Enhanced El Nino predictability from climate mode interactions

Tamás Bódai

bodai@pusan.ac.kr

Center for Climate Physics, Institute of Basic Science https://orcid.org/0000-0002-3049-107X

Article

Keywords:

Posted Date: November 8th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5336072/v1

Additional Declarations: There is **NO** Competing Interest.

Enhanced El Niño predictability from climate mode interactions

Tamás Bódai

October 26, 2024

Department of Applied Statistics, Institute for Mathematics and Physical Sciences,
Hungarian University of Agriculture and Life Sciences, Budapest, Hungary

Abstract

With further extension of the XROM introduced by Zhao et al., by including the ENSO-state-dependence of the external noise forcing as well as a seasonal modulation of both the additive and state dependent part of the forcing, I am able to improve the ENSO prediction skill considerably, especially felt for longer lead times. My new data-driven forecast model is inferred by Maximum Likelihood Estimation, from observational data in 1979-2022, which is much more costly than solving a linear regression problem by matrix inversion. This is the – completely affordable – price of obtaining the currently best forecast model of large scale features of ENSO, falsifying the claim of Zhao et al. that a reliable estimation of the state dependence is too data-intensive. I also make a few points of scrutiny by introducing a package of four concepts, those of: the apparent, theoretical maximum, climatological and true prediction skills. Finally, from the viewpoint of the philosophy of science, I examine whether Zhao et al. have delivered on their promise of explaining the enhancement of the prediction skill of the XROM in comparison with the celebrated recharge oscillator model (ROM) of ENSO.

1 Introduction

The authors of the considered article [1] (the "Article" in what follows) have made a great scientific leap by demonstrating, if not explaining, that the best model to predict ENSO, an *unphysical* AI-based model [2], owes its extra skill – elevating it above other preexisting (physical) models – from (excuse the irony: unwittingly) "taking account" of ENSO's interaction with other major modes of climate variability. The prevailing view is that ENSO is the *master* of global climate variability, although, the influences of other fairly well defined climate subsystems on ENSO also have a growing literature. Yet, a number of these articles have been "debunked", in e.g. [3, 4], showing that the influence that they found is just "apparent" – a "mirage" – and their conclusions are based on a lack of depth of the science. In this context, the article of Zhao et al. is very welcome restoring confidence in the idea that ENSO *can* be considerably influenced causally. It also gives great pleasure to the scientist to see that the new – supposedly – physical model dubbed the "XRO modell" (or XROM) rivalling the AI-based one is so simple in its basic construction, even if it has more than 500 parameters.

Yet, the Article, offered to a broad nonspecialist readership, is asking for some clarification and scrutiny. Although, I believe that many seasoned forecasters will also find some of my reasoning and analyses novel. To begin with, let us recast the XROM in a mathematical system of notation that hopefully makes the meaning clear more immediately. Subsequently, I will propose a modification and extension of it which is useful, on the one hand, and makes the model physically rather plausible, on the other. The original model consists of the coupled system of stochastic differential equations (SDE) written in the Langevin form:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = c^{ij}x_j + f_i(x_j, t; b^{il}) + \sigma_i^\xi \xi_i(t), \text{ and} \tag{1a}$$

$$\frac{\mathrm{d}\xi_i}{\mathrm{d}t} = a_i\xi_i + \eta_i(t),\ i,j = 1,\dots,J,\ l = 1,\dots,L_i, \tag{1b}$$

$J = 10$, where e.g. the tensor notation $c^{ij}x_j = \sum_{j=1}^{J} c^{ij}x_j$ encodes a linear combination in the sense of the "Einstein summation convention", $f_i(x_j, t; b^{il})$ are some (nonlinear) functions of some of the (possibly multiple j) state variables as well as time t, specified by parameters b^{il}, and the η_i are – not necessarily independent – realisations of white noise processes of zero mean and variances $2|a_i|$, i.e., $\langle \eta_i(t)\eta_i(t-s)\rangle_t = 2|a_i| > 0$ for $s = 0$ and 0 otherwise, all obeying normal distributions. We would expect an interdependence especially between η_1 and η_2 belonging to the Niño3.4 (climate) index $x_1 = T$ [°C] and the mean Equatorial Pacific thermocline depth $x_2 = h$ [m], the two variables governed by the ROM [5] that the XROM generalises, i.e., $\langle \eta_1\eta_2\rangle_t \neq 0$, because they pertain to the same location laterally. We have nonlinearities only in the equations for T and the Dipole Mode Index (DMI) $x_6 = I$ [°C] of the Indian Ocean Dipole (IOD) oscillatory mode:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = c^{1j}x_j + b_{TT}T^2 + b_{Th}Th + \sigma_T^{\xi}\xi_T, \tag{2}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = c^{6j}x_j + b_{II}I^2 + \sigma_I^{\xi}\xi_I. \tag{3}$$

Although, I estimate b_{II} from 1979-2022 observational (reanalysis) data [6] to be statistically insignificant. Furthermore, the estimates of many of c^{ij} using the same 44 years worth of observational data are also insignificant, to be addressed in Sec. 2.3. For some points of analysis, it is worth to treat the "constant parameter" (except, perhaps, for $c^{11} = c^{TT}$) model, but, in fact, in the "fully seasonal" XROM proposed by Zhao et al., the parameters b^{il}, c^{ij} are seasonally (periodically) modulated retaining two ($k = 1, 2$) harmonics. That is,

$$c^{ij} = c_0^{ij} + c_k^{ij}\cos(k\Omega t - \Phi_c^{ijk}), \tag{4}$$

while a formally identical expression applies to b^{il}. – But not to a_i and σ_i^{η}, which latter they did not justify. (In (4), the Einstein summation rule tdw. index k applies, just like $f_i(x_j, t; b^{il})$ in (1a), tdw. index i, is clearly meant to be a sum, or in the case of lag variables shortly to be introduced.) Such a temporal but periodic modulation makes the deterministic part of the system nonautonomous but cyclostationary and the pullback/snapshot ([7]/[8] and references therein) probability density distribution function (PDF) $p(x_i, t)$ time dependent but periodic.

Apart from nonlinearity, Zhao et al. claim – citing references – that the state dependence of the noise forcing of the Equatorial Pacific SST, e.g.

$$\xi_T(t) \to (1 + \beta T)\xi_T(t)$$

(with a seasonal modulation of β, potentially), could be important but they neglect it because it is data-intensive to reliably estimate. On the other hand, Olson et al. [9] for the first time, and subsequently Bódai et al. [10] independently, concluded that it plays negligible role in determining ENSO's skewness. Let us point out that these two claims are not necessarily contradictory, because Zhao et al. is presumably referring to predictability, not skewness. Also note that because of the above form of state dependence, it is also commonly referred to as "correlated additive multiplicative" (CAM) noise.

In any case, I do **propose**, first, to include seasonally modulated state dependence, in the T-equation (2), at least, such that

$$\beta = \beta_0 + \beta_k\cos(k\Omega t - \Phi_\beta^k). \tag{5}$$

Furthermore, note that the cross-correlations of η_i in the XROM entail those of the corresponding ξ_i. However, I think that it is physically implausible that ξ_i are correlated not because eqs. (1b) are coupled (they are not) but because some unmodelled processes represented by η_i are interactive. In this regard, consider that the state variables x_i are spatially very large scale quantifiers. Furthermore, no physical quantity other than temperature is considered (except for h) and only in selected locations. This means that the system of partial differential equations of the climate fluid system is drastically *coarsegrained* and "decimated by – subjective but rational – selection". It is similar to taking a vector autoregressive model (VAR) of two state variables and trying to derive an autoregressive (AR) model of just one variable being either the mean of the original two variables or just one of them, respectively. Memory terms are well known to emerge upon model reduction; see e.g. [11] and references therein. Therefore, second, instead of having equations like (1b), I retain only equations (1a) and, upon temporal discretisation (Methods 4.2; $x_n = x(t_n)$, $t_n = n\Delta t$, $\Delta t = 1$ month, $n = 0, 1, 2, \ldots$), include memory – or "delay" or "lag" – variables,

