Supplementary Information for "A data-driven model of ENSO diversity"

Cristian Martinez-Villalobos[1,2], Antonietta Capotondi[3,4], Clara Deser[5,6], Boris Dewitte[7,8,9], Neil J. Holbrook[10,11], Matthew Newman[4], Cécile Penland[4], Daniel J. Vimont[12], and Andrew T. Wittenberg[13]

[1]Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez
[2]Data Observatory Foundation, ANID Technology Center No. DO210001
[3]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA
[4]Physical Sciences Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA
[5]NSF National Center for Atmospheric Research, Boulder, CO, USA
[6]Climate and Global Dynamics Laboratory, NCAR, 1850 Table Mesa Drive, Boulder, CO, USA
[7]Center for Ecology and Sustainable Management of Oceanic Islands (ESMOI), Departamento de Biología Marina, Facultad de Ciencias del Mar, Universidad Católica del Norte, Chile
[8]Centro de Estudios Avanzados en Zonas Aridas (CEAZA), Coquimbo, Chile
[9]CECI, Université de Toulouse III, CERFACS/CNRS, Toulouse, France
[10]ARC Centre of Excellence for the Weather of the 21st Century, University of Tasmania, Hobart, Tasmania, Australia
[11]Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania, Australia
[12]Department of Atmospheric and Oceanic Sciences, University of Wisconsin-Madison, Madison, WI, USA
[13]National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

1 Text S1: Calculation of LIM deterministic and stochastic operators

A LIM can be written as (see equation 1 in the main text):

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}\mathbf{x} + \mathbf{B}\eta. \tag{1}$$

The linear deterministic operator is calculated as:

$$\mathbf{M} = \frac{1}{\tau} log(\mathbf{C}_\tau \mathbf{C}_0^{-1}), \tag{2}$$

where $\mathbf{C}_t = \langle \mathbf{x}(t)\mathbf{x}(0)^T \rangle$ is the lag-covariance matrix at lag t ($\mathbf{C}_0$ is the contemporary covariance matrix). For our calculations we used a lag $\tau = 1$ month.

The noise amplitude matrix $\mathbf{B}$ is determined from fluctuation-dissipation relationship:

$$\mathbf{B}\mathbf{B}^T = -\mathbf{M}\mathbf{C}_0 - \mathbf{C}_0\mathbf{M}^T. \tag{3}$$

2 Supplementary Figures

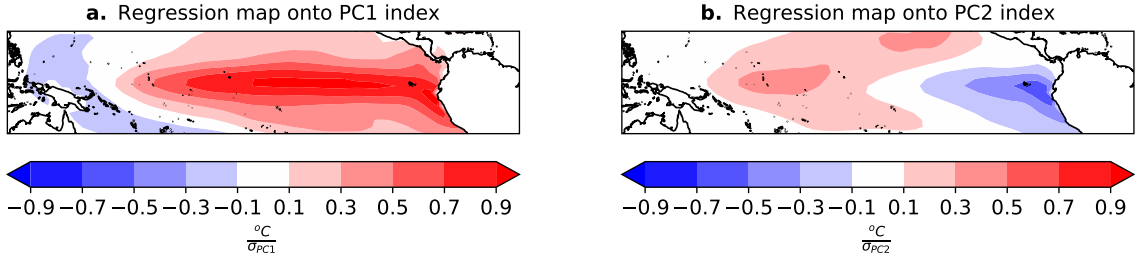

Figure S1: Spatial patterns of **a.** first EOF and **b.** second EOF of tropical Pacific SSTA. These are calculated as the SSTA regression pattern on PC1 and PC2 respectively.

