Geophysical Research Letters

Supporting Information for

Long-range Forecasting as a Past Value Problem: Using Scaling to Untangle Correlations and Causality

L. Del Rio Amador[1], S. Lovejoy[1]

[1]Physics, McGill University, 3600 University St., Montreal, Que. H3A 2T8, Canada.

Corresponding author: Lenin Del Rio Amador (delrio@physics.mcgill.ca)

Contents of this file

Introduction
Basic Theory for fGn Processes.
Text S1: Continuous-in-time fGn
Text S2: Discrete-in-time fGn
Text S3: Prediction
Text S4: Cross-correlation function
Text S5: Empirical cross-correlation matrices
Text S6: Co-predictors
Text S7: Simulations
Figures S1 to S11

Additional Supporting Information (Files uploaded separately)

Caption for Movie S1

Introduction

In (Del Rio Amador & Lovejoy, 2019, 2020), the stochastic natural variability component of globally averaged and regional temperature was represented as a fractional Gaussian noise (fGn) process. In order to determine Granger causality, we need to construct a space-time multivariate fGn process that reproduces the cross-correlations. This appendix describes the main technical details.

The main properties of fGn relevant for the present paper are summarized in the following Text S1-S3. We derive expressions for the cross-correlation function for two fGn processes in Text S4. The empirical cross-correlation matrices for the natural temperature variability and for the corresponding innovations for all the 10512 datapoints are presented in Text S5, showing that the innovations have negligible cross-correlations for nonzero lags. In Text S6 we analyze the influence of co-predictors in the forecasts under this condition. Finally, Text 7 gives some specific details for the

simulations of the tropical ocean and compares the autocorrelations and spatial structures of the simulations with global data.

Basic Theory for fGn Processes.

Text S1. Continuous-in-time fGn

An fGn process at resolution τ (the scale at which the series is averaged) has the following integral representation:

$$T_\tau(t) = \frac{1}{\tau}\frac{c_H \sigma_T}{\Gamma(H+3/2)}\left[\int_{-\infty}^{t}(t-t')^{H+1/2}\gamma(t')dt' - \int_{-\infty}^{t-\tau}(t-\tau-t')^{H+1/2}\gamma(t')dt'\right], \quad \text{(S1)}$$

where $\gamma(t)$ is a unit Gaussian δ-correlated white noise process with $\langle\gamma(t)\rangle = 0$ and $\langle\gamma(t)\gamma(t')\rangle = \delta(t-t')$ [δ is the Dirac function], Γ is the Euler gamma function, σ_T is the ensemble standard deviation (for $\tau = 1$) and

$${c_H}^2 = \frac{\pi}{2\cos(\pi H)\Gamma(-2-2H)}. \quad \text{(S2)}$$

This is the canonical value for the constant c_H that was chosen to make the expression for the statistics particularly simple. In particular, the variance is $\langle T_\tau(t)^2\rangle = {\sigma_T}^2\tau^{2H}$ for all t, where $\langle\cdot\rangle$ denotes ensemble averaging. The parameter H, with $-1 < H < 0$, is the fluctuation exponent of the corresponding fractional Gaussian noise process, the Hurst exponent, $H' = H + 1$. Fluctuation exponents are used due to their wider generality; they are well defined even for strongly intermittent non-Gaussian multifractal processes.

Equation (S1) can be interpreted as the smoothing by the fractional integral of a white noise process or as the power-law weighted average of past innovations, $\gamma(t)$. This power-law weighting accounts for the memory effects in the temperature series. The closer the fluctuation exponent is to zero, the larger is the influence of past values on the current temperature. This is evidenced by the behaviour of the autocorrelation function:

$$R_H(\Delta t) = \frac{\langle T_\tau(t)T_\tau(t+\Delta t)\rangle}{\langle T_\tau(t)^2\rangle} = \frac{1}{2}\left(\left|\frac{\Delta t}{\tau}+1\right|^{2H+2} + \left|\frac{\Delta t}{\tau}-1\right|^{2H+2} - 2\left|\frac{\Delta t}{\tau}\right|^{2H+2}\right), \quad \text{(S3)}$$

for $|\Delta t| \geq \tau$. In particular, for $\Delta t \gg \tau$ we obtain:

$$R_H(\Delta t) \approx (H+1)(2H+1)\left(\frac{\Delta t}{\tau}\right)^{2H}, \quad \text{(S4)}$$

which has the power-law behaviour mentioned earlier with the same exponent as the average squared fluctuation and due to the Wiener–Khinchin theorem, it implies a spectrum $E(\omega)\sim\omega^{-\beta}$ with exponent $\beta = 1 + 2H$. For more details on fGn processes see (Biagini et al., 2008; Gripenberg & Norros, 1996; Mandelbrot & Van Ness, 1968).

Text S2. Discrete-in-time fGn

A detailed explanation of the theory for modeling and predicting using the discrete version of fGn processes was presented in (Del Rio Amador & Lovejoy, 2019); the main results are summarized next. The analogue of Eq. (S1) in the discrete case for a finite series, $\{T_t\}_{t=1,\dots,N}$, with length N and zero mean is:

$$T_t = \sum_{j=1}^{t} m_{tj}\gamma_{t+1-j} = m_{t1}\gamma_t + \dots + m_{tt}\gamma_1, \tag{S5}$$

for $t = 1, \dots, N$, where $\{\gamma_t\}_{t=1,\dots,N}$ is a discrete white noise process and the coefficients m_{ij} are the elements of the lower triangular matrix $\mathbf{M}^N_{H,\sigma_T}$ given by the Cholesky decomposition of the autocovariance matrix, $\mathbf{C}^N_{H,\sigma_T} = {\sigma_T}^2[R_H(i-j)]_{i,j=1,\dots,N}$:

$$\mathbf{C}^N_{H,\sigma_T} = \mathbf{M}^N_{H,\sigma_T}\left(\mathbf{M}^N_{H,\sigma_T}\right)^T, \tag{S6}$$

with $m_{ij} = 0$ for $j > i$ (we assume $\tau = 1$ is the smallest scale in our system). The superscript T denotes transpose operation. In vector form, Eq. (S5) can be written as:

$$\mathbf{T}_N = \mathbf{M}^N_{H,\sigma_T}\boldsymbol{\gamma}_N. \tag{S7}$$

Equations (S5-S7) can be used to create synthetic samples of fGn with a given length N, autocorrelation function given by Eq. (S3) and set of parameters $\sigma_T > 0$ and $-1 < H < 0$ (the mean of the series is always assumed equal to zero). Conversely, given an actual temperature series with vector $\mathbf{T}_N = [T_1, \dots, T_N]^T$, we can estimate the parameters σ_T and H using the maximum likelihood method [details are given in Appendix A of (Del Rio Amador & Lovejoy, 2019)] and we can verify that it could be well approximated by an fGn model by inverting Eq. (S7) and obtaining the residual vector of innovations:

$$\boldsymbol{\gamma}_N = \left(\mathbf{M}^N_{H,\sigma_T}\right)^{-1}\mathbf{T}_N. \tag{S8}$$

If the model provides a good description of the data, the residual vector $\boldsymbol{\gamma}_N = [\gamma_1, \dots, \gamma_N]^T$ is a white noise, i.e. the elements should be $\mathrm{NID}(0,1)$ with autocorrelation function $\langle\gamma_i\gamma_j\rangle = \delta_{ij}$ (δ_{ij} is the Kronecker delta and $\mathrm{NID}(0,1)$ stands for Normally and Independently Distributed with mean 0 and variance 1). It is worth mentioning that a white noise process is a particular case of fGn with $H = -1/2$.

Text S3. Prediction

If $\{T_t\}_{t<0}$ is an fGn process, the optimal k-step predictor for T_k ($k > 0$), based on a finite number, m (memory), of past values, is given by:

$$\hat{T}_k = \sum_{j=-m}^{0} \phi_j(k)T_j = \phi_{-m}(k)T_{-m} + \dots + \phi_0(k)T_0, \tag{S9}$$

where the vector, $\boldsymbol{\phi}(k) = [\phi_{-m}(k), \dots, \phi_0(k)]^T$, satisfies the Yule-Walker equations:

$$\mathbf{R}_H\boldsymbol{\phi}(k) = \mathbf{r}_H(k), \tag{S10}$$

with the vector $\mathbf{r}_H(k) = [R_H(k-i)\,]_{i=-m,\ldots,0}^T = [R_H(m+k), \ldots, R_H(k)\,]^T$ and $\mathbf{R}_H = [R_H(i-j)]_{i,j=-t,\ldots,0}$ being the autocorrelation matrix (see Eq. (S3)) (Hirchoren & Arantes, 1998).

The root mean square error (RMSE) for the predictor at a future time k, using a memory of m values, is defined as:

$$\mathrm{RMSE}(k,m) = \sqrt{\left\langle \left[T_k - \hat{T}_k(m) \right]^2 \right\rangle}\,. \tag{S11}$$

The following analytical expression can be obtained:

$$\mathrm{RMSE}(k,m,\sigma_T,H) = \sigma_T \sqrt{1 - \mathbf{r}_H(k)^T \left(\mathbf{R}_H\right)^{-1} \mathbf{r}_H(k)}\,. \tag{S12}$$

For a given forecast horizon, k, the RMSE only depends on the parameters σ_T and H, and the memory used, m.

The theoretical mean square skill score (MSSS), is defined as:

$$\mathrm{MSSS}(k) = 1 - \frac{\left\langle \left[T(k) - \hat{T}(k) \right]^2 \right\rangle}{\left\langle T(k)^2 \right\rangle}\,. \tag{S13}$$

(the reference forecast is the mean of the series, assumed equal to zero here).

