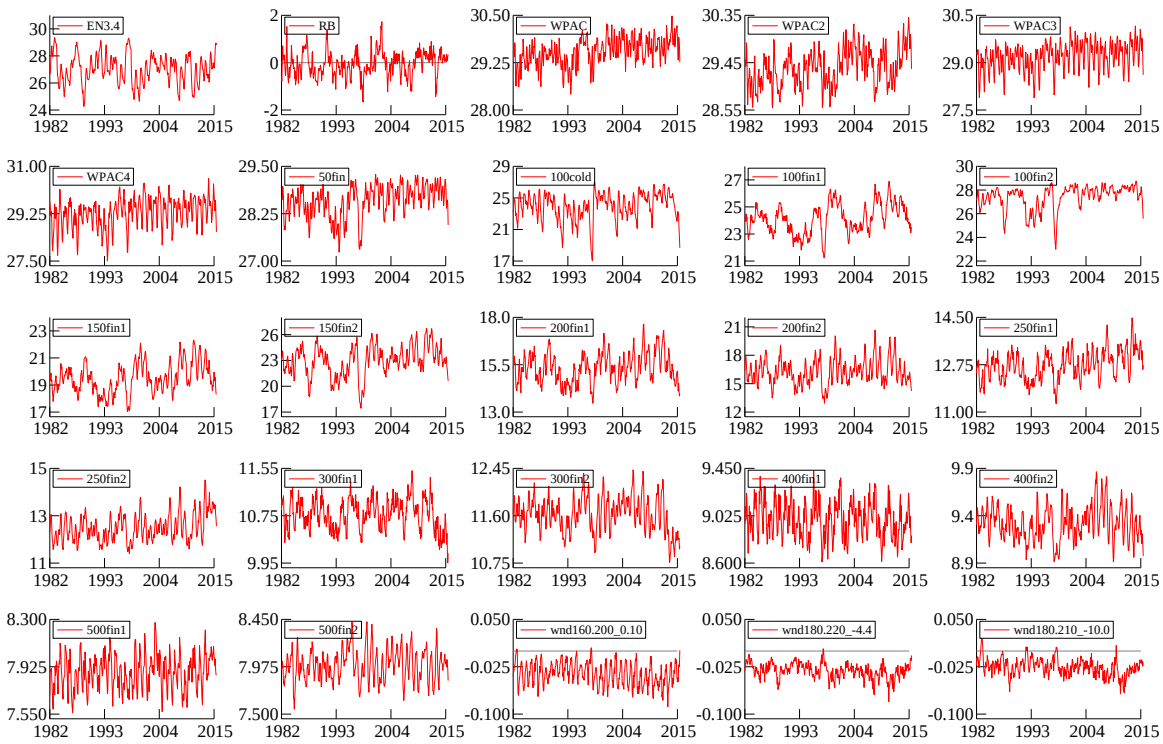
Fitting an AR(1) model to y_t gives poor forecasts because it misses the MA(1) dynamics in the reduced-form. As a result, the $ARMSE_h$ given by AR(1) is approximately 1.08 times larger than ARMA(1,1). But once coupled with dynamic factor simulations, AR(1)DFS delivers $ARMSE_h$ that is clearly smaller than AR(1) for all h except for $h = 1$, 2 and 3, and relative accuracy seems to increase with. This is also shown in the right panel of Figure 8 which gives the number of rejections for Diebold-Mariano (DM) test

(Diebold and Mariano, 1995) based on a squared forecast error loss function. Out of the 100 replications, fewer than 70 replications show significant improvement for $h = 5$ and 6 but it exceeds 85 when $h \geq 8$ at 5% level. Although AR(1)DFS fails to perform better than the correctly specified ARMA(1,1) model it does improve from AR(1) and the resulted $ARMSE_h$ converges to ARMA(1,1) when $h \geq 14$. Lastly, ARMA(1,1)DFS performs almost identically as ARMA(1,1), which shows dynamic factor simulations neither harm nor benefit a correctly specified Step (i) univariate model.

B Data

Figure 9

Data Graphs



The figure presents the time series plots of all variables in our Data Set.

Table 5

Details of Data Set

Description of the dependent and explanatory regression variables used in our study. The acronyms, the variable types, the regions of where the variables are measured and calculated, and the sources from which the variables are obtained. The sources are the NOAA-OI-SST-V2 database (OISST) as provided by NOAA/OAR/ESRL PSD (see `http://www.esrl.noaa.gov/psd/`), the NCEP/NCAR reanalysis by Kalnay et al. (1996), NCEP, Subsurface Temperature And Salinity Analyses (ISHII) by Ishii et al. (2005), as archived at the National Center for Atmospheric Research, Computational and Information Systems Laboratory (see `http://www.rda.ucar.edu/datasets/ds285.3/`), and, finally, the Hadley Centre EN4.0.2 (EN4) as analysed by Good et al. (2013).