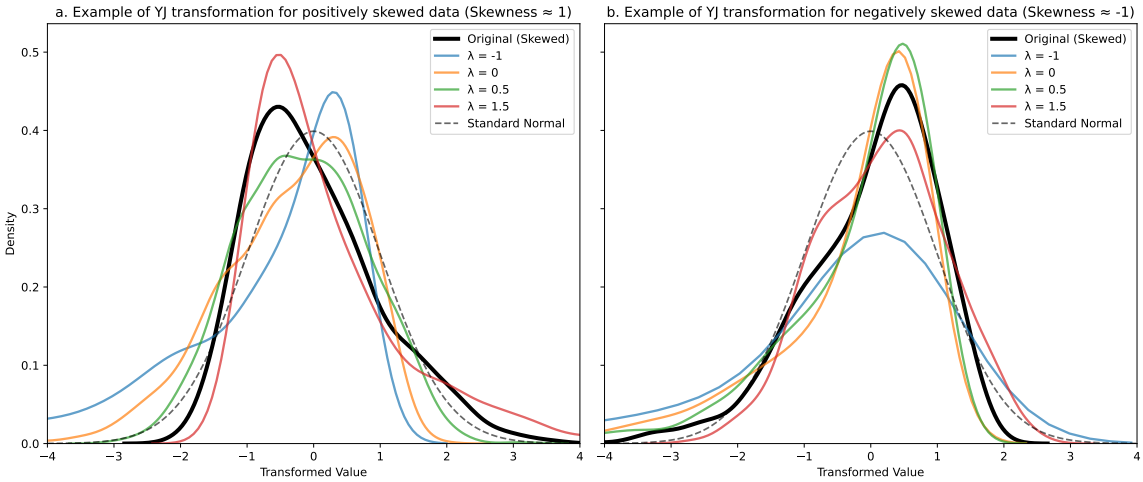

Figure S2: Examples of probability distributions of transformed data when applying the YJ transformation with exponents $\lambda = -1, 0, 0.5, 1.5$ on **a.** positively skewed data ($S \approx 1$) and **b.** negatively skewed data ($S \approx -1$) created synthetically. The original data has mean zero and standard deviation of 1. The dashed line shows a normal distribution for reference.

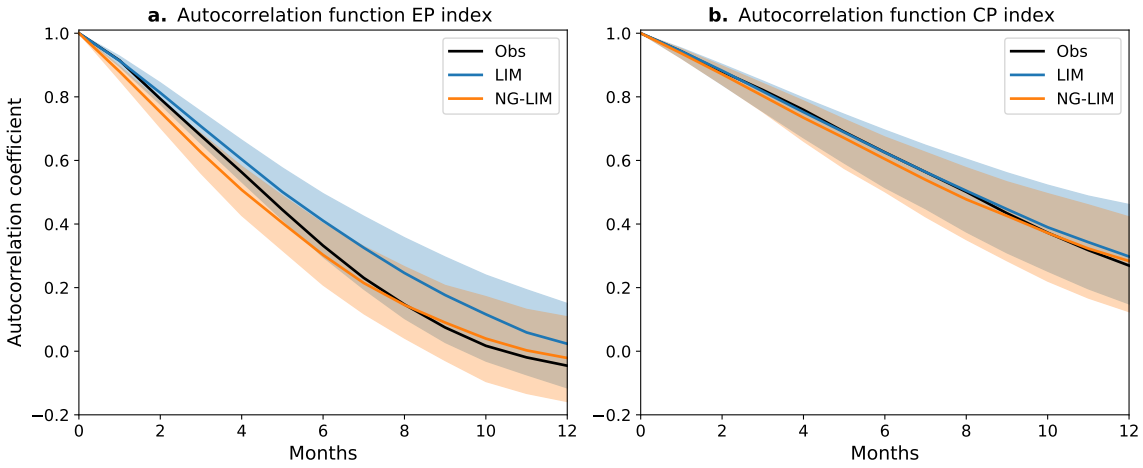

Figure S3: Autocorrelation function of **a.** EP and **b** CP indices in observations (black), median estimation generated by the standard LIM (blue) and generated by the NG-LIM (orange). The shading encompasses the 5th-95th percentiles of the autocorrelation function across epochs of the same length as observations.