From the definition of the RMSE, Eq. (S11), we obtain the theoretical value:

$$\mathrm{MSSS}(k,m,H) = 1 - \frac{\mathrm{RMSE}(k,m,\sigma_T,H)^2}{{\sigma_T}^2}, \tag{S14}$$

or, replacing Eq. (S12):

$$\mathrm{MSSS}(k,m,H) = \mathbf{r}_H(k)^T \left(\mathbf{R}_H\right)^{-1} \mathbf{r}_H(k) = \boldsymbol{\phi}(k) \cdot \mathbf{r}_H(k)\,. \tag{S15}$$

For $H = -1/2$, the fGn process is a white noise process and $\mathrm{MSSS} = 0$. The skill increases with H and the process becomes perfectly predictable when $H \to 0$.

Text S4. Cross-correlation function

Let $T_i(t)$ and $T_j(t)$ be two fGn processes with zero mean and respective parameters σ_{Ti}, H_i and σ_{Tj}, H_j, which could represent, for example, the natural temperature variability at locations "i" and "j", respectively. The cross-covariance function:

$$C_{ij}(t_1,t_2) = \left\langle T_i(t_1) T_j(t_2) \right\rangle, \tag{S16}$$

can be found using the integral representation (Eq. (S1)) for each process:

$$C_{ij}(t_1,t_2) = \frac{1}{\tau^2} \frac{c_{Hi} c_{Hj} \sigma_{Ti} \sigma_{Tj}}{\Gamma\left[H_i + 3/2\right] \Gamma\left[H_j + 3/2\right]} \int_{-\infty}^{t_1} \int_{-\infty}^{t_2} (t_1 - t')^{H_i + 1/2} (t_2 - t'')^{Hj + 1/2} \Big[\left\langle \gamma_i(t') \gamma_j(t'') \right\rangle + \\ + \left\langle \gamma_i(t' - \tau) \gamma_j(t'' - \tau) \right\rangle - \left\langle \gamma_i(t') \gamma_j(t'' - \tau) \right\rangle - \left\langle \gamma_i(t' - \tau) \gamma_j(t'') \right\rangle \Big] dt' dt'' \tag{S17}$$

where we have changed variables in each of the respective second integrals in Eq. (S1).

Let us assume that the white-noise innovations satisfy:

$$\left\langle \gamma_i(t')\gamma_j(t'')\right\rangle = a_{ij}\delta(t'-t''), \tag{S18}$$

where $-1 < a_{ij} < 1$. If we assume that $T_i(t)$ and $T_j(t)$ are jointly wide-sense stationary, then, without loss of generality, we can replace $t_1 = 0$, $t_2 = \Delta t \geq \tau$, in Eq. (S17) to obtain:

$$C_{ij}(\lambda) = c_{H_i}c_{H_j}\sigma_{T_i}\sigma_{T_j}a_{ij}\tau^{H_i+H_j}\frac{\cos(\pi H_j)\Gamma(-2-H_i-H_j)}{\pi} \left[(\lambda+1)^{2+H_i+H_j} + (\lambda-1)^{2+H_i+H_j} - 2\lambda^{2+H_i+H_j}\right], \tag{S19}$$

where $H_i + H_j \neq -1$, $\lambda = \Delta t/\tau$ and c_{H_i} and c_{H_j} are given by Eq. (S2).

For the cross-correlation function:

$$R_{ij}(\Delta t) = \frac{\left\langle T_i(t)T_j(t+\Delta t)\right\rangle}{\sqrt{\left\langle T_i(t)^2\right\rangle\left\langle T_j(t+\Delta t)^2\right\rangle}} = \frac{C_{ij}(\Delta t)}{\sigma_{Ti}\tau^{H_i}\sigma_{Tj}\tau^{H_j}}, \tag{S20}$$

where we replaced Eq. (S16), $\langle T_i(t)^2\rangle = {\sigma_{Ti}}^2\tau^{2H_i}$ and $\langle T_j(t+\Delta t)^2\rangle = {\sigma_{Tj}}^2\tau^{2H_j}$, we obtain:

$$R_{ij}(\lambda) = \frac{a_{ij}}{2}\frac{\Gamma(-2-H_i-H_j)}{\sqrt{\Gamma(-2-2H_i)\Gamma(-2-2H_j)}}\sqrt{\frac{\cos(\pi H_j)}{\cos(\pi H_i)}} \left[(\lambda+1)^{2+H_i+H_j} + (\lambda-1)^{2+H_i+H_j} - 2\lambda^{2+H_i+H_j}\right]. \tag{S21}$$

The integral in Eq. (S17) can also be evaluated for $t_1 = t_2 = \Delta t = 0$. The final result for all cases is:

$$R_{ij}(a_{ij},H_i,H_j,\lambda) = \frac{a_{ij}}{2}\frac{\Gamma(-2-H_i-H_j)}{\sqrt{\Gamma(-2-2H_i)\Gamma(-2-2H_j)}}F_{ij}(\lambda,H_i,H_j) \left[|\lambda+1|^{2+H_i+H_j} + |\lambda-1|^{2+H_i+H_j} - 2|\lambda|^{2+H_i+H_j}\right] \tag{S22}$$

where $\lambda = (t_2 - t_1)/\tau$ is the lag of process "i" with respect to "j" in units of τ and

$$F_{ij}(\lambda,H_i,H_j) = \begin{cases} \sqrt{\dfrac{\cos(\pi H_j)}{\cos(\pi H_i)}}; & \text{if} \quad \lambda \geq 1 \\ \sqrt{\dfrac{\cos(\pi H_i)}{\cos(\pi H_j)}}; & \text{if} \quad \lambda \leq -1 \\ \dfrac{1}{2}\left[\sqrt{\dfrac{\cos(\pi H_i)}{\cos(\pi H_j)}} + \sqrt{\dfrac{\cos(\pi H_j)}{\cos(\pi H_i)}}\right]; & \text{if} \quad \lambda = 0 \end{cases}. \tag{S23}$$

This expression is equivalent to the one obtained by (Coeurjolly et al., 2010), except that they use a different normalization. For $\lambda \gg 1$, we obtain the asymptotic approximation:

$$R_{ij}\left(a_{ij},H_i,H_j,\lambda\right) \approx \varphi_{H_i,H_j} a_{ij} \lambda^{H_i+H_j} . \tag{S24}$$

where

$$\varphi_{H_i,H_j} = \frac{\Gamma\left(-2-H_i-H_j\right)\left(2+H_i+H_j\right)\left(1+H_i+H_j\right)}{2\sqrt{\Gamma\left(-2-2H_i\right)\Gamma\left(-2-2H_j\right)}} \sqrt{\frac{\cos\left(\pi H_j\right)}{\cos\left(\pi H_i\right)}} . \tag{S25}$$

Note that the function $F_{ij}\left(\lambda,H_i,H_j\right)$ satisfies $F_{ij}(\lambda) = F_{ji}(-\lambda)$. The other factors in Eq. (S22) are symmetric with respect to a permutation of the indexes or to the sign of the lag. Hence, the cross-correlation function satisfies the symmetry property for jointly wide sense stationary-processes: $R_{ij}(\lambda) = R_{ji}(-\lambda)$. If $a_{ij} = 1$ and $H_i = H_j$, we recover the autocorrelation function Eq. (S3).

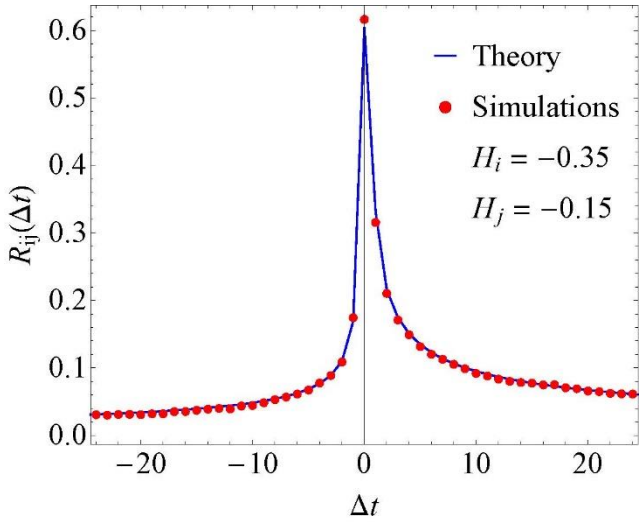

Figure S1. Cross-correlation function for $a_{ij} = 0.7$ and fluctuations exponents $H_i = -0.35$ and $H_j = -0.15$. The blue line is the graph of the theoretical expression Eq. (S22) and the red dots were obtained from a set of 500 pairs of fGn series each $N = 1000$ timesteps long.