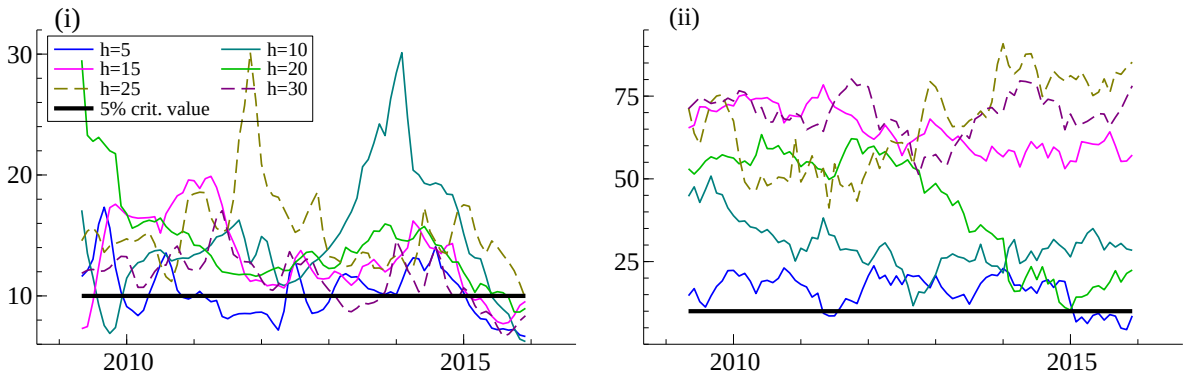
	Acronym	*Region*	*Source*
	Sea surface temperature		
	EN3.4 – Niño3.4	$[190e - 240e] \times [5s - 5n]$	OISST
1	RB	$[180e - 280e] \times [65s - 50S]$	OISST
2	WPAC	$[140e - 160e] \times [5s - 5n]$	OISST
3	WPAC2	$[140e - 180e] \times [10s - 5n]$	OISST
4	WPAC3	$[120e - 170e] \times [10s - 5n]$	OISST
5	WPAC4	$[140e - 160e] \times [10s - 0]$	OISST
	Subsurface temperature		
6	50fin – Temp.50m	$[120e - 170e] \times [10s - 5n]$	ISHII, EN4
7	100cold – Temp.100m "cold"	$[140e - 210e] \times [5n - 10n]$	ISHII, EN4
8	100fin1 – Temp.100m R1	$[120e - 140e] \times [10s - 5n]$	ISHII, EN4
9	100fin2 – Temp.100m R2	$[150e - 180e] \times [7s - 7n]$	ISHII, EN4
10	150fin1 – Temp.150m R1	$[120e - 140e] \times [10s - 5n]$	ISHII, EN4
11	150fin2 – Temp.150m R2	$[150e - 180e] \times [7s - 7n]$	ISHII, EN4
12	200fin1 – Temp.200m R1	$[120e - 140e] \times [10s - 7n]$	ISHII, EN4
13	200fin2 – Temp.200m R2	$[150e - 180e] \times [7s - 7n]$	ISHII, EN4
14	250fin1 – Temp.250m R1	$[120e - 140e] \times [7s - 7n]$	ISHII, EN4
15	250fin2 – Temp.250m R2	$[140e - 170e] \times [7s - 7n]$	ISHII, EN4
16	300fin1 – Temp.300m R1	$[120e - 140e] \times [7s - 7n]$	ISHII, EN4
17	300fin2 – Temp.300m R2	$[160e - 200e] \times [10s - 3n]$	ISHII, EN4
18	400fin1 – Temp.400m R1	$[120e - 140e] \times [5s - 5n]$	ISHII, EN4
19	400fin2 – Temp.400m R2	$[150e - 170e] \times [10s - 0]$	ISHII, EN4
20	500fin1 – Temp.500m R1	$[120e - 140e] \times [5s - 5n]$	ISHII, EN4
21	500fin2 – Temp.500m R2	$[150e - 170e] \times [10s - 0]$	ISHII, EN4
	Zonal wind stress		
22	wnd160.200_0.10 – WND1	$[160e - 200e] \times [0 - 10n]$	NCEP
23	wnd180.220_-4.4 – WND2	$[180e - 220e] \times [4s - 4n]$	NCEP
24	wnd180.210_-10.0 – WND3	$[180e - 210e] \times [10s - 0]$	NCEP

C Fluctuation test

We have adopted the fluctuation test of Rossi and Sekhposyan (2016) to investigate the stability of the superior performance in forecast accuracy of our preferred method UCDFS, against the two best-performing competing methods UCM and LASSO. The Wald test is based on the regression (10) and the null hypothesis whether the forecasts of UCDFS contribute significantly to the forecast error of UCM or LASSO. For each method we have 100 forecast errors. The regression window size is set equal to 30. Hence we have 70 rolling windows with the corresponding Wald statistics W_m, for $m = 1, \ldots, 70$, which are depicted in Figure 10. The fluctuation test statistic is the supremum over the 70 Wald statistics, that is $\max_m(W_m)$. Further details, including those concerning the computation of the critical values, are given by Rossi and Sekhposyan (2016). The test for no information advantage of UCFDS over what is achieved by UCM or LASSO for a forecast horizon h is rejected when the maximum of W_m lies above the critical value (the solid black line). The time paths of the Wald statistic helps gauge what period leads to the main difference (or no difference) in the forecasting performances. We can conclude that UCDFS improves significantly on the forecasts of UCM and LASSO when h becomes larger. Also this superior performance is persistent over time. The choice for another window size (say, 20 or 40) does not change these conclusions.