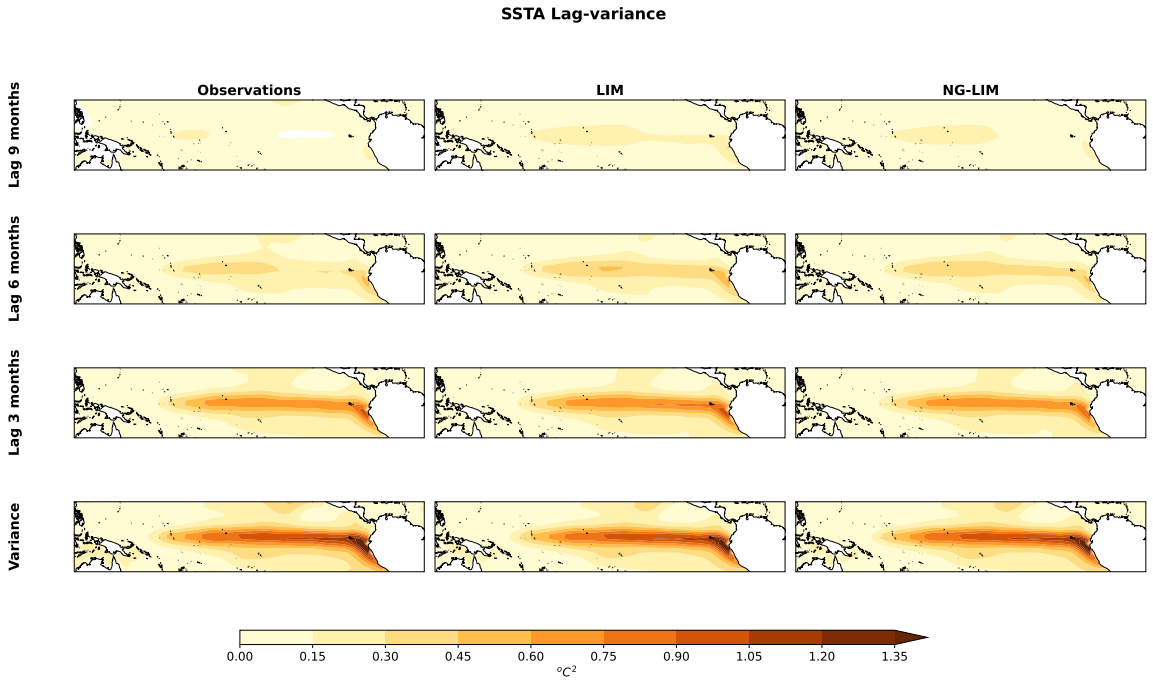

Figure S4: Tropical Pacific SSTA lag-variance spatial pattern in **a.** observations, **b.** generated by the standard LIM, and **c.** generated by the NG-LIM, for lags of 9, 6, 3 and 0 months.

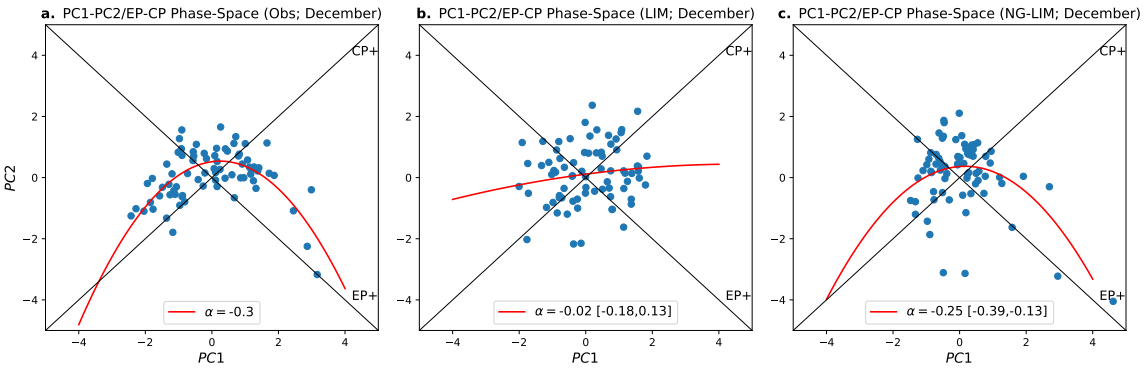

Figure S5: Same as Figs. **2a,b,c** in the main text but only for December.