Figure S1 shows an example of the cross-correlation function for $a_{ij} = 0.7$ and fluctuations exponents $H_i = -0.35$ and $H_j = -0.15$. The blue line is the graph of the theoretical expression Eq. (S22) and the red dots were obtained from a set of 500 pairs of fGn series each $N = 1000$ timesteps long. The simulations were produced using Eqs. (S5-S7) where the white noise series of innovations, $\{\gamma_i\}$ and $\{\gamma_j\}$, were generated from a multivariate Gaussian distribution in such a way that: $\langle\gamma_i(t')\gamma_i(t'')\rangle = \langle\gamma_j(t')\gamma_j(t'')\rangle = \delta_{t't''}$ and $\langle\gamma_i(t')\gamma_j(t'')\rangle = \langle\gamma_j(t')\gamma_i(t'')\rangle = a_{ij}\delta_{t't''}$ (Notice that we replaced the Dirac by the Kronecker δ for discrete-in-time series). The estimated cross-correlation function for the simulations (red dots) was computed as:

$$R_{ij}(\Delta t) = \frac{\sum_{t=1}^{N-\Delta t} T_i(t) T_j(t+\Delta t)}{\sqrt{\sum_{t=1}^{N} T_i(t)^2 \sum_{t=1}^{N} T_j(t)^2}} . \quad \text{(S26)}$$

Text S5. Empirical cross-correlation matrices

For the dataset described in Sect. 2.1, we estimated the natural temperature variability, $T_i(t)$, and, using the theory presented in Text S2, we obtained the series of innovations for each location, $\gamma_i(t)$. Using Eq. (S26), we estimated the lagged cross-correlation matrices (shown in Fig. S2) for the innovations and for the natural temperature variability: $[\rho_{ij}(\lambda)]$ and $[R_{ij}(\lambda)]$, respectively ($\lambda = \Delta t/\tau$ is the lag in units of $\tau = 1$ month). In panel (a) we show the full cross-correlation matrix for lag $\lambda = 0$ including the 10512 grid points (73 latitudes × 144 longitudes) for the innovations (left) and for the temperature (right). The pixels were indexed as: 1→{90°S, 0°E}, 2→{90°S, 2.5°E}, …, 10511→{90°N, 5°W}, 10512→{90°N, 2.5°W}. These 10512×10512-pixel images are too big to appreciate the detailed structure of the teleconnections. The large autocorrelation values are visible only along the main diagonal, as well as in the top-left and bottom-right corners corresponding to the poles where the grid points are very close to each other and where they share the same temperature values. Large correlations are also observed in the tropical region for the temperature anomalies.

To discern the details of the cross-correlations, we blew-up the regions shown as black squares in panel (a). In panel (b), we show the lagged cross-correlation matrices for the innovations for $\lambda = 0$, 1, 2 and 3 months (left to right), only for these 576 grid points between latitudes 42.5°N and 52.5°N. In panel (c), we show similar cross-correlation matrices as in (b), but now for the natural temperature variability. In the figure captions in panels (b) and (c), we show the values of the average cross-correlation $\pm$ one standard deviation and the maximum absolute value for each matrix (i.e. out of more than $3 \cdot 10^5$ values).

For $\lambda = 0$, the elements $\rho_{ij}(0) = a_{ij}$ shown in Fig. S2(b), are relatively large. The maximum values are evidently 1 along the main diagonal, but very large values are also obtained along the diagonals separated by 144 pixels because they represent places only 2.5° away in latitude. For the temperature (panel (c)), the correlations decrease with the lag, but some of the structure is preserved and relatively large values are obtained even for $\lambda = 3$ months, mainly along the diagonals. The temperature cross-correlation, $R_{ij}(\lambda)$ is proportional to a_{ij}, but it also depends on the fluctuation exponents H_i and H_j for every λ, following Eqs. (7) and (8). For the cross-correlation of the innovations (panel (b)), the values decrease much faster. Even for $\lambda = 1$, we can see that almost all the correlation is lost (see the distribution values in the figure captions). This indicates that the innovation series closely satisfy the time-independence condition given by Eq. (S18) (actually, its discrete version where we replace the Dirac by the Kronecker delta). The same analysis exemplified here for the small sample square region of 576×576 pixels,

was performed in the full 10512×10512 correlation matrices obtaining similar distributions for each respective lag.

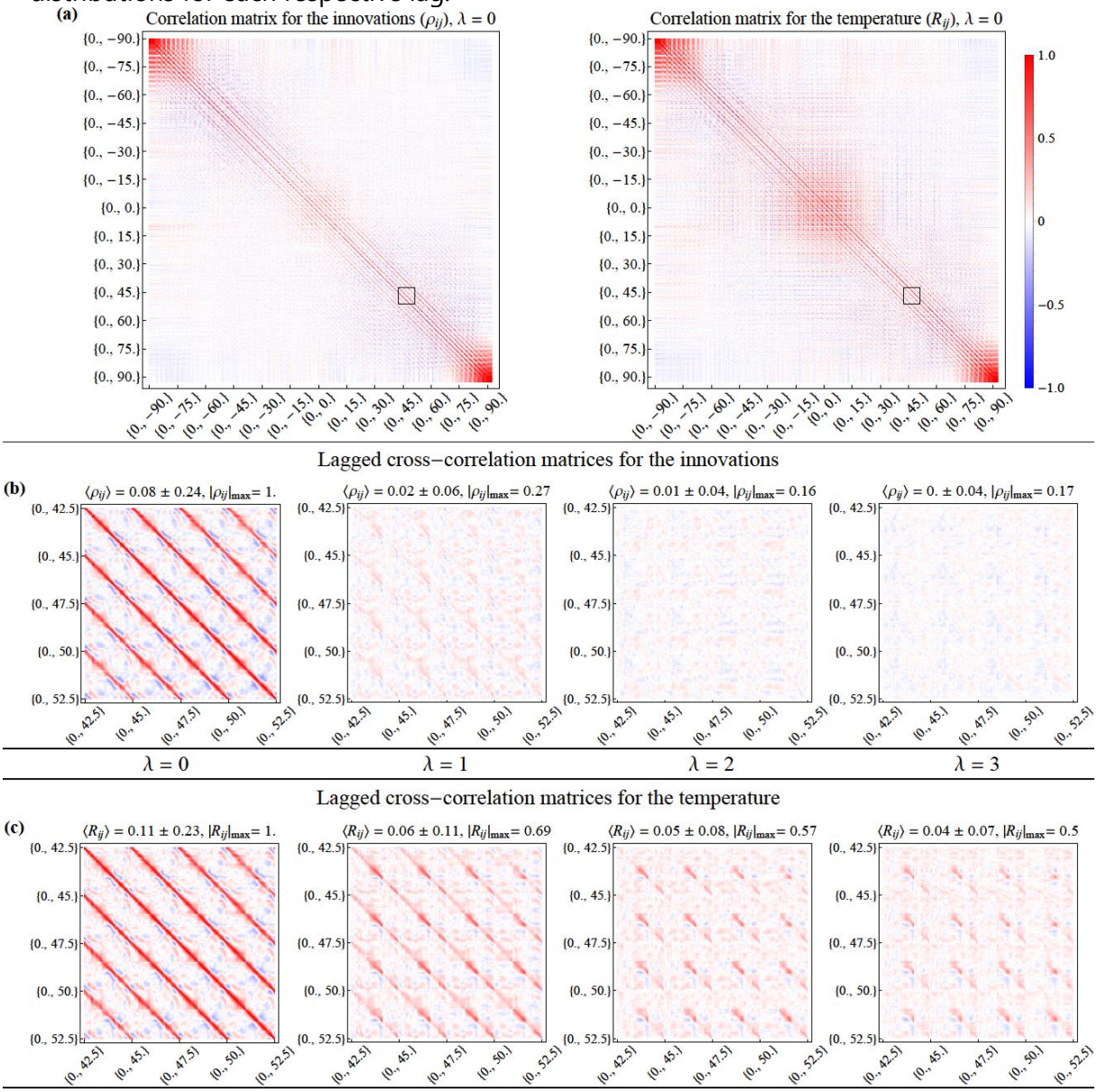

Figure S2. Lagged cross-correlation matrices for the innovations and for the natural temperature variability: $[\rho_{ij}(\lambda)]$ and $[R_{ij}(\lambda)]$, respectively. (a) Full cross-correlation matrices for lag $\lambda = 0$ including the 10512 grid points for the innovations (left) and for the temperature (right). (b) Lagged cross-correlation matrices for the innovations for $\lambda =$ 0, 1, 2 and 3 (left to right), only for the 576 grid points between latitudes 42.5°N and 52.5°N. They correspond to the small square region shown in the respective matrix in panel (a). (c) Same as in (b), but now for the natural temperature variability. In the figure captions in panels (a) and (b), we show the values of the average cross-correlation ± one standard deviation and the maximum absolute value for each matrix.

Text S6. Co-predictors

In Text S3 we mentioned that, given m datapoints, the *k*-step predictor for the temperature at location "*i*", with $T_i(t)$ an fGn process with parameters σ_i and H_i, is in vector form:

$$\hat{T}_i^0(k) = \boldsymbol{\phi}_i^0(k) \cdot \mathbf{T}_i, \tag{S27}$$

where $\mathbf{T}_i = [T_i(-m), \dots, T_i(0)]^T$ and the vector of coefficients $\boldsymbol{\phi}_i^0(k) = \left[\phi_{-m}^{i0}(k), \dots, \phi_0^{i0}(k)\right]^T$, satisfies the Yule-Walker equations (Eq. (S10)). The superscript "0" indicates that only the location "*i*" is considered.