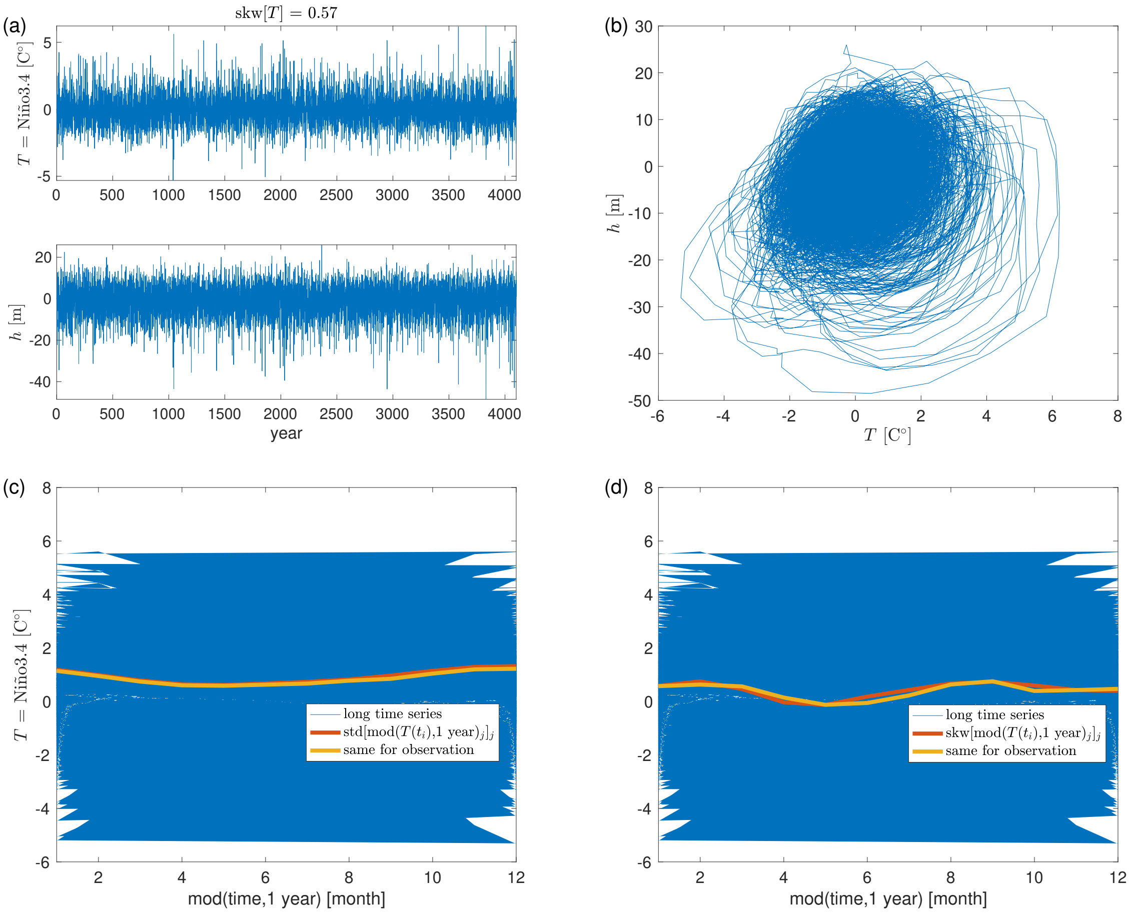

Figure 1: A very long simulation of the XDROM+ fitted to the ORA (1979-2022) observational data by MLE as described in Methods. 4.2. Clearly, the process is stationary in view of long time series (a); and the synthetic true skewness of Nino3.4 is very close to the observed skewness of 0.52, while the seasonal dependences of the standard deviation (c) and skewness (d) are both very satisfactory. Panel (b) displays the trajectory projected onto the 2D phase plane of the ROM.

x_{n-d}, $d = 0, \ldots, D$. But only to linear/first order for $d \geq 1$. That is, e.g. in eq. (2), or rather in its time-discrete version, we have $c_n^{1j} x_{nj} \to c_n^{1dj} x_{(n-d)j} = c_n^{10j} x_{nj} + c_n^{11j} x_{(n-1)j} + \cdots + c_n^{1Dj} x_{(n-D)j}$. I will refer to a model with delay variables as an XDROM. For reliable model/parameter inference (Methods 4.2), we need to include so many D delay variables that the auto-covariances of (raw) fitting residuals $\langle \hat{\xi}_n \hat{\xi}_{n-m} \rangle_n$ are statistically insignificant for any m. I find that the inclusion of as few as a single delay variable (per x_i), $D = 1$, satisfies this requirement in the case of the 10D XDROM. This is fortunate, indeed, because further delay variables would immensely inflate the model parameter space. Finally, third, similarly to β, I allow for a seasonal modulation of the CAM noise strength σ_i^{ξ}. I will refer to a model that has $\beta \neq 0$ and includes the seasonal modulation of it and that of the σ_i^{ξ} as an XDROM+.

The XDROM+ inferred from observational data in the period of 1979-2022 [6] seems in view of Fig. 1 very authentic, just like the XROM. However, in quantitative terms, a main novel result in this article is that the XDROM+ well outperforms the XROM wrt. prediction skill – when enough data is fed to it – increasingly more so for longer forecast lead times (Fig. 5). This makes the XDROM+ the best model currently for forecasting large scale features of ENSO.

On a note of scrutiny, otherwise, I argue that what is calculated by Zhao et al. as for quantifying the prediction skill, Pearson's (linear) correlation coefficient, something that I call here the "observed" or "apparent prediction skill", is useful only in comparing different models or getting a ballpark figure of what I call the "climatological skill". – Although a rather uncertain ballpark figure because of the (power) law of large numbers (of a very slow decay). I also demonstrate that the apparent skill calculated from historical data is completely unrelated to what I regard the true skill of a prediction made for the immediate future. The latter cannot be defined by a correlation coefficient but rather an ensemble-wise root mean square error (E-RMSE or REMSE) of the *deterministic* prediction, where the ensemble is a *probabilistic* forecast ensemble with all members stemming identically from the current observed state.

All the points of technical scrutiny pertain to a mathematical and statistical understanding/interpretation of the results of Zhao et al. These analyses are aided by *ensemble experiments* that simulate a *synthetic truth*, taken to be the XDROM in some cases and the XDROM+ in others fitted to observational data provided by Sen Zhao. Without such a context, most readers will misunderstand the results of the Article, I believe, which opinion is the result of a "casual extrapolation" from my perception of the referee reports and editorial handling as *incompetent*.

Finally, I also report on my attempts to further improve the prediction skill (or just "skill" in the following) by reducing the complexity of the model. To my surprise, I could make only meagre improvement this way.

2 Analysis

2.1 Climatological, theoretical maximum, apparent and true prediction skill

It is rather informative to define and determine a "theoretical maximum skill" – the second (ii) item of the section title. The "number one" basis of the concept is that 1) we have the *exact* model that governs the process to predict. Therefore, apart from defining it, we can only determine the theoretical maximum for a "synthetic truth" – ergo, unfortunately, not the actual truth. Hence, this theoretical maximum is *not a strict reference* for the observed skill. But it might well be – and I assume so – a very accurate one. Another assumption is that 2) no other model can predict the truth of a specific XDROM better than this very XDROM can predict itself. Perhaps this assumption is obviously true (but I have a lingering doubt). Third, 3) a further matter that makes the calculated theoretical maximum not an accurate reference for the observed skill is that we do not know the parameters of the XDROM even if it governs ENSO (and the other $x_i, i > 2$) deadly accurately in terms of 1). The best that we can do is that we fit the available observational data by the best model that we can think of (the XDROM or the XDROM+, depending on the purpose) and use the estimated parameter values to specify the synthetic truth. Still, we can have an idea of the uncertainty of the reference in this regard by simulating the model generating many independent realisations of $2022 - 1979 + 1 = 44$ years time span, fitting the synthetic data to have a synthetic truth for each realisation, determining the theoretical maxima for all synthetic truths and, finally, getting the variance of the theoretical maximum skill with respect to the different realisations/"alternative truths"/spaghetto (singular, à la Michael Ghil). This is rather computationally expensive, however, and thus I choose to rely on my faith/confidence instead. The calculation of the theoretical maximum skill takes a Monte Carlo (MC) experiment. For the MC, I make $R_s = 1000$ runs, which are initialised from independent December-January (DJ, $t = t_0$) states (remember: $D = 1$), two years apart each DJ initial condition (IC), of the synthetic truth of an XDROM obtained from a long simulation of it. From each IC, I run the XDROM over 44 years (until t_f). At an earlier point, say, $t_t = 25$ years, where the subscript t in t_t stands for *training*, the simulations are branched out, "forked": as for an "offshoot", applying no random innovations for pair-wise corresponding (deterministic) forecasts. This offshoot is of $\tau_{max} = \max[\tau] = 20$ months span only, the maximal forecast lead time that is of interest here. Finally, for each of the lead times $\tau = 1, \ldots, 20$ separately, I calculate the maximum skill as a correlation coefficient $\rho_{max}(\tau) = \rho[T_{r,p}(t_t + \tau), T_{r,t}(t_t + \tau)]_r$ wrt. the variability over the MC runs/realisations. (Subscripts $r,\ p,\ t$ of $T =$ Niño3.4 stand for 'realisation', 'prediction', 'truth', respectively.)