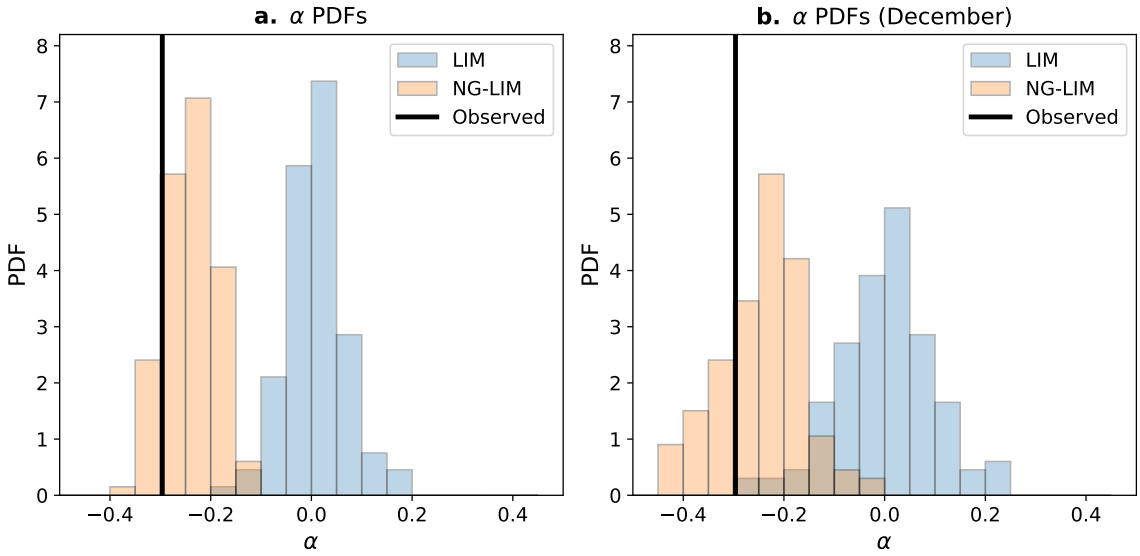

Figure S6: Probability distribution of estimated parameter α across epochs of same length as observations in the standard LIM (blue) and NG-LIM (orange). The black line shows the observed value. The estimations use **a.** annual data, and **b.** only Decembers.

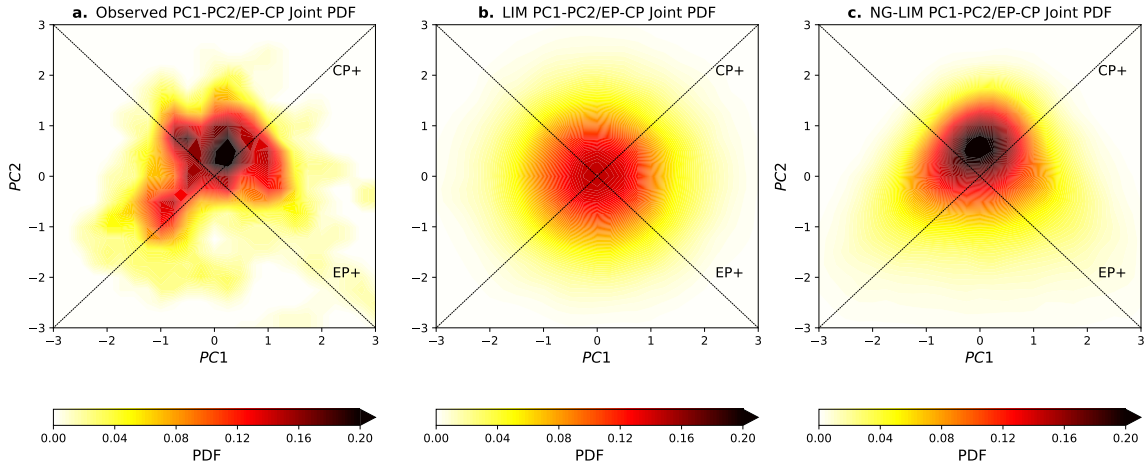

Figure S7: Joint probability distributions of normalized PC1 and PC2 indices in **a.** observations, **b.** standard LIM (whole integration), **c.** NG-LIM (whole integration). All these joint PDFs integrate to 1. The joint EP-CP indices PDF can be seen at the diagonals.