Let us assume that we have another time series at location "*j*", $T_j(t)$ (fGn process with parameters σ_j and H_j) and we want to add this information to improve the predictor for the temperature at location "*i*". The optimal *k*-step predictor is now given by the sum of $2m + 2$ terms ($m + 1$ in the scalar product for each location):

$$\hat{T}_i(k) = \boldsymbol{\phi}_i(k) \cdot \mathbf{T}_i + \frac{\sigma_{Ti}}{\sigma_{Tj}} \boldsymbol{\phi}_j(k) \cdot \mathbf{T}_j, \tag{S28}$$

where the vectors of coefficients, $\boldsymbol{\phi_i}(k)$ and $\boldsymbol{\phi_j}(k)$ satisfy the Yule-Walker equations:

$$\begin{pmatrix} \mathbf{R}_{ii} & \mathbf{R}_{ij} \\ \mathbf{R}_{ji} & \mathbf{R}_{jj} \end{pmatrix} \begin{pmatrix} \boldsymbol{\phi}_i(k) \\ \boldsymbol{\phi}_j(k) \end{pmatrix} = \begin{pmatrix} \mathbf{r}_{ii}(k) \\ \mathbf{r}_{ij}(k) \end{pmatrix}. \tag{S29}$$

The matrices $\mathbf{R}_{ii} = [R_{ii}(1, t_1 - t_2)]_{t_1,t_2=-m,\dots,0}$ and $\mathbf{R}_{jj} = \left[R_{jj}(1, t_1 - t_2)\right]_{t_1,t_2=-m,\dots,0}$ are the autocorrelation matrices for processes "*i*" and "*j*", respectively, $\mathbf{R}_{ij} = \mathbf{R}_{ji}{}^T = \left[R_{ij}(a_{ij}, t_1 - t_2)\right]_{t_1,t_2=-m,\dots,0}$ are the cross-correlation matrices and the vectors $\mathbf{r}_{ii}(k) = [R_{ii}(1, k - s)\,]_{s=-m,\dots,0}^T$ and $\mathbf{r}_{ij}(k) = \left[R_{ij}(a_{ij}, k - \mathrm{s})\,\right]_{s=-m,\dots,0}^T$, (the elements are obtained from Eq. (S22)).

For the case where we only have one time series at location "*i*", Eq. (S15) gives:

$$\mathrm{MSSS}_i^0(k, m, H_i) = \mathbf{r}_{ii}(k)^T \left(\mathbf{R}_{ii}\right)^{-1} \mathbf{r}_{ii}(k) = \boldsymbol{\phi}_i^0(k) \cdot \mathbf{r}_{ii}(k). \tag{S30}$$

The MSSS for the forecast at location "*i*", considering now the information from the two locations "*i*" and "*j*" is:

$$\mathrm{MSSS}_i(k, m, a_{ij}, H_i, H_j) = \boldsymbol{\phi}_i(k) \cdot \mathbf{r}_{ii}(k) + \boldsymbol{\phi}_j(k) \cdot \mathbf{r}_{ij}(k). \tag{S31}$$

The skill score for horizon k is a function of the memory, m, the intrinsic spatial correlation of the innovations, a_{ij}, (independent of the scaling) and the fluctuation exponents, H_i and H_j.

The main question is how much the new location helps to improve the accuracy of the forecast at position "*i*". This can be quantified by the difference $\Delta\mathrm{MSSS}_i = \mathrm{MSSS}_i - \mathrm{MSSS}_i^0$:

$$\Delta\mathrm{MSSS}_i\left(k,m,a_{ij},H_i,H_j\right)=\left[\boldsymbol{\phi}_i\left(k\right)-\boldsymbol{\phi}_i^0\left(k\right)\right]\cdot\mathbf{r}_{ii}\left(k\right)+\boldsymbol{\phi}_j\left(k\right)\cdot\mathbf{r}_{ij}\left(k\right) \tag{S32}$$

which can also be written as:

$$\Delta\mathrm{MSSS}_i\left(k,m,a_{ij},H_i,H_j\right)=\boldsymbol{\phi}_j\left(k\right)\cdot\left[\mathbf{r}_{ij}\left(k\right)-\mathbf{R}_{ij}{}^T\boldsymbol{\phi}_i^0\left(k\right)\right]. \tag{S33}$$

This is the normalized projection of the predictor from the new location,

$$\hat{T}_j\left(k\right)=\boldsymbol{\phi}_j\left(k\right)\cdot\mathbf{T}_j, \tag{S34}$$

in the direction of the error from the first predictor only:

$$\Delta\mathrm{MSSS}_i\left(k\right)=\left\langle\hat{T}_j\left(k\right)\left[T_i\left(k\right)-\hat{T}_i^0\left(k\right)\right]\right\rangle. \tag{S35}$$

This is in agreement with the orthogonality principle that states that the error of the predictor at location "i" is orthogonal to the data used to build that predictor. Location "j" can only contribute with new information that gives some component along this orthogonal direction.

For continuous-in-time infinitely long series (infinite memory), the predictor $\hat{T}_j(k)$ for $k>0$ is a linear combination of past innovations, $\{\gamma_j(t)\}_{t<0}$, while the error, $T_i(k)-\hat{T}_i^0(k)$, is a linear combination of future innovations $\{\gamma_i(t)\}_{t>0}$ (see the development in Sect. 2.5 of (Lovejoy et al., 2015)). That means that, if the condition Eq. (S18) is satisfied (future and past innovations are independent), then $\Delta\mathrm{MSSS}_i=0$ in Eq. (S35) and the new location does not help to improve the forecast. That does not mean that the two series at location "i" and "j" are independent; they are still correlated with the correlation function given by Eq. (S22). It is just that this correlation is already included in the information obtained from the past at location "i", which is enough for obtaining the optimal prediction for that location.

For discrete-in-time finite series, there is some improvement in the prediction from using a co-predictor, but this improvement decreases with the memory, m, and is very small if enough past data points are used to build the predictor. In Fig. S3 we show contour plots of the relative difference $\Delta\mathrm{MSSS}_i/\mathrm{MSSS}_i^0$ (in %) as a function of H_i and H_j for $k=1$, $m=50$ and values of $a_{ij}=0.6$, 0.7 0.8 and 0.9. Notice that for a wide range of values, when H_i and H_j are relatively close (dark blue region in the plots), the second location brings almost no new information to the forecasts. The relative gain increases with a_{ij}, but even for $a_{ij}=0.6$ (which is a fairly high correlation already) it remains lower than 1% for all values of H_i and H_j (see top-left panel of Fig. S3). That is why we did not include plots for $a_{ij}<0.6$. In fact, even for highly correlated locations at the level of innovations with $a_{ij}=0.9$ (bottom-right panel), the maximum relative improvement is lower than 4%. This maximum improvement is obtained when $H_i\approx-0.5$, for which the original MSSS_i^0 is very low, so the difference is actually at the noise level and is not statistically significant. This means that, in practice, for any set of fluctuation exponents H_i and H_j and values of a_{ij} as large as 0.9, we only gain less than 2% of the original MSSS by using a co-predictor.

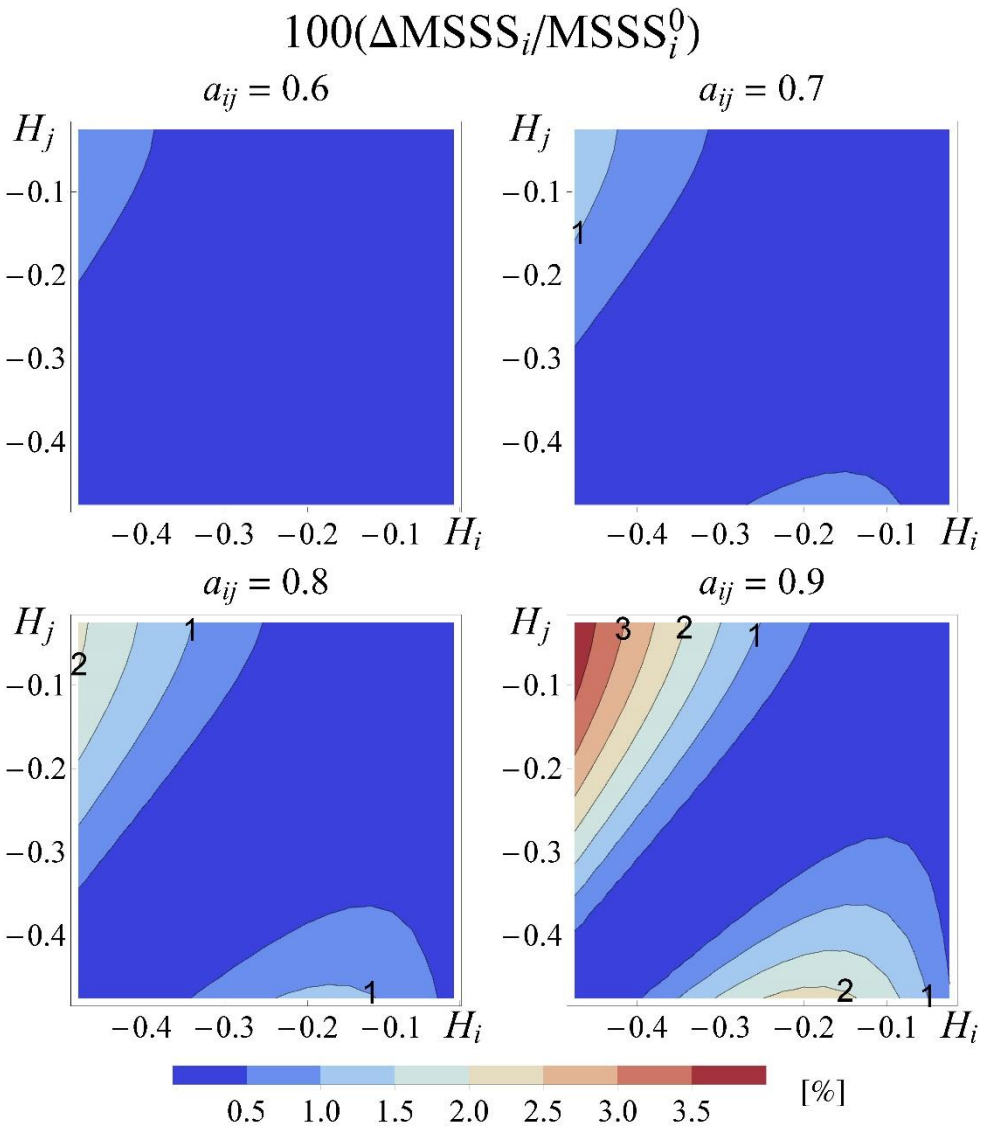

Figure S3. Contour plots of the relative improvement $\Delta\mathrm{MSSS}_i/\mathrm{MSSS}_i^0$ (in %) as a function of H_i and H_j for one step ($k = 1$), with 50 past values ($m = 50$) and correlations $a_{ij} = 0.6$, 0.7 0.8 and 0.9.