Because of the independence of the realisations, they well sample the snapshot PDF that provides a *sound* definition of the climate [8, 12, 13]. In other words, the MC *ensemble* well represents the climatological distribution. Hence, I refer to such a skill as a "*climatological* skill", the first item (i) in the section title.

Because of the periodicity/cyclostationarity of the climatology [14, 9], one would expect the predictability to also depend on the season, i.e., that $\rho_{max}(\tau)$ are periodic functions of a variable t_t. Zhao et al. did demonstrate the seasonality of skill in their Fig. 2 (p); here I consider the theoretical maximum in addition and use a different *ad hoc* visual representation. To this end, I actually conduct a set of 12

MC experiments with forking at all different months of the year, $t_t = 300 + n$ [month], $n = 1, \dots, 12$. The seasonality of predictability is nicely confirmed by Fig. 2 (a). The different curves showing the lead time dependence start their nosedive at different times/months because of their relative "distance" from spring which poses a barrier to predictability [15, 16]. Apart from this short range of fast change, the predictability can persist quite well for a number of months ahead. Taking an (weighted; see φ below) annual average of the skill for each lead time τ separately, marked by the black dot-dash line, the decline of this average skill with increasing τ is very steady.

When trying to determine the theoretical maximum skill, the approximation can suffer for two reasons. One is, in reference to assumptions 1) and 3) above, that the wrong model is used. Say, if we use forecast models inferred from data spans of $t_t = 25$ years in each of the said MC experiment ($\rho_{T_t,max}(\tau) = \rho[T_{r,T_t,p}(t_t + \tau), T_{r,t}(t_t + \tau)]_r$, $t_t = \max[T_t]$), we obtain the picture seen in Fig. 2 (b). The approximation suffers in the word's practical sense: the skill would always be seen *worse* ($\rho_{T_t,max}(\tau) < \rho_{max}(\tau)$).

The other reason why the approximation would suffer is the imperfect representation of the climatology. This is entailed by the very real situation of a finite observational time span. In fact, in a data-driven forecasting approach [17, 18], the latter entails both causes, also the first one, as the observational data is used not only for evaluating the skill but also for inferring the forecast model. However, one important difference between the two causes is that – by chance – finite data can make the skill look better than what it actually is. Assumption 2) above, that the true model is the best predictor, implies that a better than real skill is an artefact. It is only fair, then, to call the skill determined from finite observational data the "observed skill", $\rho_{T_o}(\tau)$ (notation regarding the subscript inspired by [19]), where $T_o = \{t \in [t_0, t_f]\}$ refers to the time period of observation, a.k.a. "apparent skill" in the parlance of [4], the third item (iii) of the section title. This is – and can only be – based on *temporal* (versus ensemble-wise) statistics: $\rho_{T_o}(\tau) = \rho[T_p(t + \tau), T_t(t + \tau)]_{t \in T_v}$. In the latter, T_v denotes a "model verification" time period, say, the leftover upon booking a training time period T_t from the observational time period T_o. In fact, $T_v = T_o \setminus T_t \setminus T_p$, the backslash denoting set difference and T_p the time period of the latest forecast – or, actually, "hindcast" – time span. Obviously, $|T_p| = \tau_{max}$. And T_t is commonly a compact, single piece set. Because the sets T_v and T_t are disjunct, the said definition of ρ_{T_o} is often called the (conservative) "out-of-sample" skill (in contrast with the – in the case of overfitting that is not "benign" [20], stupidly – liberal "in-sample" skill). I will refer to this definition, used in Fig. 2 (d), as method #1. However, if the model fitting method allows for a "gap" in T_t, then method #1 is wasteful. Instead, we can have $T_v = T_p$ and place it every possible way within T_o. I will refer to this definition of ρ_{T_o}, used in Fig. 2 (c), as method #2. This way the training data set is considerably larger, $|T_t| = |T_o| - |T_p|$, besides that we have a larger sample of pairs of the (synthetic) truth and (deterministic) prediction for evaluating the (apparent) skill as a sample correlation coefficient. If we can allow for more than one gap in T_t, then we can push things to the limit of $|T_t| = |T_o| - 1$. Although it is probably not that useful given that $|T_p| = \tau_{max} \ll T_o$.

Because $\rho_{T_o}(\tau)$ does not differentiate between the months when forecasts are made, it is representing some sort of an annual average, which is – considering the dot-dash line in Fig. 2 (b) – more likely to have a decay of possibly a rather even rate, just like it is shown by the ensemble/spaghetti diagram in Fig. 2 (d) sampling possible realisations of $\rho_{T_o}(\tau)$ for the XDROM process. (The observed skills have been evaluated only for 200 of the $R_s = 1000$ realisations.) For this reason, the said apparent skill can be well above the *relevant* theoretical maximum *considering* the month when the forecast is made. Alas, even the "envelop" $\max[\rho_{T_t,max}]_{t_t}(\tau)$ (black dashed curve) is "breached" for a rather large proportion of realisations/lead times.

It clearly pays to use method #2 in evaluating the apparent skill. In Fig. 2 (c), the E-mean (black solid) approximates the annual mean theoretical maximum quite well already for $|T_o| = 44$ years. Although, in fact, it is not a straightforward E-mean but $\varphi^{-1}\langle\varphi(\rho_{T_o})\rangle_r$, where $\varphi(\rho) = \operatorname{atanh}(\rho)$ is the variance-stabilising Fisher transformation. However, we see that there is still considerable E-wise variability of the apparent skill. Fig. 2 (e) displays the results of an analysis of the dependence of the approximation of the true value of the climatological skill and the estimation variance (or standard deviation) on the time series length. Clearly, in the limit of infinite length, the theoretical value is approached (the estimate is "asymptotically unbiased"). The results show that, first, the respective tendencies are *slow*, governed by scale free power laws (as opposed to fast exponential decays of well defined time scales), and, second, the approach of the mean is much faster than the vanishing of variability. The latter is characterised by an exponent of about $-1/2$, although definitely larger in modulus than that, as indicated by the standard error in round brackets following the last stable significant digit of the nonlinear regression estimate. This result could be expected because the standard deviation of the sample correlation coefficient upon