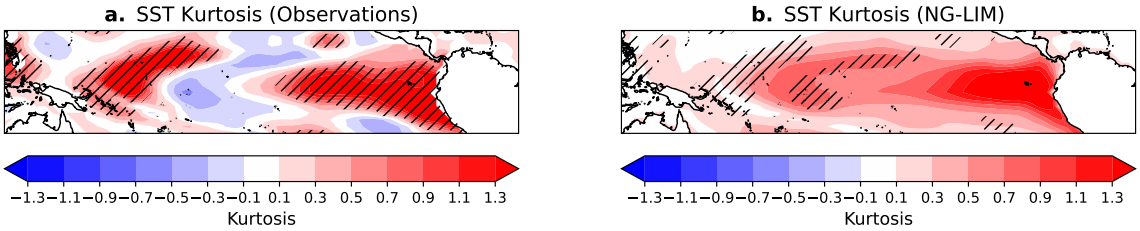

Figure S8: Same as Figures 3a,b in the main text but showing excess kurtosis. Excess kurtosis K is defined as $K(x) = \frac{\langle x^4 \rangle}{\langle x^2 \rangle^2} - 3$. Typically, distributions with $K > 0$ ($K < 0$) have heavier (lighter) tails than Gaussian. The cross-hatching in **a** (**b**) shows regions where the observed SSTA excess kurtosis is outside the 5th-95th percentile range of excess kurtosis generated by the standard LIM (NG-LIM) across epochs

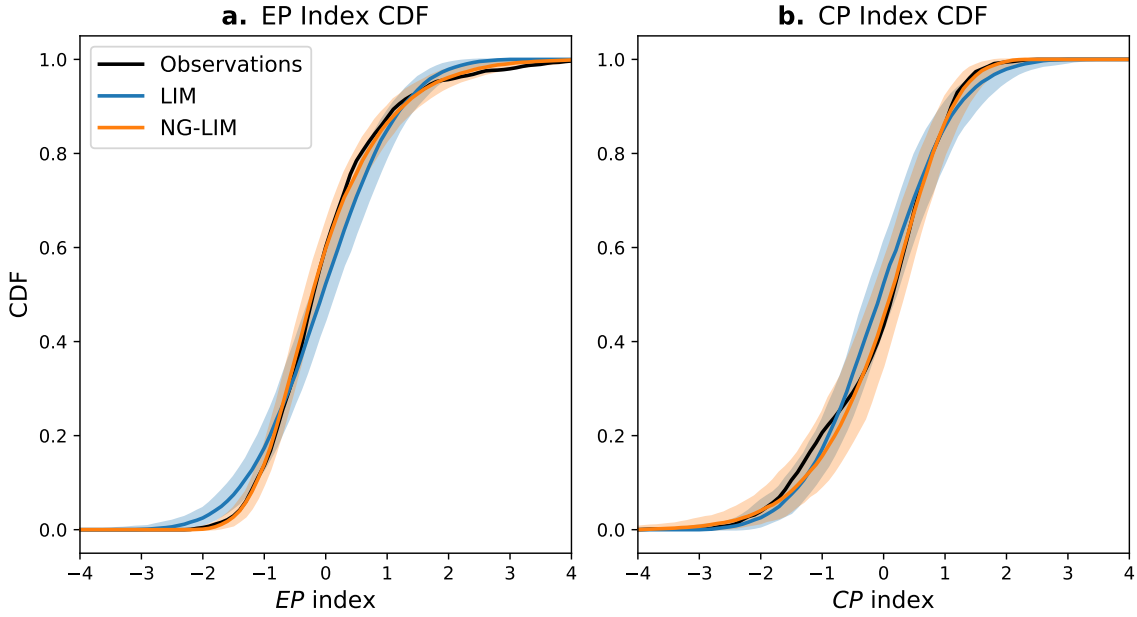

Figure S9: **a.** (**b.**) Cumulative distribution function (CDF) of the EP (CP) index in observations (black), median estimation generated by the standard LIM (blue) and generated by the NG-LIM (orange). The shading encompasses the 5th-95th percentiles of the CDF estimation across epochs of the same length of observations.