If MSSS_i^0 is the score obtained by predicting the temperature series at location i independently and MSSS_i is the score considering also the information from another location j, then the improvement $\Delta\mathrm{MSSS}_i = \mathrm{MSSS}_i - \mathrm{MSSS}_i^0$ is, by definition, a measure of the Granger causality between series i and j. It can be proven that $\Delta\mathrm{MSSS}_i$ is the projection of the predictor at time t from the new location, $\hat{T}_j(t)$, in the direction of the error from the first predictor only:

$$\Delta\mathrm{MSSS}_i(t) = \left\langle \hat{T}_j(t)\left[T_i(t) - \hat{T}_i^0(t)\right]\right\rangle. \tag{S36}$$

The orthogonality condition Eq. 12 implies that, for infinite series, if the new predictor is a linear combination of past data, then $\Delta\mathrm{MSSS}_i = 0$. If only a few memory

steps are used, then larger improvements are obtained by borrowing memory from co-predictors. Figure S4 shows the relative improvement $\Delta\text{MSSS}_i/\text{MSSS}_i^0$ (in %) as a function of H_i and H_j for $k = 1$, $a_{ij} = 0.8$ and values of $m = 20$, 5, 3 and 1. As the number of autoregressive steps used (memory) decreases, the larger the relative improvement becomes from using a co-predictor. In all cases, the long-memory predictor skill from a single location, MSSS_i^{∞} (with $m = 50$), is larger than the combined short memory one for $m = 5$, 3, 2 and 1 (see Fig. S5).

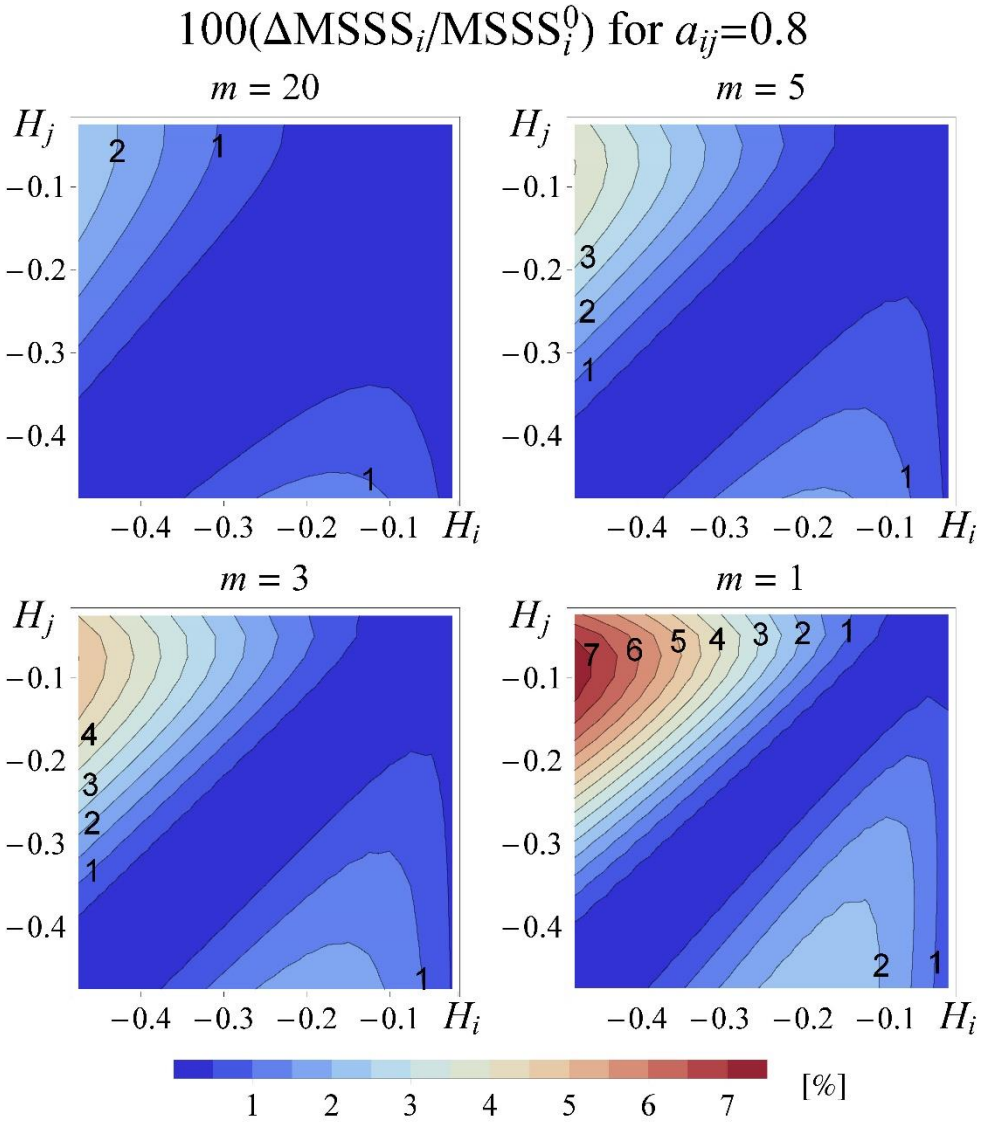

Figure S4. Contour plots of the relative improvement $\Delta\text{MSSS}_i/\text{MSSS}_i^0$ (in %) as a function of H_i and H_j for one step ($k = 1$), with $a_{ij} = 0.8$ and $m = 20$, 5, 3 and 1.

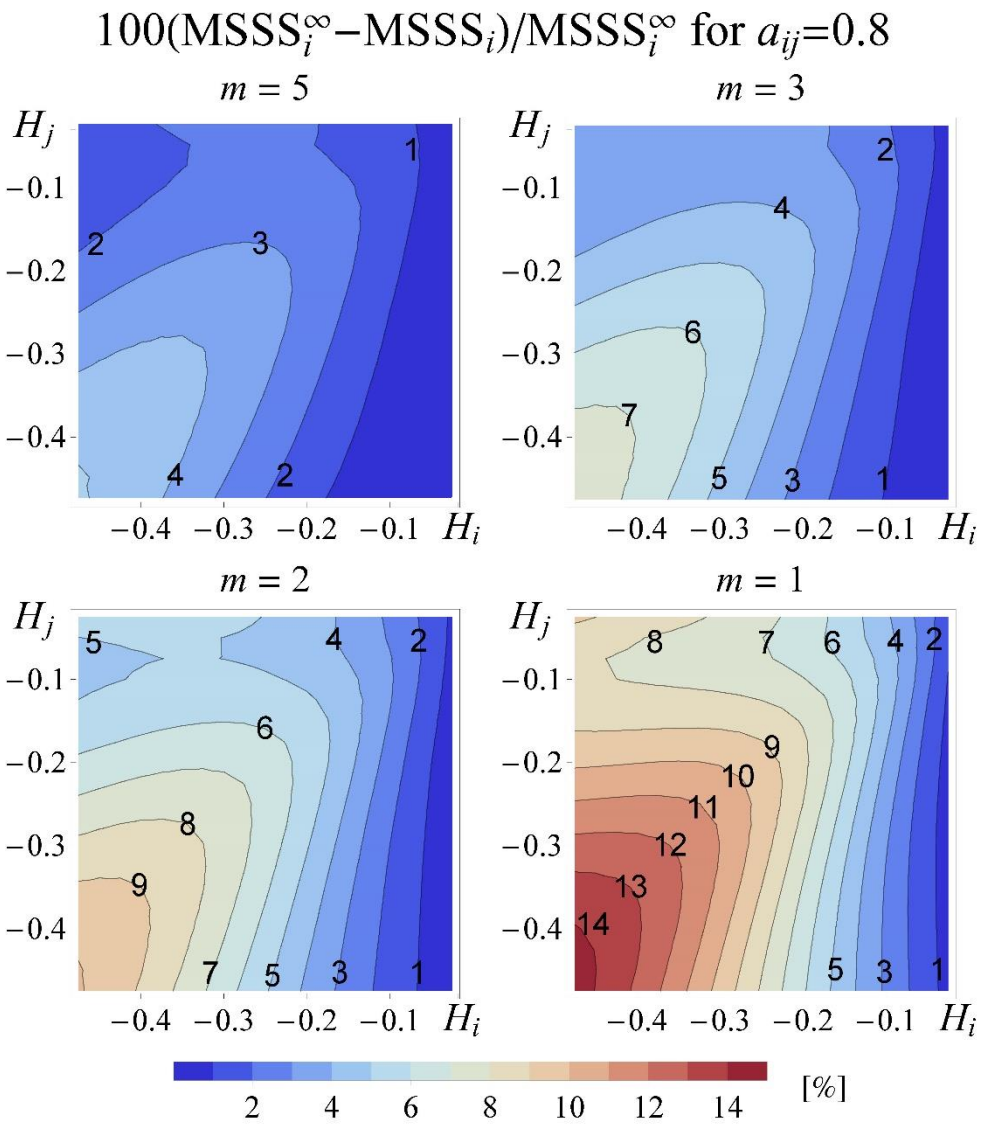

Figure S5. Contour plots of the relative difference $(\mathrm{MSSS}_i^\infty - \mathrm{MSSS}_i)/\mathrm{MSSS}_i^\infty$ as a function of H_i and H_j for one step ($k = 1$), with $a_{ij} = 0.8$. We used $m = 50$ for the single location long-memory MSSS_i^∞ and $m = 5$, 3, 2 and 1 for the combined skill MSSS_i.