Figure 10

Fluctuation test for UCDFS against UCM and LASSO



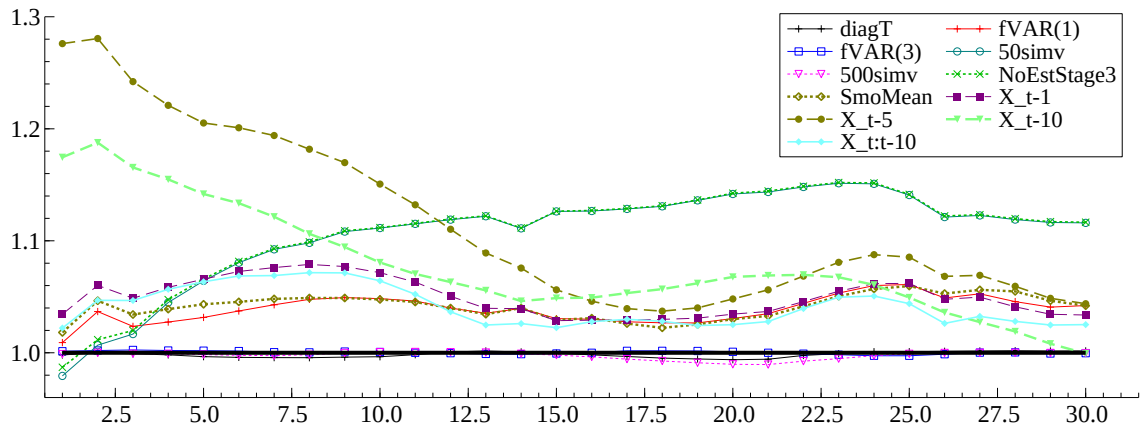
The fluctuation test of Rossi and Sekhposyan (2016). Illustrated are the time path of Wald statistic for the tests that UCFDS has no information advantage (additional forecasting power) on top of what can be achieved by (i) UCM or (ii) LASSO for different h-step ahead forecasts. The instability-robust test statistic is the maximum of all Wald statistics.

D RMSE among different specifications of UCDFS

Figure 11 shows the ratio of RMSE of different UCDFS specifications to that of our base line UCDFS used in Section 4. "diagT" indicates diagonal transition matrices for the VAR(2) dynamics used for the Step (*ii*) DFM. Some extent of parsimony is achieved by specifying diagonal transition matrices for the factor dynamics but the results are literally the same as full matrices are used. "fVAR(1)" and "fVAR(3)" indicate a VAR(1) and VAR(3) model used for the DFM. It can be seen that both RMSEs are higher than that of the baseline specification for all h, but the difference is rather small. "simv" means the number of ensemble residuals generated from the DFM. While "50simv" underperforms the baseline specification with 200 ensembles, increasing this number to 500 suggest no difference. "NoEstStage3" means no estimation is carried out in Step (*iii*); "SmoMean" does not simulate ensembles $v_t^{(i)}$ from $p_{\text{dfm}}(v_t^{uc}|X_1, ..., X_T)$ but use the smoothed mean $E(v_t^{uc}|X_1, ..., X_T)$. Both of the two specifications perform worse than the baseline specification, highlighting the benefit from re-estimation and averaging which aim at reducing estimation errors. "X_t-j" means we do not use the contemporaneous X_t in the DFM, but use the j-th period lagged X_{t-j}. "X_t:t-10" is the collection of $(X_t, X_{t-1}, ..., X_{t-10})$. Except for "X_t:t-10", the forecasts from other specifications are clearly inferior to those obtained from our baseline specification. Therefore, one should use the contemporaneous X_t to in Step (*ii*). The fact that one fails to reject the null of equal predicative accuracy between "X_t:t-10" and the baseline specification sheds light on that naive use of older observation such as the S&W and CDFM may not always be a good choice for improving El Niño forecasts.

Figure 11

Relative RMSE of alternative model specifications of UCDFS



The baseline specification has VAR(2) process with full transition matrices for the factors in the Step (ii) DFM $(v_t^{uc}, X_t')'$, and 200 ensemble residuals $v_t^{(i)}$. All third step ensemble series $y_t^{(i)}$ for $i = 1, ..., 200$ are re-estimated.