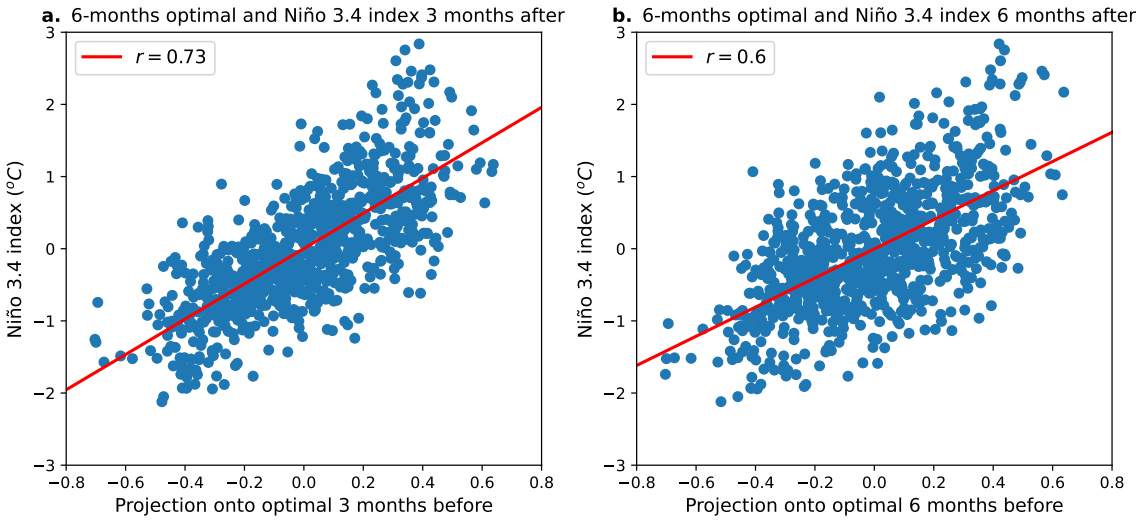

Figure S10: Scatter plot of the projection of SSTA onto the warm optimal pattern and the Niño 3.4 index **a** 3-months later and **b** 6-months later. The optimal projection ranges from -1 to 1, with 1 implying that the SSTA look exactly as in Fig. 4a, up to a rescaling factor.

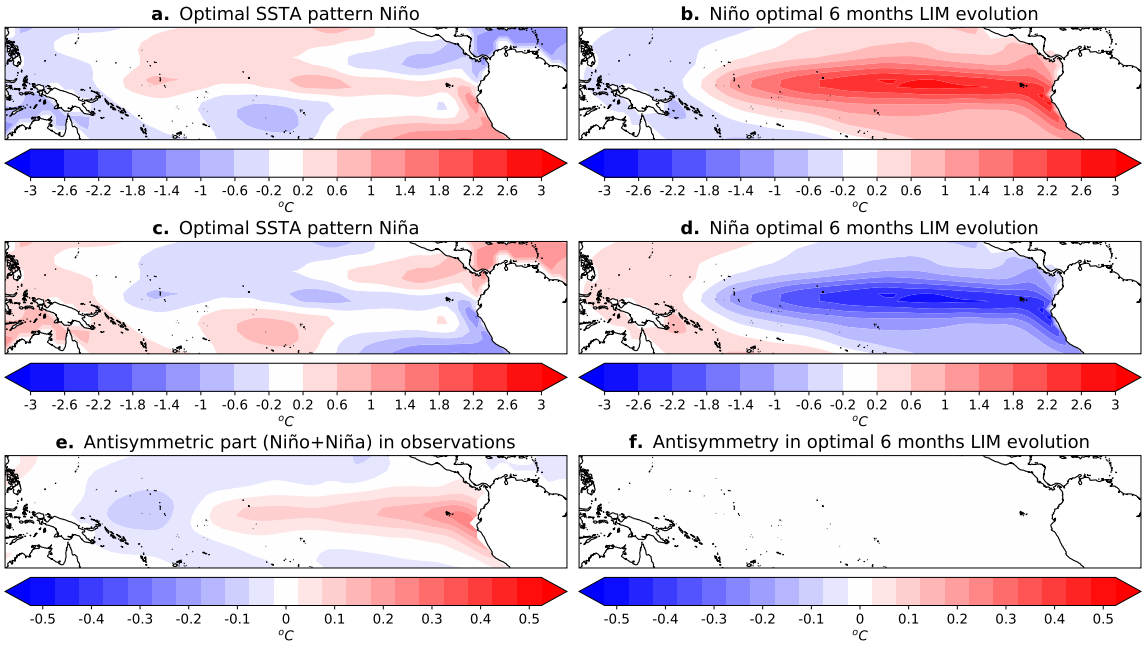

Figure S11: Same as panels 4a to 4f in the main text, but for the standard LIM. Note that as expected, there is no spatial asymmetry in the evolution of the warm and cold optimals in this case (panel f).