Following Eq. (S33), we computed a map of the maximum relative improvement $\Delta\mathrm{MSSS}_i/\mathrm{MSSS}_i^0$ (Fig. S6) based on the empirical parameters. We see that the contribution from any other location is very small, reaching a maximum of 2% only in a few places, which is in the noise level of the skill estimates. This empirically confirms the lack of Granger causality between the series. As we show in Fig. S3, the maximum improvement at location i comes from a place j with large correlation, a_{ij}, and fluctuation exponent H_j very different than H_i. This gives the largest component of the co-predictor orthogonal to the m-dimensional space defined by the m past values of temperature at location i used to build the predictor. To produce the forecasts used in Fig. S6, we used $m = 20$. The relative contribution from co-predictors can be decreased by just increasing the

number m of past temperature values used to build the original predictor, in the limit $m \to \infty$, the relative improvement vanishes.

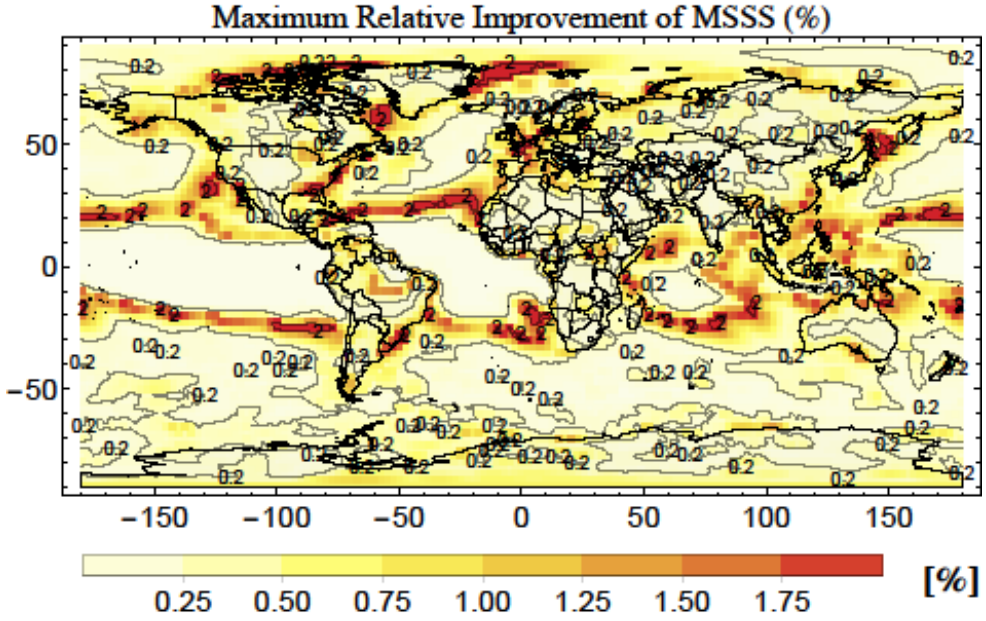

Figure S6. Maximum relative improvement, $\Delta\mathrm{MSSS}_i/\mathrm{MSSS}_i^0$, that could be obtained from using any other location as co-predictor, using $m = 20$ past values for the forecasts.

Text S7. Simulations

For most of the globe, the natural temperature variability has a transition from the weather (characterized by fluctuations increasing with the time scale) to the macroweather (with decreasing fluctuations) at a transition time τ_w lower than one month. This means that, for monthly averaged temperature, only the macroweather regime is present with the corresponding $H < 0$ (see Fig. 1(a)) and the temperature can be modeled and simulated using the theory presented in sections S1 and S2.

As presented in (Del Rio Amador & Lovejoy, 2020), for the places shown in red and yellow in Fig. 1(b) (generally over the tropical ocean), the weather-macroweather transition occurs at τ_w generally between 1 and 2 years, corresponding to the longer predictability limit of the ocean. For monthly averages, there is a biscaling behaviour: for time scales lower than τ_w, $H > 0$, the temperature can be modelled as a fractional Brownian motion (fBm) process (first differences are fGn) and for time scales larger than τ_w, $H < 0$, the anomalies behave as fGn. Lovejoy (2019, 2020; Lovejoy et al., 2020) shows how this biscaling process – called fractional relaxation noise (fRn) – emerges naturally as the solution to a fractional energy balance equation (FEBE), and τ_w can be identified as a characteristic relaxation time.

In this paper, we follow a more pragmatic approach to reproduce the two scaling regimes by still using the theory for fGn processes. This is achieved by expressing the natural temperature variability as a smoothed fGn process with a window τ_w:

$$T_{\tau_w}(t) = \int_{t-\tau_w}^{t} T_\tau(t')\,dt', \tag{S37}$$

where $T_\tau(t)$ is the resolution τ fGn defined by Eq. (S1), which is the τ increments of the corresponding fBm process, $B_{H'}(t)$, with Hurst exponent $H' = H_{\mathrm{fBm}} > 0$ (the fluctuation exponent for the fGn is $H_{\mathrm{fGn}} = H' - 1 < 0$ as $0 < H' < 1$):

$$T_\tau(t) = \frac{1}{\tau}\left[B_{H'}(t) - B_{H'}(t-\tau)\right]. \tag{S38}$$

For $\tau \ll \tau_w$ (or exactly in the discrete case where we make $\tau = 1$ and we replace the integral by the corresponding sum in Eq. (S37)), it can be shown that $T_{\tau_w}(t)$ is the τ_w increment of the same fBm process:

$$T_{\tau_w}(t) = B_{H'}(t) - B_{H'}(t-\tau_w). \tag{S39}$$

Therefore, it is stationary with variance $\langle T_{\tau_w}(t)^2\rangle = {\sigma_T}^2 {\tau_w}^{2H}$ and correlation function:

$$R_{H,\tau_w}(\Delta t) = \frac{1}{2}\left(\left|\frac{\Delta t}{\tau_w}+1\right|^{2H} + \left|\frac{\Delta t}{\tau_w}-1\right|^{2H} - 2\left|\frac{\Delta t}{\tau_w}\right|^{2H}\right), \tag{S40}$$

with H being the fluctuation exponent for the high frequencies (i.e. $0 < H < 1$). The only difference with Eq. (S3) is that now we do not have the restriction $|\Delta t| \geq \tau_w$, thus the new process has two scaling regimes. For $\Delta t \ll \tau_w$ it can be approximated by:

$$R_{H,\tau_w}(\Delta t) \approx 1 - \left(\frac{\Delta t}{\tau_w}\right)^{2H}, \tag{S41}$$

while for $\Delta t \gg \tau_w$ it follows a power-law equivalent to Eq. (S4):

$$R_{H,\tau_w}(\Delta t) \approx H(2H-1)\left(\frac{\Delta t}{\tau_w}\right)^{2H-2}. \tag{S42}$$

Similar asymptotic behaviours are obtained for the high and the low frequency approximations of the fRn process, solution of the FEBE. Also, notice how the cross-correlations (Eqs. (7) and (8)) satisfy similar equations if we take $2H = H_i + H_j$.

Figure S7 shows the dependence of the autocorrelation function with the lag $\lambda = \Delta t/\tau_w$ (Eq. (S40)) for different values of H. The high and the low frequency approximations for $H = 0.25$, given by Eqs. (S41) and (S42), were included as dashed and dotted lines, respectively.

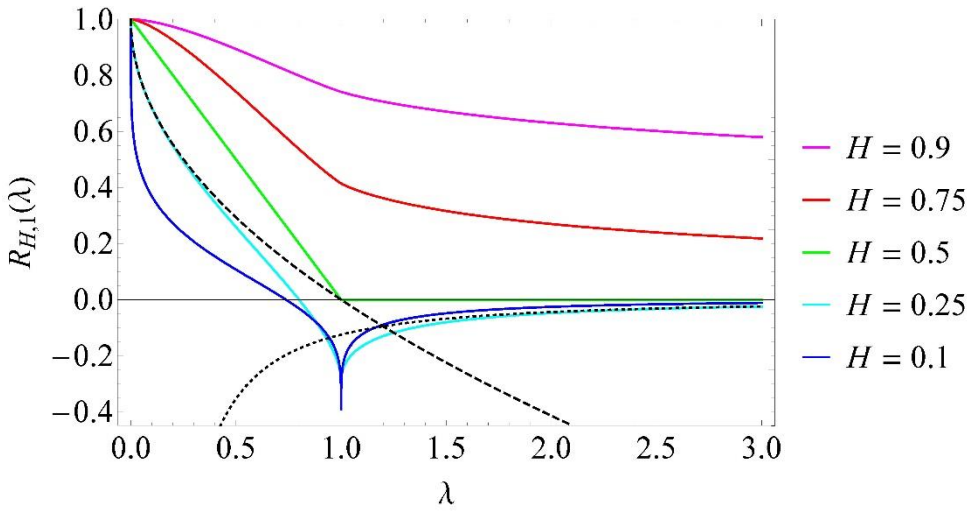

Figure S7. Autocorrelation function vs. $\lambda = \Delta t/\tau_w$ (Eq. (S40)) for different values of H. The high and the low frequency approximations for $H = 0.25$, given by Eqs. (S41) and (S42), were included as dashed and dotted lines, respectively.