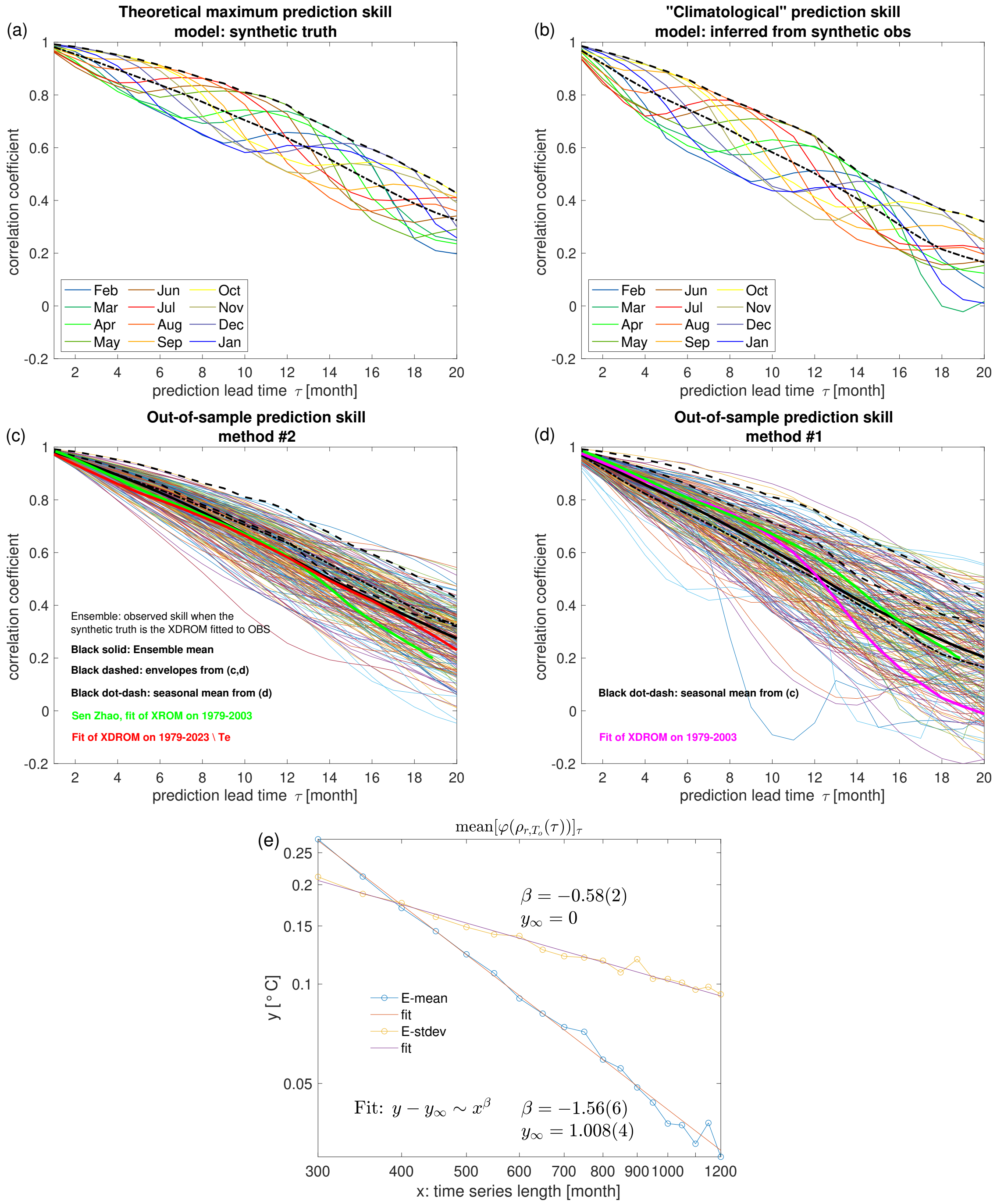

Figure 2: Theoretical maximum and apparent prediction skill of the fully seasonal XDROM. (a,b) Seasonality of the theoretical maximum prediction skill defined in two different senses. (c,d) Compare the result using the fitting method [5] of Zhao et al. [1] (green) with three things: my attempts of reproducing it (red, magenta), the theoretical maximum defined in two different ways (dashed envelopes; see the main text) and contingencies (the ensemble of thin lines belonging to different independent realisations) subject to the assumption that the XDROM fitted to 1979-2022 observations by LSQ (Methods 4.2) faithfully represents ENSO and its teleconnection network. Seasonality of the prediction skill is also shown by Fig. 2 (p) of the Article, but they use temporal correlation coefficients not E-wise and that figure does not function to show the envelop. (e) Scaling laws for the apparent skill approximating the E-wise climatological skill. For this diagram, the apparent skill was evaluated for 500 realisations (for each choice of the time series length – horizontal axis), whereas doing this only for a 100 realisations leaves the scaling laws barely recognisable.

applying the Fisher transformation only depends on the sample size N (not e.g. the true value of the correlation coefficient) being $(N+3)^{-1/2}$, something like the law of large numbers, however, in our case we expect a favourable effect from the fact that a longer time series would yield on average a more accurate forecast model. The power exponent of the much faster approach of the mean or true value (y_∞), on the other hand, is about $-3/2$.

In panels (c,d), I display my attempts (red, magenta, respectively) of reproducing what the method of Zhao et al. gives (new XROM simulation output provided kindly be Sen Zhao, as I am not able to use their Python code) relying on data between years 1979-2022 (as available in their Zenodo archive [6]), $T_t = 25$ years (green). It turns out that their method fares better with method #1, and I can only get similar results when relying on more training/evaluation data using method #2. I must think that this is because my model, the XDROM, is more data-hungry given that it has about twice as many parameters owing to featuring delay variables. (But this is just naive speculation.) On the other hand, their fitting method, following [5], would not allow for a gap in T_t and, so, they can only employ method #1 to calculate the apparent skill. (Neither does Matlab's `nlarx` allow for a gap, besides that there seems to be a bug in that code, which, as I understand, Mathworks is now working on to eliminate.)

Finally, we consider the concept of the "true skill". I think, it is the most meaningful to define this skill in the practical context, one when we intend to make a (data-driven) forecast for the immediate future relying on as much historical data as we trust to make us a favour instead of working against us. That is, we infer the forecast model from data in T_o (of our choice) and make a forecast for the next few months, $\tau = 1, 2, \ldots$ The statistical quantity that could define **the true skill cannot be a correlation coefficient**, because the forecast is deterministic and hence it has no variance, in which case the correlation coefficient is undefined. It can be, however, an ensemble-wise root-mean-square-error, E-RMSE or REMSE, $\mathrm{REMSE}_{T_o}(\tau) = \mathrm{RMSE}[T_{r,T_o,p}(t_f + \tau), T_{r,t}(t_f + \tau)]_r$, $t_f = \max[T_o]$, in which case the ensemble (of realisations, r) is a forecast ensemble defined by all possible realisations of the noise forcing. That is, the REMSE can be evaluated by performing an appropriate new MC experiment.

2.2 Is the apparent prediction skill useful?

Let us, then, consider the dependence of the true skill on t_f, $\mathrm{REMSE}_{T_o}(t_f, \tau)$, as new data is becoming available with the passing of time. We can think of this dependence in terms of an expanding window with a fixed t_0 or a moving window. The choice would not make a material difference had the $\mathrm{REMSE}_{T_o}(t_f, \tau)$ had a short persistence. This is indeed the case, as shown by Fig. 3. Intending to answer the question of the section heading, consider that the apparent skill $\rho_{T_o}(t_f, \tau)$ as a temporal statistic surely has a considerable persistence in stark contrast with the true skill. Therefore, in ways of scrutinising the Article, I conclude that **the apparent skill is not indicative of the true skill**. Any reader and forecaster should be clear on this point.

The apparent skill is useful, however, in comparing the predictive power of different models. This claim can be backed by the fact that the apparent skill calculated by methods #1 and #2 correlate. Relying on the data that produced Fig. 2 (c,d), I evaluated the correlation coefficient for each τ separately and found extremely highly significant figures stably around $1/2$. Typically, method #2 yields a larger apparent skill, clearly, because it is based on more data and, hence, the inferred forecast model tends to be more authentic. This fact is instrumental in proving in the next section that the XDROM+ model is superior to the XROM of Zhao et al.

But before that, let us supply evidence for the claim that the apparent skill is not really useful for other things than the comparison of models – if not for providing a ballpark figure of the prediction skill. Fig. 4 (a) shows that $\rho_{T_o}(t_f, \tau)$ (method #2) and $\mathrm{REMSE}_{T_o}(t_f, \tau)$ do not correlate at all for any τ. The correlation is taken ensemble-wise, having generated "possible synthetic observations" (alternative realities). Then, one might think that the apparent skill has a persistence that is not wholly an artefact to do with the sliding window size, say, that of T_v. But, say, the apparent skill could be related in adjacent, non-overlapping windows of T_v. To examine this proposition, we use a third method #3 of evaluating the apparent skill. This is done by fixing $t_0 = \min[T_t] = 1979$ and, as T_v, $|T_v| = 19$, is moved forward in time, taking such T_t's that $\max[T_t] = \min[T_v]$. Then, let us compare the apparent skills evaluated for a pair of T_v's such that $\max[T_v] = 2022$ for one and $\min[T_v] = 2022$ for the other, in terms, again, of an E-wise correlation coefficient. The result of this is shown in Fig. 4 (b). That is, any persistence of the apparent skill is wholly an artefact, indeed.