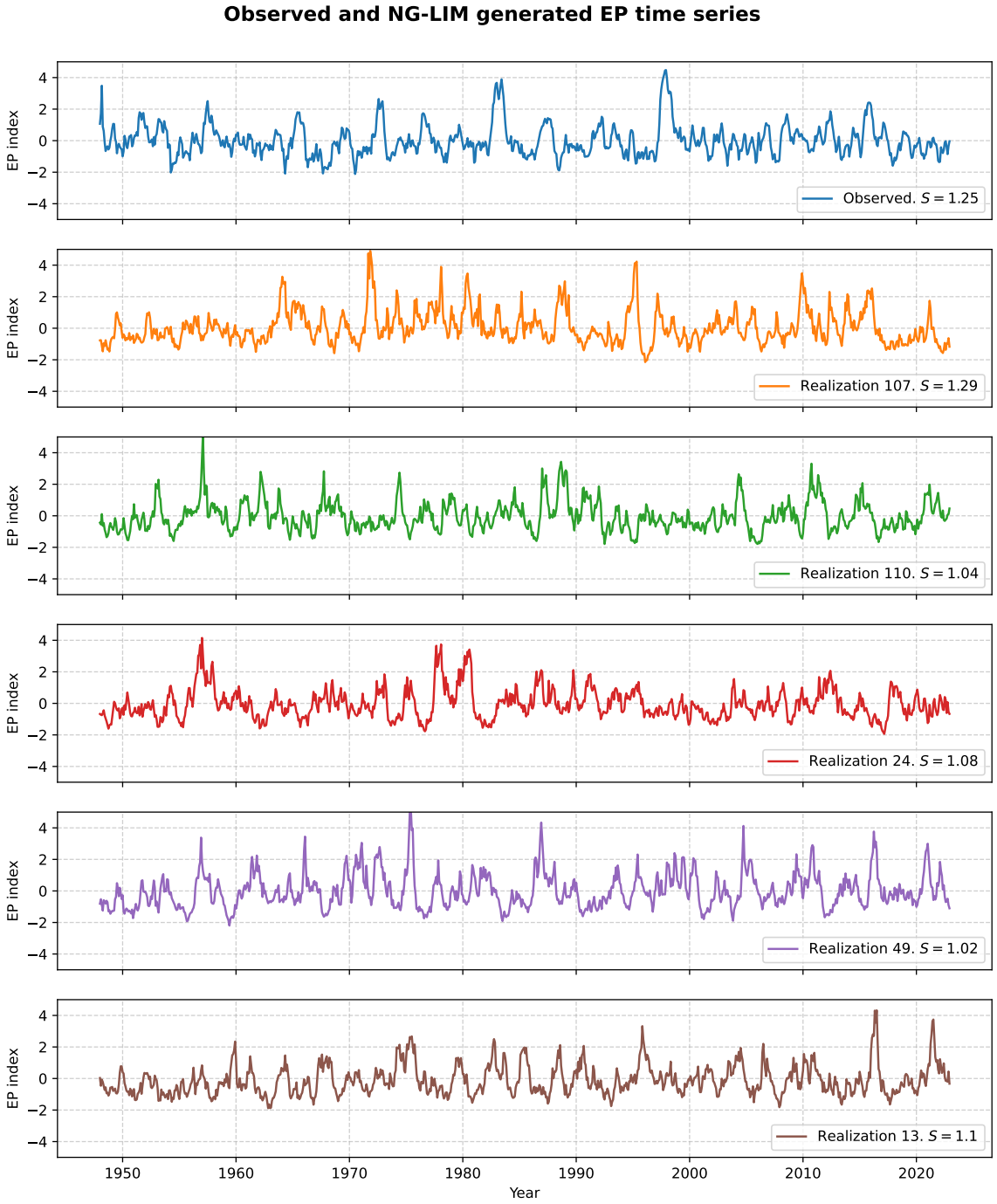

Figure S12: Time series of the observed EP index (top panel) and examples of randomly selected EP realizations generated by the NG-LIM. The legend indicates which realization (out of 133) is shown, along with its corresponding EP skewness coefficient, S