To estimate the transition time, τ_w, in the places where $\tau_w > 1$ month ($H > 0$ in Fig. 1(a)), we can fit the theoretical autocorrelation function (ACF) (Eq. (S40)) to the empirical one obtained from the data. Figure S8 shows examples of the ACF for both cases: pure fGn with $H < 0$ and biscaling process with transition at τ_w from the weather, $H > 0$ (fBm-like) to the macroweather, $H < 0$ (fGn-like). The ACF's for the reference dataset (marked as "+"), for one set of simulations (marked as "o") and calculated using Eqs. (S3) or (S40) with the corresponding values of H and τ_w (solid curve) are shown in Fig. S8 for: (a) grid point in the North Atlantic at (55°N, 22.5°W), with estimated $H = -0.2$; (b) same as in (a) but with logarithmic scales in both axes to highlight the scaling; (c) grid point in the Tropical Pacific at (5°S, 177.5°W), with estimated high frequency $H = 0.38$ and $\tau_w = 16$ months and (d) the monthly mean temperature for the Niño 3.4 region (5°N-5°S, 170°W-120°W), with average fluctuation exponent $\langle H \rangle = 0.54$ and $\langle \tau_w \rangle = 10$ months. There is good agreement between the empirical, the simulated and the theoretical ACF's in all cases.

The empirical reproduction of the ACF validates the realism of the stochastic model on a pixel-by-pixel basis, but we must also reproduce the coupling between different pixels, i.e.: the cross-correlation structure. This is verified by comparing the empirical orthogonal functions (EOF's) obtained from the decomposition of the cross-correlation matrix. In Fig. S9 we show the first five EOF's for the reference dataset (left) and for a single realization simulations (right). To avoid strong multiplicative seasonality effects in the higher latitudes, we only considered the anomalies between 60°S and 60°N. In Fig. S10 we show the EOF's using only sea surface temperatures (SST's). The stochastic simulations reproduce very well the main modes of variability of the reference dataset. In the main test, we show how the cross-correlation structure for non-zero lags is also well reproduced by comparing the empirical and simulated ratio of global influence (RGI) – equivalent to the area weighted connectivity (AWC) for zero lags.

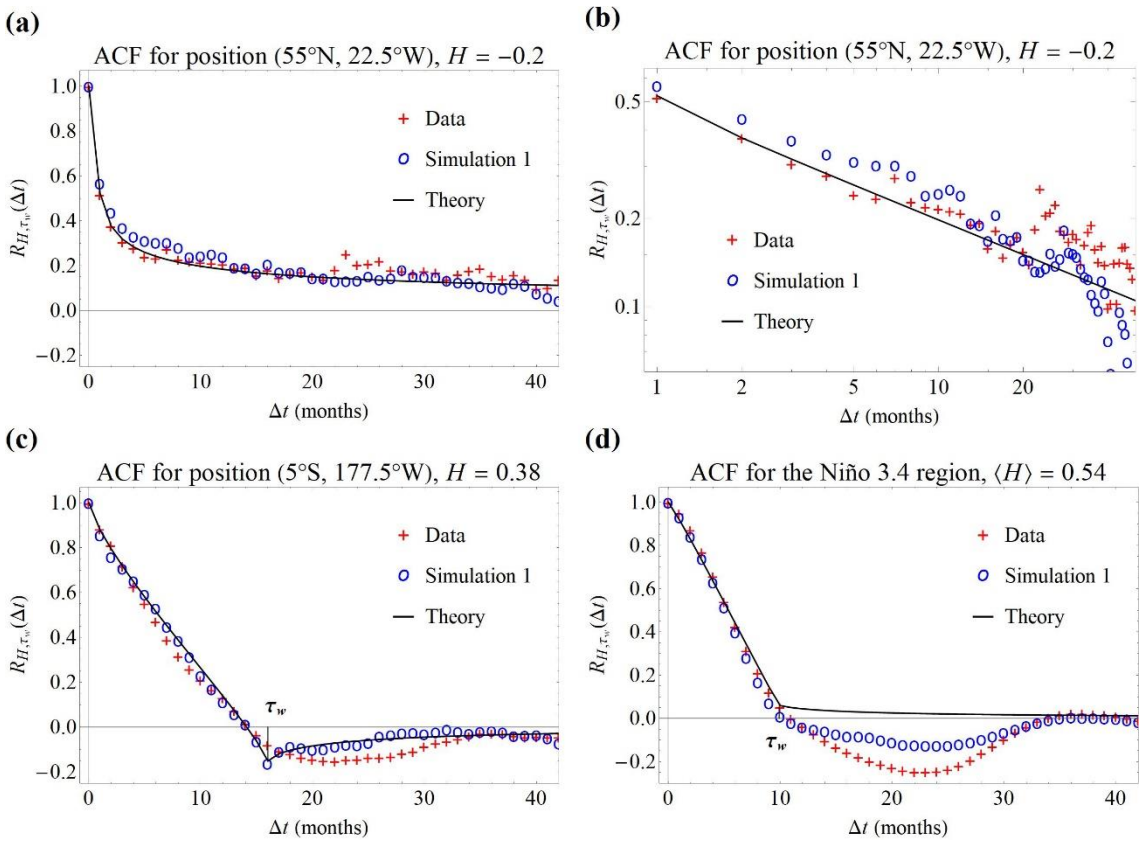

Figure S8. Autocorrelation functions (ACF's) for the reference dataset (marked as "+"), for one set of simulations (marked as "o") and obtained theoretically using Eqs. (S3) or (S40) (solid curve) for the temperature series at different positions: (a) grid point in the North Atlantic at (55°N, 22.5°W), with estimated $H = -0.2$; (b) same as in (a) but with logarithmic scale in both axes to highlight the scaling; (c) grid point in the Tropical Pacific at (5°S, 177.5°W), with $H = 0.38$ and $\tau_w = 16$ months and (d) the monthly mean temperature for the Niño 3.4 region (5°N-5°S, 170°W-120°W), with average $\langle H \rangle = 0.54$ and $\langle \tau_w \rangle = 10$ months.

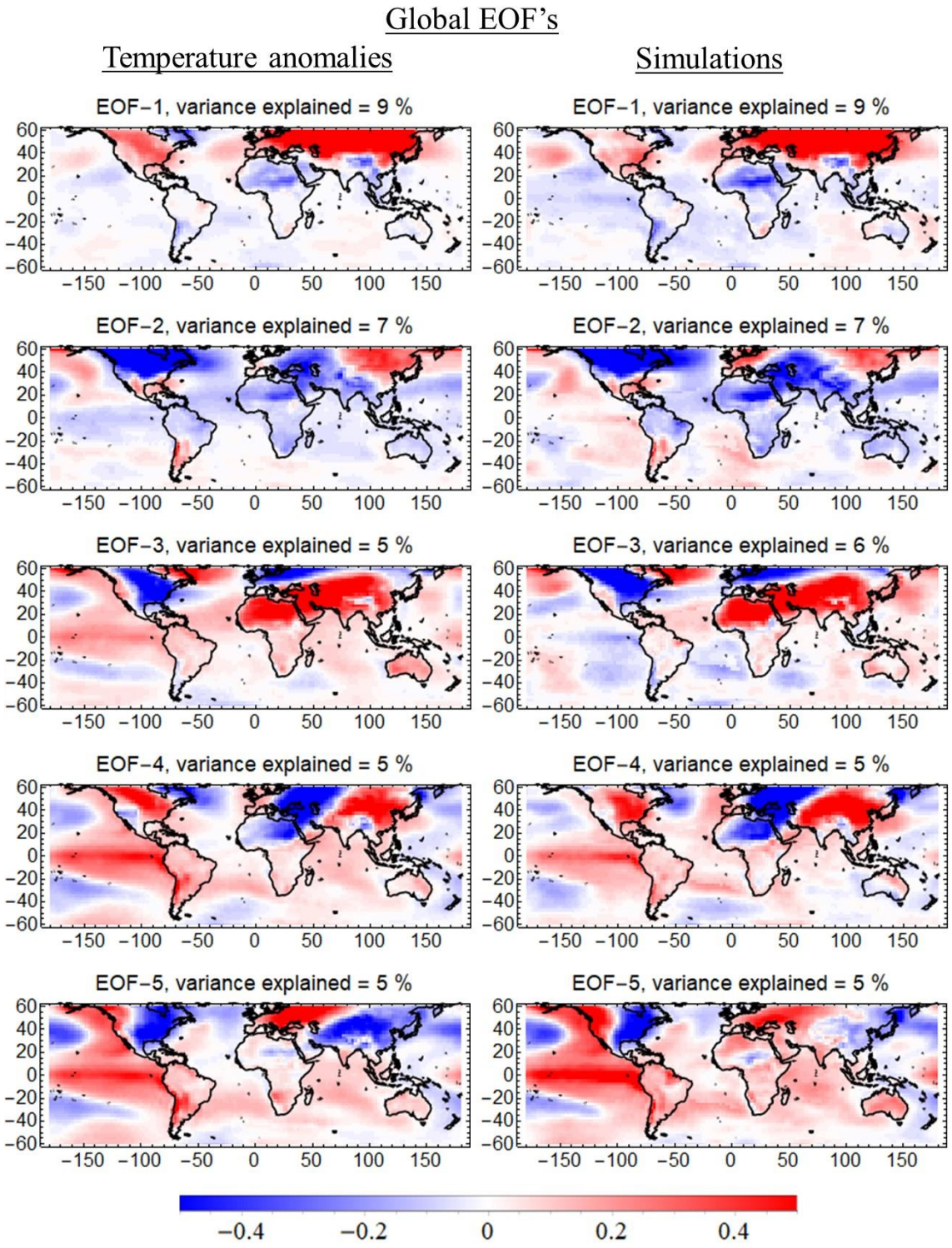

Figure S9. First five EOF's for the reference dataset (left) and for one simulation (right). We only considered the anomalies between 60°S and 60°N to avoid strong multiplicative seasonality effects in the higher latitudes.