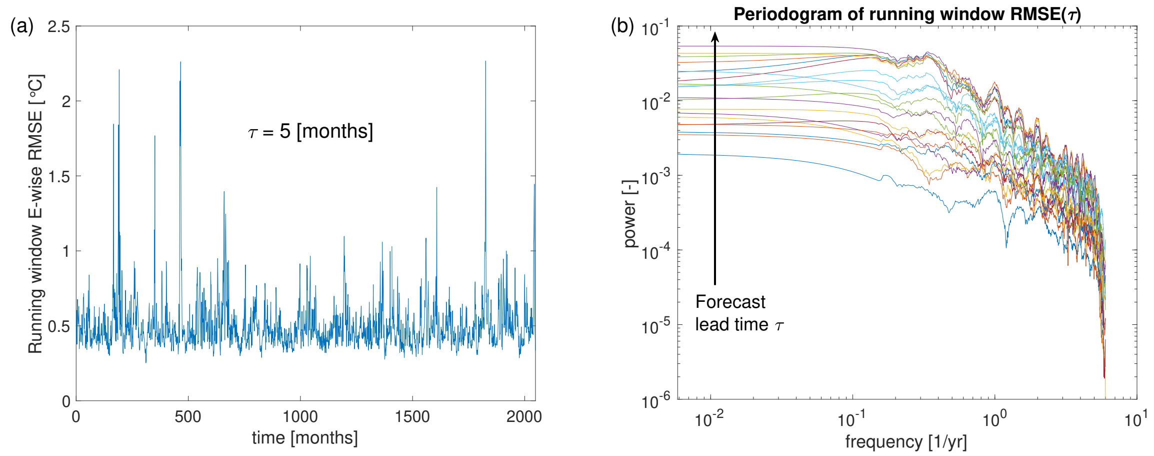

Figure 3: The true prediction skill in association with a period of observation T_o: the ensemble-wise RMSE (or REMSE) REMSE$_{T_o}(t_f, \tau)$. (a) A time series of the REMSE by running the window T_o and (b) corresponding power spectra. The spectra are obtained by smoothing the raw estimate using a Savitzky-Golay filter with a 50 data point window, calling Matlab's `smoothdata`, and the raw spectra are calculated by Matlab's `periodogram`. For obtaining time series like that in panel (a), I simulate 100 possible futures τ_{max} ahead for each point in time (month) along a long reference trajectory. At times, some of the 100 E-members blow up, presumably because of the existence of an attractor at infinity; I discard those escaping trajectories when calculating the REMSE. When T exits the range of $[-5, 10]$ [°C], the realisation is omitted.

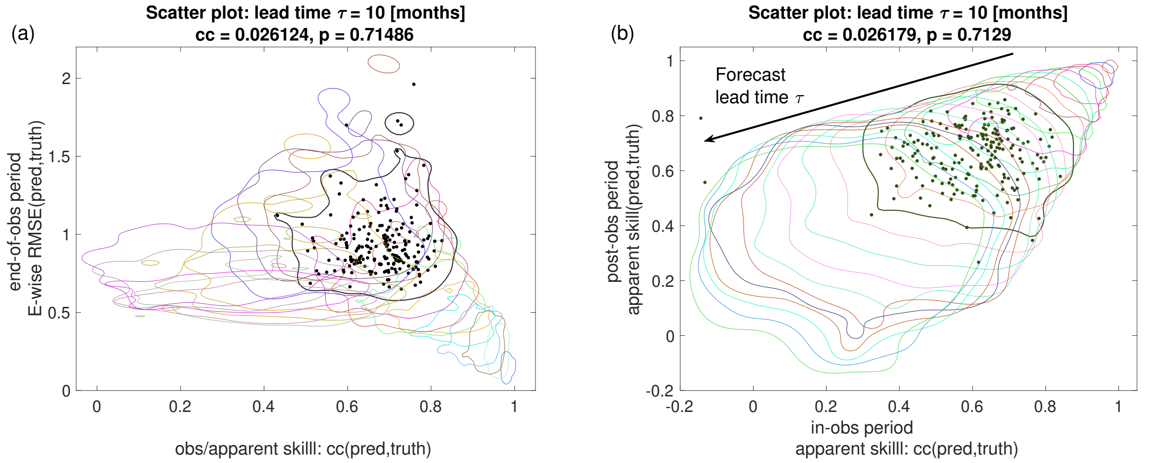

Figure 4: The lack of relationship of the true skill with the apparent skill (a) and the absence of a physical persistence of the apparent skill (b). Scatter plots are shown for $\tau = 10$ months, and for all other lead times considered I just plot the level contour of the PDF inside which 95% of the probability mass is contained as obtained from a (not very accurate) kernel density estimate of the PDF. These contours are good enough to show that data points scatter in an area extended in all directions, i.e., not along one specific direction.

2.3 Improvement

The XDROM+ seems to have a **superior predictive power** over the X(D)ROM by a **considerable margin** in view of the observed skill; see Fig. 5 (d). To dispel any doubt that the XDROM+ is really a superior model in predicting ENSO, I conducted a "small-ensemble" experiment. It is neither necessary to go for more than five ensemble members to have statistical confidence, nor is it tempting because of the hefty run time of calculating the observed skill using method #2, as the XDROM+ has to be fitted to data by the expensive MLE a number of $(T_o - \tau_{max})$ [month] times for each E-members. For the synthetic truth, to be predicted by both the XDROM and XDROM+, I used the XDROM+ fitted to the actual observations 1979-2022. Two diagrams in Fig. 5 (a,b) show the results using method #1 and #2, respectively. In addition, the diagram in panel (c) shows these results collectively in a different representation (see the figure caption). We can see how the XDROM fares better using method #1. However, using method #2, the XDROM+ proves considerably more powerful (all circle markers situated below the solid diagonal line of breaking even). The improvement is felt increasingly more at longer lead times τ. Panel (d) shows that the skill can be improved further a tiny bit by including $K = 3$ instead of 2 harmonics (purple vs gold) for the seasonal modulation of parameters.

For the purpose of further improving the skill, two simple ways of model complexity reduction are described in Methods 4.3. One is by rewriting the 10D X(D)ROM(+) into a 2D or even 1D DROM(+) retaining T and h or just T, respectively. The 1D DROM+ has a weaker prediction skill (result not shown) and the 2D DROM+ also has a much worse prediction skill than the X(D)ROM for lead times up to 15 months. This is so even upon an attempt of optimising for the memory (D) and number (K) of harmonics retained. In Fig. 5 (d), the solid sky blue curve shows the best predictability that I have seen of a 2D DROM+. And reducing model complexity does not have a favourable effect. It is not clear to me why even the 1D and 2D DROM perform comparatively poorly when they are mathematically equivalent with the X(D)ROM. Otherwise, it is interesting, even if probably not very useful, that the 2D DROM+ can outperform the 10D X(D)ROM for longer lead times. On the other hand, I found that ridding about 70% of the parameters of the XDROM+ $(K = 3)$ yields the best result. However, the improvement is meagre compared to the impressive advantage of the XDROM+ itself over the X(D)ROM. This result suggests to me that we have a "benign overfitting" [20] in the case of the X(D)ROM(+).

3 Discussion

In this paper I scrutinised what we can imply from the single most important result of the Article, namely, what I call here the observed or apparent prediction skill. I demonstrated that it is not indicative of what I regard as the true skill of a data-driven forecast (Fig. 4).

Yet, I make the observation that this true skill, the ensemble-wise RMSE, or REMSE, has a persistence – even if short – as the window of data from which the forecast model is inferred is extending/moving with the passing of time (Fig. 3). Persistence always indicates some determinism and, so, predictability [18]. Therefore, it would be worthwhile to try to predict the true prediction skill. Interesting questions in this regard might be, e.g.: what (hopefully easily interpretable) states of ENSO are responsible for better or worse skill (especially the horrible spikes of the REMSE seen in Fig. 3 (a)); how long memory of ENSO states does the skill have?

The apparent skill evaluated from observations of the unique history of our Earth climate can be useful in comparing models wrt. their predictive power; but also in providing a ballpark figure of skill for a single model. Relating to the latter, I discovered power scaling laws for the ensemble-mean and ensemble-standard-deviation of the apparent skill depending on the observational time series length. The E-mean approaches the climatological skill much faster than the vanishing of the E-std. Unfortunately, though, we have a bottle neck problem regarding the reliability of the ballpark figure of the observed skill: the slow vanishing of the E-std, the variability, would keep our approximation of the climatological skill by the apparent skill uncertain, *inaccurate*. Mind how the "technological" (ISO 5725) definition of "accuracy" [21] combines the "trueness" (no bias, E-mean) and "precision" (variability, E-std).