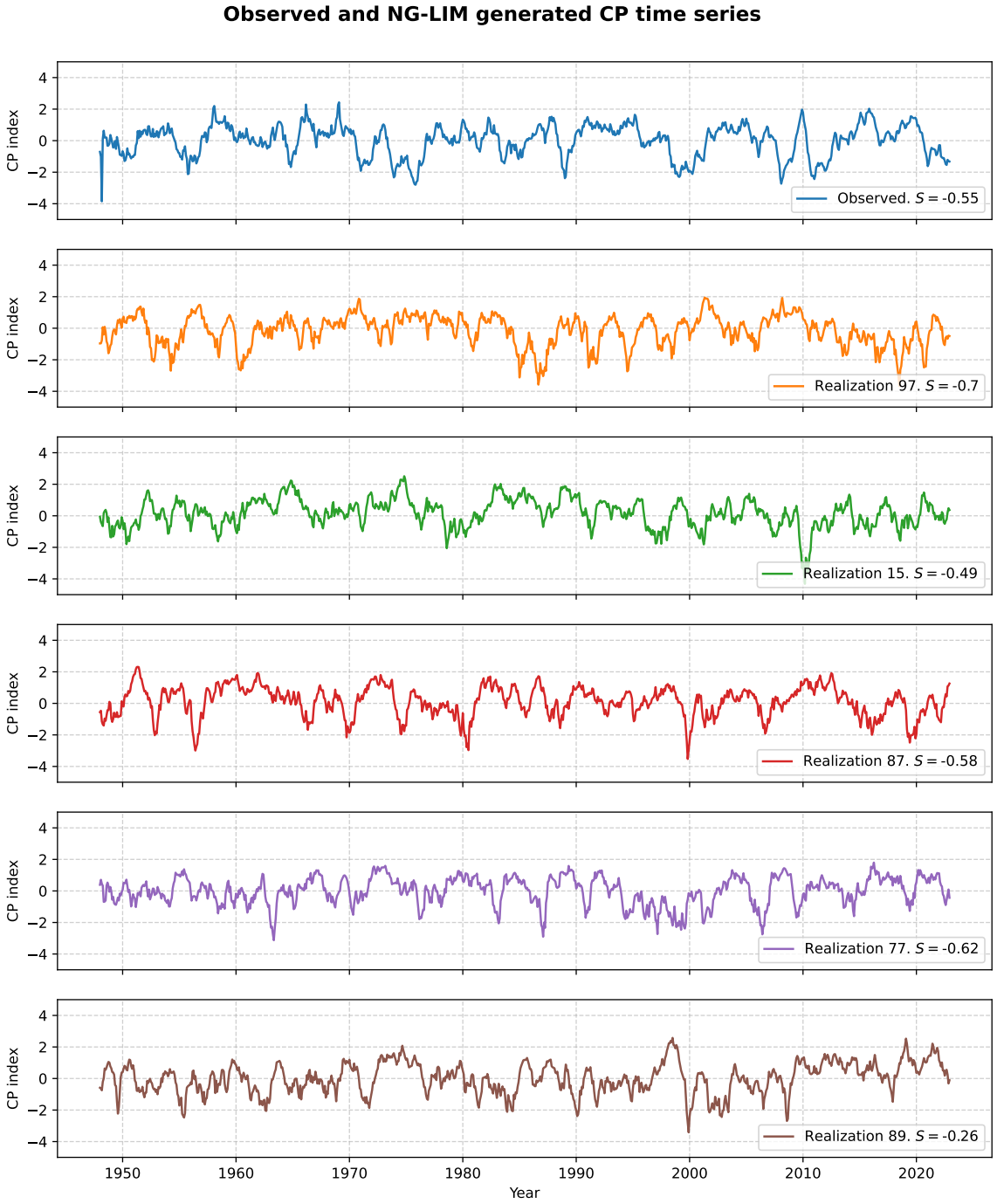

Figure S13: Time series of the observed CP index (top panel) and examples of randomly selected CP realizations generated by the NG-LIM. The legend indicates which realization (out of 133) is shown, along with its corresponding CP skewness coefficient, S