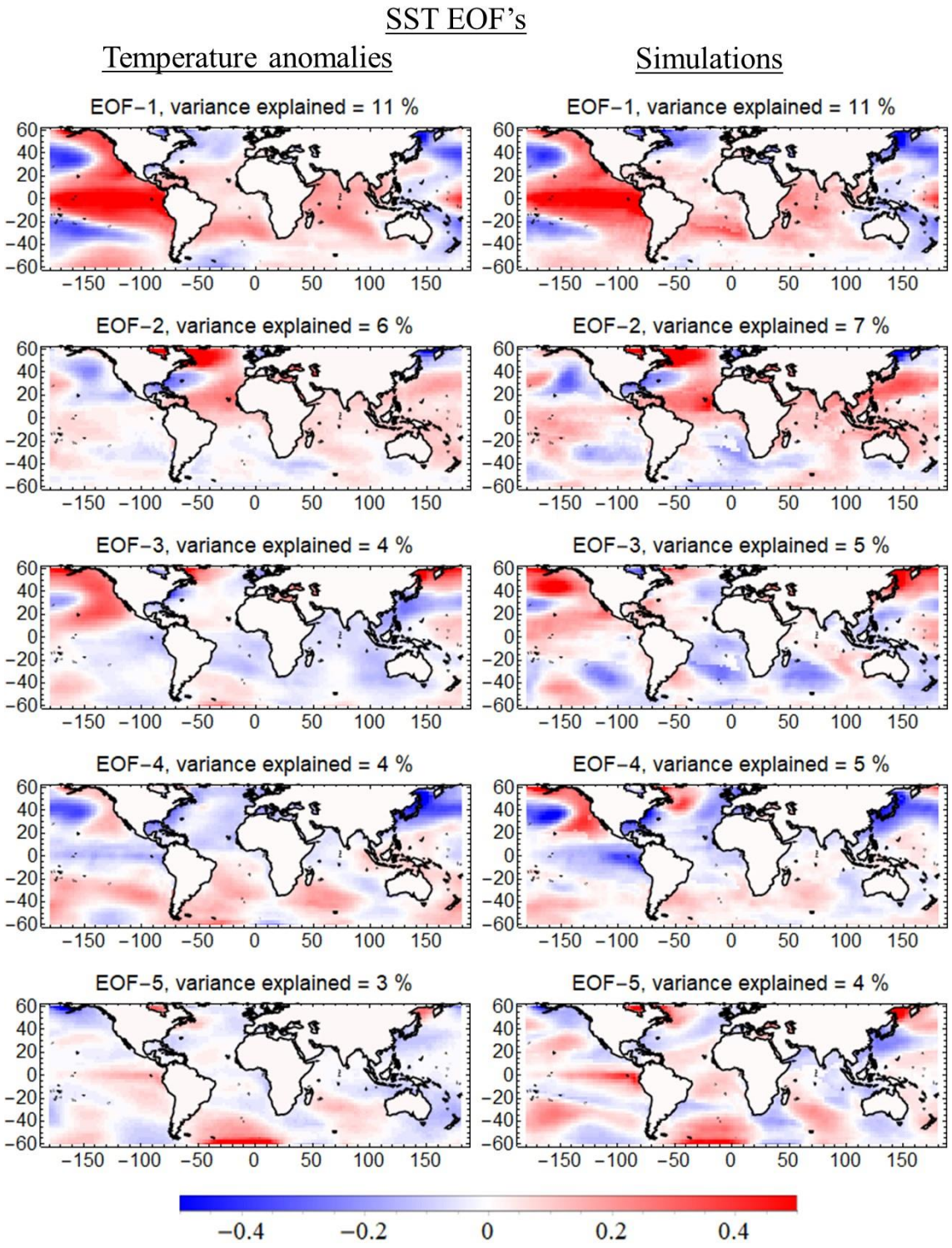

Figure S10. First five EOF's for the reference dataset (left) and for one simulation (right) only considering sea surface temperature anomalies (SST) between 60°S and 60°N.

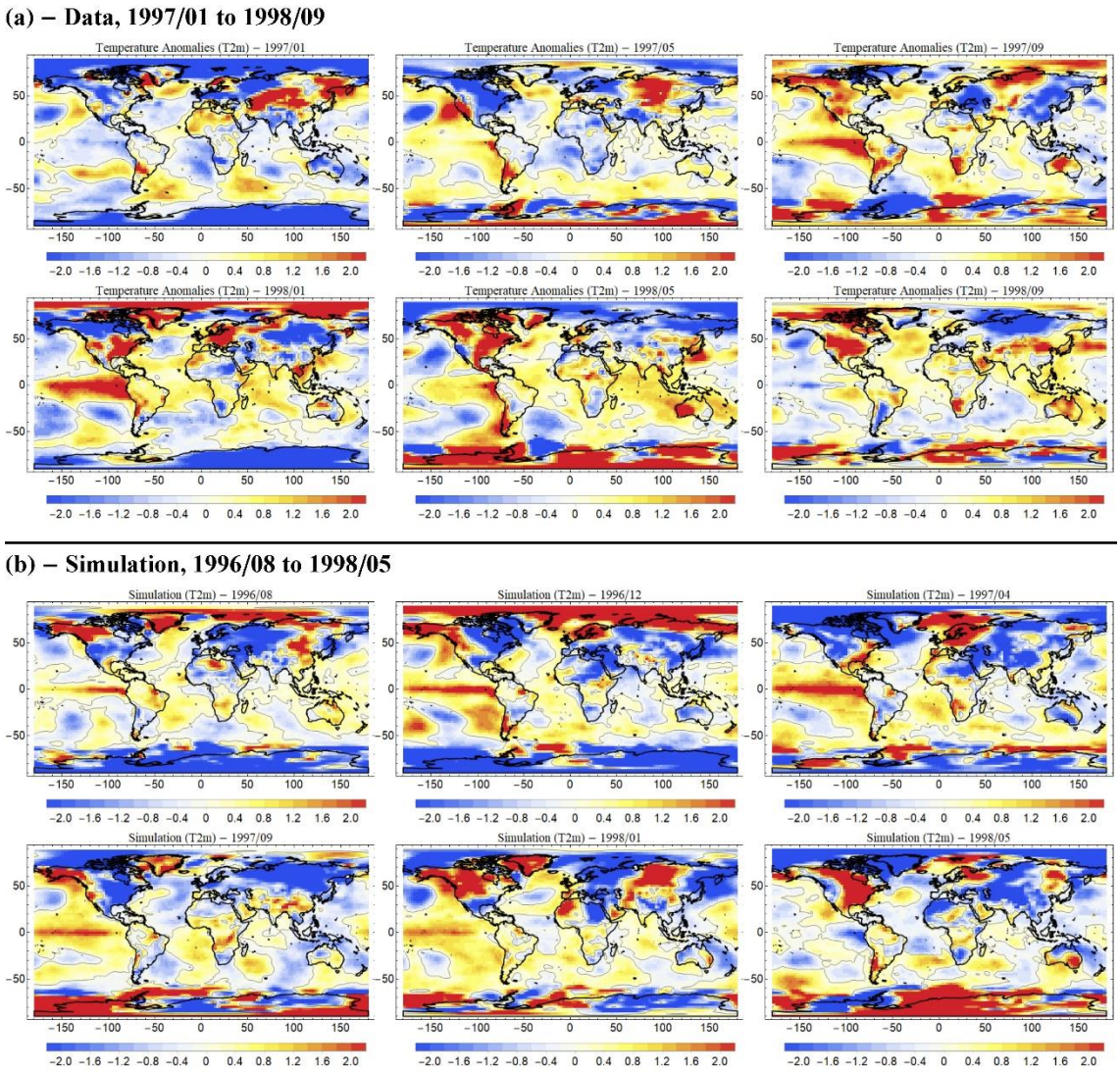

Figure S11. Sequence showing the evolution of one El Niño event (see the central and east-central equatorial Pacific) since January 1997 to September 1998 for the reference dataset (a) and from August 1996 to May 1998 for the simulations (b).

Additional Supporting Information (Files uploaded separately)

The evolution of the detrended temperature anomalies obtained from NCEP/NCAR Reanalysis 1 and one simulation for the period January 1948 – December 2019 is shown in Movie S1. It is important to realize that the simulation (right) is not supposed to be the same as the reality (left), it is only supposed to have the same type of variability (the date above the simulation is totally fictional). Among other realistic features, the simulation reproduces huge regional patterns including El Niño and La Niña events that are emergent properties of the model. One example of the evolution of the El Niño event is

shown in Fig. S11 with a sequence since January 1997 to September 1998 for the reference dataset (a) and from August 1996 to May 1998 for the simulations (b).

Movie S1. Evolution of the detrended temperature anomalies obtained from NCEP/NCAR Reanalysis 1 and one simulation for the period January 1948 – December 2019. It is important to realize that the simulation (right) is not supposed to be the same as the reality (left), it is only supposed to have the same type of variability (the date above the simulation is totally fictional). Among other realistic features, the simulation reproduces huge regional patterns including El Niño and La Niña events that are emergent properties of the model.