Finally, I remark that from Fig. 2 (m) of the Article it is unclear if the XROM beats the AI-based model [2] wrt. prediction skill. Indeed, the main contribution of the Article that it offers, instead, as per the title itself, is the *explanation* of the enhanced prediction skill of the XROM over the ROM. – We have a binary alternative and we were promised to get an explanation for the outcome. However, I do not see much of an explanation why one or the other model – inferred from data – would have better prediction skill. Or we can see the situation in the way that the substance of the explanation is exhausted by the statement of the title. Because if we can take it for granted that some major modes of climate variability can causally influence ENSO, then it should go without saying that the prediction skill of a model that includes those modes – ideally, i.e., regarding the theoretical maximum skill – is better than what has only the local ENSO variables (T, h) (say, the ROM). And I think we can take it for granted, because in a fully coupled system we can know *apriori* that there is information flow both ways concerning any two state variables; the chances for a one-way flow/coupling, i.e., a master-slave relationship, is zero. It is

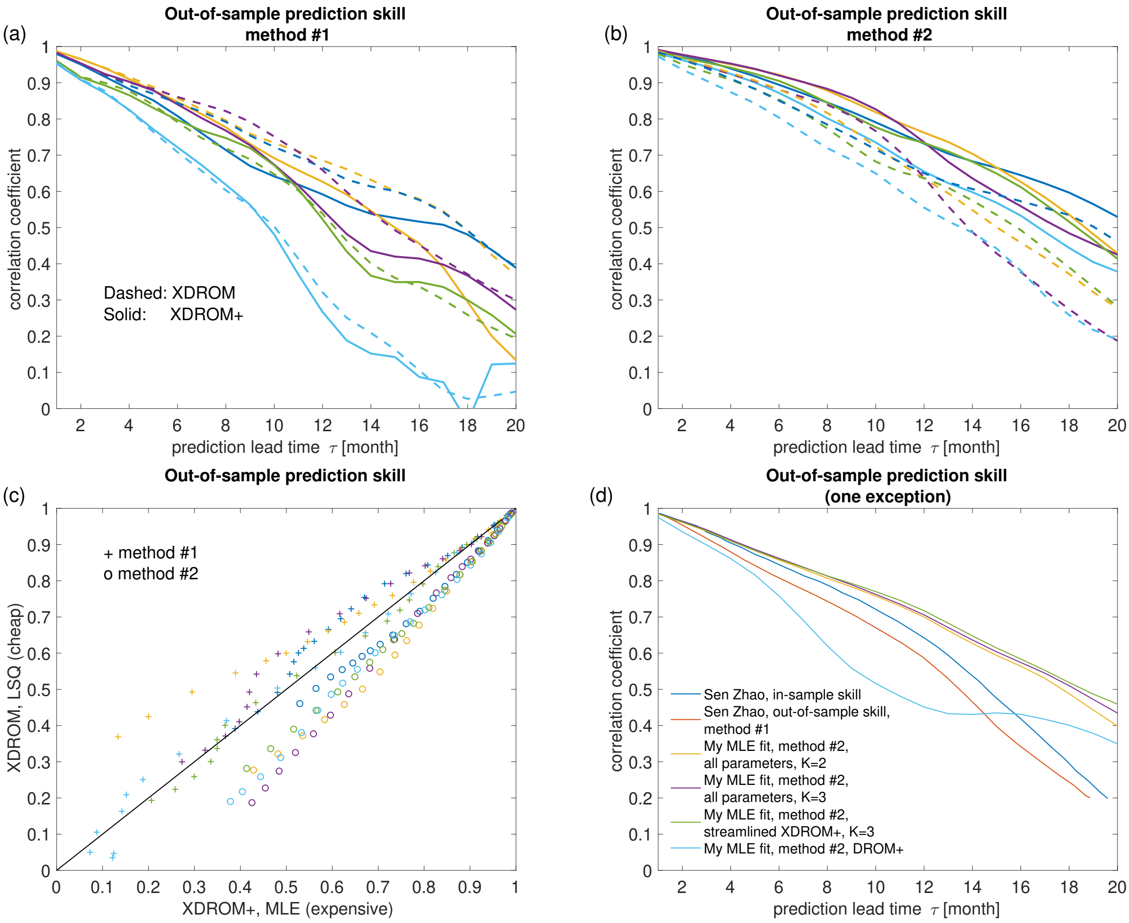

Figure 5: Improvement of the prediction skill in view of the apparent skill for a small ensemble of synthetic observations. (a,b) XDROM vs XDROM+ and method #1 vs #2. (c) A scatter plot collecting all data from (a,b). The colour of markers in (c) correspond to the colour of curves in (a,b). The lead time τ is not explicitly indicated in (c) but can be inferred since the apparent skill is monotonically decreasing with τ. (d) Improving skill by model complexity reduction.

akin to the situation with, say, a one-sample t-test for the mean of a population in a *practical* situation, when we know apriori that the null-hypothesis is false, because the chances that it is true are zero. (Thus, if we are not able the reject the null-hypothesis, it is because we do not have enough data.)

Much of what the authors present are results in terms of numerical figures for the "contributions" of different modes of variability to the prediction skill. They come up with these by performing three sets of experiments as follows (I quote): "(1) uninitialized experiments (referred to as U_j), (2) decoupled experiments (D_j) and (3) relaxation towards observations experiments (R_j)". Mind that we can speak about "contributions" only if the principle of superposition applies [10], and the authors claim that in the case of the *nonlinear* XROM "...the uninitialized experiment framework is a suitable approach to quantify the nearly additive relative contributions of each basin to ENSO forecast skill". It is unfortunate that the authors do not make their exposition in this regard in the context of response theory which is the mathematical basis of the principle of superposition. Thus, we are left unsure what is the forcing and what is the system that responds to a forcing (linearly). Anyhow, what has it to do with the promised explanation?

In his monumental yet slim volume "The Structure of Scientific Revolutions" [22], Thomas S. Kuhn posits that *a theory has only so much explanatory power as predictive power*. Had the numerical figures of the said contributions been meant to be explained – which is certainly ***not*** what the Article's title refers to – then there is no such explanation in the Article in the sense of Kuhn. I would be desperate if I was tasked to come up with such explanations, I admit. Vague references to physical processes have no chance of suggesting approximate numerical figures for contributions. Making such references to physical processes in an attempt of an explanation is usually called a “handwaving explanation”. I am afraid that this is not what people believe the journal *Nature* would knowingly target.

I do think that the rather achievable task is the explanation promised indeed by the Article’s title, i.e., explaining the outcome of the said binary alternative. However, I think, it has to do with mathematics rather than geophysics. (See also e.g. how some of the same authors purport to explain [23] a statistical fluke [24], which had been well known to some forecasting experts [17]. And the cherry on top is the anecdote that I shared on Linkedin [25].) Otherwise, as I argued above, any explanation is lacking because we are not so much interested in the theoretical maximum skill, but rather the skill of a model inferred from finite data. Considering that the promised explanation should be mathematical in nature, references to geophysical effects by Zhao et al. are based on hindsight, which could at best be regarded as *interpretations* – instead of explanations. What are not trivial and remain open questions, waiting for appropriate mathematical analyses, are: why is the overfitting of the X(D)ROM(+) benign (had [20] not actually had it explained already); is there a certain time series length when e.g. the prediction skills of the XROM fitted to 10 time series and the ROM to 2 time series of that same length break even (similarly to the case of the XDROM versus the XDROM+); if so, why *that* length; why does the XDROM have a memory of roughly only one month; why does e.g. the 2D DROM of any large D have worse skill than the 10D XDROM, $D = 1$?

I would like to point out that I myself did not supply any explanation as to why the XDROM+ can outperform the XDROM only if fed more data than a certain critical amount. I do not quite know yet how to tackle that challenge, actually. A further question that I pose is: why is the AI-based model completely unable to account for the ingredients of the XDROM+ over the XROM?

My critique of Zhao et al. that they did not deliver on their promise implies that *the authors misrepresent their merit.* (I will not speculate if they do this knowingly or not.) I think their actual merit derives from, ad 1, coming up with the XROM, and, ad 2, providing the numerical figures for the contributions. To me, the former is far more admirable than the rather straightforward analysis of the latter – even if the latter is also genuine scientific novelty given that the contributions cannot be guessed, but one needs to perform calculations to find them out. Otherwise, the latter seems to be used to “beef up” the paper, on the one hand, and to “have a story to spin”, on the other, perhaps out of a sense of insecurity that coming up with the XROM alone is not seen valuable enough by an editor of a “luxury journal”. Unfortunately, I observe a “gold rush” in the Earth sciences where scientists make each other believe that so-called physical explanations are the holy grail of science. (Does the sound of a title “Unravelling the physical mechanism of [*phenomenon*]” ring familiar? Use the word “unravelling”, and you are one step closer to success; if you unravel a physical mechanism, actually or purportedly, then you might well be in, had you ticked other important boxes [26].) Thus, scientists will feed editors and their peers with what they expect. No wonder, then, that in this “echo chamber” a lot of serious issues with the quality of science are overlooked. And the few who could and would point it out often do not have a voice.

I think, we see an article here that exemplifies a grave problem with our current system of scientific publishing including – if the manuscript is not desk-rejected [26] – peer review. None of the three reviewers of the Article noticed A.) the said misrepresentation of merit and B.) that the promised explanation will be mathematical not physical. This story turned out so partly because the editor himself is not competent enough, inviting only Earth scientists to reviewing the Article. I think the lesson here is that we need to train ourselves, and the next generation of scientists, to look deeper at problems, and from various angles. Earth scientists too often lack sufficient or even basic mathematical competence [23, 24], let alone some training in philosophy, because of which science evidently badly suffers. I cannot resist pointing out the irony that despite of this trend, the primary scientific title also in the STEM subjects is still called PhD: Doctor of Philosophy.

4 Methods

4.1 Variables and data used

No new data sets and variables have been considered for this study, only those of the original study, archived online on Zenodo [6].

4.2 Model inference of the XDROM/XDROM+ by LSQ/MLE and model simulation

Using the Euler-Maruyama SDE numerical integration scheme, the stochastic version of the "forward Euler" scheme, we can turn the system of SDEs (1) into a first-order (1) nonlinear (NL) vector (V) autoregressive (AR) (discrete time, stochastic) model. It can be codenamed an NLVAR(1) model. Likewise, the XDROM can be turned into an NLVAR(D) model. Matlab's `nlarx` would be a perfect tool to handle such a model wrt. model inference, and, then, there is `simulate` and `predict`, all to serve our purposes. Unfortunately, `nlarx` yields a nonphysical model whose simulation quickly blows up. Likely, there is a bug in the code, or the handling of a model with a very high-dimensional parameter space is not robustly done. Fortunately, the method of ordinary least squares (LSQ) can be applied for model inference, instead. As Zhao et al. suggested, instead of searching for phase parameters like Φ_c^{ijk} in eq. (4), we would better turn it into the form of

$$c^{ij} = c_0^{ij} + C_k^{ij} \cos(k\Omega t) + S_k^{ij} \sin(k\Omega t).$$

This way we are left with a *linear* regression problem for the inference of model parameters, which, instead of an expensive numerical minimum search, can be solved by matrix inversion (to do with the fact that the objective function, the "sum of squares", is a convex function of the model parameters with a single minimum, in which case the initialisation of the minimum search can be arbitrary). Matlab's nothing less than awesome "backslash operator" \, also known as `mldivide`, handles the XDROM, $D = 1$, featuring > 1k parameters with ease and accuracy. – There is no need for the convoluted method of Chen & Jin (2021) trying to reduce the size of a matrix to be numerically inverted. For the purpose of model complexity reduction (Methods 4.3), we can obtain standard errors for the parameter estimates by resorting to a minimum search using Matlab's `fminunc` that outputs the Hessian matrix of entries h_{ij}. Then, the standard errors are: $SE_i = \sqrt{h_{ii}^{-1}}$, taking the square roots of the diagonal entries of the inverse of the Hessian. One can do themselves a favour and initialise the minimum search by the solution of the LSQ problem obtained by using the backslash operator. Unfortunately, the inversion of the Hessian by Matlab's `inv` can fail for a too large matrix. That is, increasing the number K and D of harmonics and delay variables, respectively, face a limit.

The coefficients σ_i^ξ are simply obtained from the raw fitting residuals, as usual. The raw residuals $\hat{\xi}_i$, in ways of model/fit diagnostics, are typically found serially not significantly correlated already with $D = 1$, as assumed, but significantly cross-correlated in some cases. Especially $\rho[\hat{\xi}_T, \hat{\xi}_h]$ is very significant, and figuring as about 0.4. When it comes to simulating the model, for authenticity, we ought to prescribe cross-correlated white noise forcing: $\sigma_i^\xi \xi_{i,n} = l_i^j \zeta_{j,n}$, where $\zeta_{i,n}$ are uncorrelated normally distributed white noise realisations of zero mean and unit variance and the square *lower* triangular matrix l_i^j can be obtained by a Cholesky decomposition (e.g. using Matlab's `chol`) because $l_i^k l_k^j = q_i^j$, where $q_i^j = \hat{\xi}_{i,n}\hat{\xi}^{j,n}/(N-1)$ is the sample covariance matrix. (Ask for a proof from ChatGPT.)

The XDROM+ cannot be inferred by LSQ. Violation of LSQ's assumption of additive noise can result in surprising effects, such as the sporadic fractality of the objective function [10]. I have managed to perform the fitting of XDROM+ to data instead by Maximum Likelihood Estimation, MLE, eq. (7.83) of [19] providing the likelihood function. (See also the Methods section of [10].) Here, we cannot avoid performing numerical minimum search, which is very expensive. In general, for a model of the complexity of the XDROM+, it would be most likely unfeasible. The minimum search needs to be *initialised*, and I was not able to obtain a solution even with the most sensible initial conditions. Instead, it turned out to be key to ***start near the solution***. Fortunately, in the case of the XDROM+ it is possible, because it is a rather modest extension of the XDROM, which we could fit to data with ease. Then, initial conditions for the few more parameters that the XDROM+ features over the XDROM can be chosen simply as zeros. Simulation of the XDROM+ too requires generating cross-correlated noise forcing time series as described above.

See Matlab scripts in the Zenodo archive [27] that should be helpful in reproducing the results reported here.

4.3 Model complexity reduction

Two simple approaches have been considered for model complexity reduction. One is via eliminating state variables. As a benchmark model for the Article itself, they considered a commonly used 2D recharge oscillator model (ROM) of ENSO featuring only the first two "ENSO variables" of the XROM, T and h. A first-order (as in "VAR(1)") ROM of ENSO is employed most commonly. The rewriting of the XROM would, however, lead rather to an infinite-order 2D NLVAR: $\lim_{\omega\to\infty}\mathrm{NLVAR}(\omega)$, in which we have delay variables of order d higher than 1, i.e., the immediate future depends on the entirety of the past. – Memory terms are well known to emerge upon such model reduction; see e.g. [11] and references therein. The "distribution" of coefficients of the delay variables encodes the signature of the teleconnection network. If we neglect the (quadratic) nonlinearity in eq. (3) for the IOD (which I have checked to have no detectable effect on the model's prediction skill), then rewriting the XROM of a 10D NLVAR to the 2D NLVAR is fairly tractable. We gain further insight if we keep only T, omitting even h, to start with. Substituting formulae for the (delay) variables of x_i, $i > 1$, into the equation for x_1, recursively, the formulae being given by the equations (1) for those variables, one would end up with time dependent coefficients $\cos^p(\Omega t)\sin^q(\Omega t)$ for $x_{1,(n-d)}$, the larger p, q, the larger d. Using trigonometric "power-reduction formulae" [28] (see also `https://www.youtube.com/watch?v=jClj4S4qJ8M` at e.g. 13:40), the coefficients of $x_{1,(n-d)}$ are typically periodic functions of time of an increasing number of harmonics proportional with d. Looking farther into the past, this would result in an exponentially increasing number of model parameters, which is clearly unaffordable very quickly. Conveniently, however, conditions farther in the past matter less, therefore, we can apply a memory cutoff, desirably at a number of time steps when coefficient estimates are still significant. In fact, this needs not to be checked directly; it is enough if we optimise for the cutoff in terms of the best prediction skill. These considerations certainly apply when state variable h is retained. One might want to resolve to cutoff the number of harmonics uniformly for delay variables and retain more delay variables/memory. I say this also because even LSQ, using Matlab's backslash operator, faces a limit of the dimensionality of the model parameter space. It is not clear to me, though, if in the case of 44 years worth of data we would meet sooner the latter problem or the saturation of the number of significant parameters as we increase the model parameter space (by increasing D and K).

The other simple approach is discarding parameters of the XROM/XDROM/XDROM+ whose estimates are not statistically significant. The aim is to find an optimal model of some reduced nontrivial number of parameters, with the best prediction skill. In order to eliminate parameters p one by one, I look at their estimate's ($\hat{p}$) relative distance from zero: $\lambda|\hat{p}|/SE$, which is something like the reciprocal of the well known "coefficient of variation". In here, SE is the standard error of estimation. Clearly, ordering the parameters wrt. this distance will yield the same result whatever the value of λ is. Therefore, having in mind statistical significance, which is defined in a way that the confidence interval (proportional with the SE) excludes zero, I take this order for a step by step model reduction.

See Matlab scripts in the Zenodo archive [27] that should be helpful in reproducing our results reported here.

Acknowledgements

I acknowledge that no language assistance software was used to write this article other than basic spell and grammatical checking provided by the TeXShop free document development system. I am grateful for Richard Koch and his many co-workers for gifting scientists and other authors such an amazing tool. I used a ("very recent") thesaurus, though, the Google Chrome browser engine, for which Alphabet, the parent company of Google, whose online search monopoly has been ruled illegal in the US recently (`https://www.bbc.com/news/articles/c0k44x6mge3o`), is reservedly acknowledged. I also acknowledge OpenAI and Quora Inc., the respective makers of the general purpose chat bots of ChatGPT and Poe, who still provide free services at the time of writing on 28.08.2024., but already impose meagre daily allowances. I thank Soong-Ki Kim and Roman Olson for useful discussions about the XROM and the Article. I'm especially thankful to Roman who introduced me to the MLE estimation of SDEs (Methods. 4.2) before I started to work with him on the paper [10], which method was the key to beat the predictive power of the previous record holder XROM. I thank an anonymous Mathworks staff member for advising me about the usage of Matlab's `nlarx` and for the information that the sum of squares is a convex function of the model parameters. I am indebted to Sen Zhao for providing me simulation output data using his Python code [6] and also for feedback on an earlier version of my manuscript. The feedback/rebuttal was apparently coordinated by Fei-Fei Jin and Malte Stuecker whom I therefore also thank. This work received funding from the Science Excellence Programme of the Hungarian University of Agriculture and Life Sciences (`https://research.uni-mate.hu/hu/class-of-2024`).

Competing Interests

I declare that I have no competing financial interests.

Open research

Matlab scripts and data files including the ORA reanalysis data in [6] for validating and reproducing, etc., the results in this paper are publicly available [27]. The data for the green line in Fig. 2, provided kindly by Sen Zhao upon my request, is also included in this archive. I grabbed the data from the received PNG file using the fantastic Matlab tool `grabit` by Jiro Doke [29]. The most expensive figure to compute was Fig. 2 (e), which took an overnight simulation on 8 cores of an Apple Silicon M3 processor. Fig. 5 (a-c) took a similar run time.

References

[1] Sen Zhao, Fei-Fei Jin, Malte F. Stuecker, Philip R. Thompson, Jong-Seong Kug, Michael J. McPhaden, Mark A. Cane, Andrew T. Wittenberg, and Wenju Cai. Explainable El Niño predictability from climate mode interactions. *Nature*, 630(8018):891–898, 2024.

[2] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775):568–572, 2019.

[3] Malte F. Stuecker, Axel Timmermann, Fei-Fei Jin, Yoshimitsu Chikamoto, Wenjun Zhang, Andrew T. Wittenberg, Esther Widiasih, and Sen Zhao. Revisiting ENSO/Indian Ocean Dipole phase relationships. *Geophysical Research Letters*, 44(5):2481–2492, 2017.

[4] Tamás Bódai, Sundaresan Aneesh, June-Yi Lee, and Sun-Seon Lee. Decadal Indian Ocean Influence on the ENSO-Indian Monsoon Teleconnection Mostly Apparent. *Journal of Geophysical Research: Atmospheres*, 128(15):e2023JD038673, 2023.

[5] Han-Ching Chen and Fei-Fei Jin. Simulations of ENSO Phase-Locking in CMIP5 and CMIP6. *Journal of Climate*, 34(12):5135 – 5149, 2021.

[6] Sen Zhao. Extended nonlinear Recharge Oscillator (XRO) model for "Explainable El Niño predictability from climate mode interactions", February 2024. `https://doi.org/10.5281/zenodo.10681114`.

[7] Mickaël D. Chekroun, Eric Simonnet, and Michael Ghil. Stochastic climate dynamics: Random attractors and time-dependent invariant measures. *Physica D: Nonlinear Phenomena*, 240(21):1685–1700, 2011.

[8] Gábor Drótos, Tamás Bódai, and Tamás Tél. Probabilistic Concepts in a Changing Climate: A Snapshot Attractor Picture. *Journal of Climate*, 28(8):3275 – 3288, 2015.

[9] Roman Olson, Soon-Il An, Soong-Ki Kim, and Yanan Fan. A novel approach for discovering stochastic models behind data applied to El Niño–Southern Oscillation. *Scientific Reports*, 11(1):2648, Jan 2021.

[10] Tamás Bódai, Soong-Ki Kim, and Roman Olson. Climatology and natural and forced changes of ENSO variability, 2024. unpublished manuscript.

[11] Christian L. E. Franzke, Terence J. O'Kane, Judith Berner, Paul D. Williams, and Valerio Lucarini. Stochastic climate theory and modeling. *WIREs Climate Change*, 6(1):63–78, 2015.

[12] T. Tél, T. Bódai, G. Drótos, T. Haszpra, M. Herein, B. Kaszás, and M. Vincze. The Theory of Parallel Climate Realizations. *Journal of Statistical Physics*, 2019.

[13] G. Drótos and T. Bódai. On defining climate by means of an ensemble. *ESS Open Archive*, 2022.

[14] Tamás Bódai and Tamás Tél. Annual variability in a conceptual climate model: Snapshot attractors, hysteresis in extreme events, and climate sensitivity. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(2):023110, 04 2012.

[15] Michelle L'Heureux. The Spring Predictability Barrier: we'd rather be on Spring Break. *NOAA,* climate.gov, 2015. https://www.climate.gov/news-features/blogs/enso/spring-predictability-barrier-we%E2%80%99d-rather-be-spring-break.

[16] Wansuo Duan and Chao Wei. The 'spring predictability barrier' for ENSO predictions and its possible mechanism: results from a fully coupled model. *International Journal of Climatology*, 33(5):1280–1292, 2013.

[17] Sarah Hallerberg, Eduardo G. Altmann, Detlef Holstein, and Holger Kantz. Precursors of extreme increments. *Phys. Rev. E*, 75:016706, Jan 2007.

[18] Tamás Bódai. Predictability of threshold exceedances in dynamical systems. *Physica D: Nonlinear Phenomena*, 313:37–50, 2015.

[19] Thomas Ljung. *System Identification: Theory for the User.* Prentice Hall, 1999.

[20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[21] Wikipedia contributors. Accuracy and precision — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Accuracy_and_precision&oldid=1246807706, 2024. [Online; accessed 5-October-2024].

[22] Thomas S. Kuhn. *The Structure of Scientific Revolutions.* University of Chicago Press, 1962.

[23] Wenju Cai, Benjamin Ng, Tao Geng, Lixin Wu, Agus Santoso, and Michael J. McPhaden. Butterfly effect and a self-modulating El Niño response to global warming. *Nature*, 585(7823):68–73, 2020.

[24] Wenju Cai, Benjamin Ng, Tao Geng, Lixin Wu, Agus Santoso, and Michael J. McPhaden. Addendum: Butterfly effect and a self-modulating El Niño response to global warming. *Nature*, 591(7849):E14–E15, 2021.

[25] Bódai, Tamás. Reluctant to call out corruption and lack of integrity on a grand scale. Linkedin, https://www.linkedin.com/posts/tamas-bodai-65236261_thousands-of-scientists-publish-a-paper-every-activity-7018021292314415104-vHju?utm_source=share&utm_medium=member_desktop, 2024. [Online; accessed 8-October-2024].

[26] Bódai, Tamás. Answering "Why do some papers get rejected by journals without being reviewed? Is it considered unfair to not inform authors of the reason for rejection?". Quora, https://qr.ae/psVgPM, 2024. [Online; accessed 6-October-2024].

[27] Tamás Bódai. Matlab codes implementing the XDROM+ data-driven ENSO forecast model and some analysis of it, October 2024.

[28] Wikipedia contributors. List of trigonometric identities — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_trigonometric_identities&oldid=1238105268, 2024. [Online; accessed 27-August-2024].

[29] Jiro Doke. GRABIT, 2024. https://www.mathworks.com/matlabcentral/fileexchange/7173-grabit.