Climate Dynamics (2019) 53:4373–4411
https://doi.org/10.1007/s00382-019-04791-4

Predicting the global temperature with the Stochastic Seasonal to Interannual Prediction System (StocSIPS)

Lenin Del Rio Amador[1] · Shaun Lovejoy[1]

Received: 25 July 2018 / Accepted: 25 April 2019 / Published online: 22 May 2019

Abstract

Many atmospheric fields—in particular the temperature—respect statistical symmetries that characterize the macroweather regime, i.e. time-scales between the ≈ 10 day lifetime of planetary sized structures and the (currently) 10–20 year scale at which the anthropogenic forcings begin to dominate the natural variability. The scale-invariance and the low intermittency of the fluctuations implies the existence of a huge memory in the system that can be exploited for macroweather forecasts using well-established (Gaussian) techniques. The Stochastic Seasonal to Interannual Prediction System (StocSIPS) is a stochastic model that exploits these symmetries to perform long-term forecasts. StocSIPS includes the previous ScaLIng Macroweather Model (SLIMM) as a core model for the prediction of the natural variability component of the temperature field. Here we present the theory for improving SLIMM using discrete-in-time fractional Gaussian noise processes to obtain an optimal predictor as a linear combination of past data. We apply StocSIPS to the prediction of globally-averaged temperature and confirm the applicability of the model with statistical testing of the hypothesis and a good agreement between the hindcast skill scores and the theoretical predictions. Finally, we compare StocSIPS with the Canadian Seasonal to Interannual Prediction System. From a forecast point of view, GCMs can be seen as an initial value problem for generating many "stochastic" realizations of the state of the atmosphere, while StocSIPS is effectively a past value problem that estimates the most probable future state from long series of past data. The results validate StocSIPS as a good alternative and a complementary approach to conventional numerical models. Temperature forecasts using StocSIPS are published on a regular basis in the website: http://www.physics.mcgill.ca/StocSIPS/.

1 Introduction

When taken beyond their deterministic predictability limits of about ten days, the output of General Circulation Models (GCMs) can no longer be usefully interpreted in a deterministic sense; they are at least implicitly stochastic and if they use stochastic parameterizations, they are explicitly so. In this "macroweather" regime, successive fluctuations tend to cancel each other out so that in control run mode, each GCM converges ultra slowly (Lovejoy et al. 2013) to its own climate. Assuming ergodicity, the control run climate is deterministic because it is the long-time average climate state, but the fluctuations about this state are stochastic.

Although each GCM climate may be different—and different from that of the real world—various studies (see e.g. the review (Lovejoy et al. 2018)) have indicated that the space–time statistics of fluctuations about the climates are statistically realistic—that they are of roughly the same type as the fluctuations observed in the real climate about the real climate state. For example, over wide ranges, and with realistic exponents, they exhibit scaling in both space and in time and at least approximately, they obey a symmetry called "statistical space–time factorization" (Lovejoy and de Lima 2015) that relates space and time. This suggests that the main defect of GCMs is that their fluctuations are around unrealistic model climates.

Many different stochastic processes can yield identical statistics. This leads to the possibility—developed in (Lovejoy et al. 2015)—that a simple model, having the same space–time statistical symmetries as the GCMs and the real world, could be used to directly model temperature fluctuations. If in such a model, the long term behaviour and the statistics of the fluctuations are forced to match that of real-world data in the past, the model would thus combine realistic fluctuations with a realistic climate, leading to

✉ Lenin Del Rio Amador
delrio@physics.mcgill.ca

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

significantly improved forecasts. Indeed, using this ScaLIng Macroweather Model (SLIMM), (Lovejoy 2015) gave some evidence for this by accurately forecasting the slow-down in the warming after 1998.

Starting with (Hasselmann 1976), various stochastic macroweather and climate models have been proposed. Today, these approaches are generally known under the rubric Linear Inverse Modelling (LIM), e.g.: (Penland and Matrosova 1994; Penland and Sardeshmukh 1995; Winkler et al. 2001; Newman et al. 2003; Sardeshmukh and Sura 2009). However, they all are based on integer order (stochastic) differential equations and these implicitly assume the existence of characteristic time scales associated with exponential decorrelation times; such models are not compatible with the scaling. To obtain models that respect the scaling symmetry, we may use fractional differential equations that involve strong, long-range memories; it is these long-range memories that are exploited in SLIMM forecasts. From a mathematical point of view, the fractional differential operators are of Weyl type (convolutions from the infinite past) so that they are not initial value problems, but rather past value problems.

In this paper we present the new Stochastic Seasonal to Interannual Prediction System (StocSIPS), that includes SLIMM as the core model to forecast the natural variability component of the temperature field, but also represents a more general framework for modelling the seasonality and the anthropogenic trend and the possible inclusion of other atmospheric fields at different temporal and spatial resolutions. In this sense, StocSIPS is the general system and SLIMM is the main part of it dedicated to the modelling of the stationary scaling series. The original technique that was used to make the SLIMM forecasts was basically correct, but it made several approximations (such as that the amount of data available for the forecast was infinite) and it was numerically cumbersome. Here, for the developing of StocSIPS, we return to it using improved mathematical and numerical techniques and validate them on ten different global temperature series since 1880 (five globally-averaged temperature series and five land surface average temperature). We then compare hindcasts with Canada's operational long-range forecast system, the Canadian Seasonal to Interannual Prediction System (CanSIPS) and we show that StocSIPS is just as accurate for 1-month forecasts, but significantly more accurate for longer lead times.

2 Theoretical framework

2.1 SLIMM

Since the works of (Hasselmann 1976), there have been many stochastic climate theories based on the idea that the high-frequency weather drives the low-frequency climate as a stochastic forcing [for a review, see Franzke et al. (2014)]. The first and simplest approaches for solving the stochastic climate differential equations deduced from these theories were made through linear inverse models (LIM). The theoretical justification of LIM methods is based on extracting the intrinsic linear dynamics that govern the climatology of a complex system directly from observations of the system (inverse approach). However, they implicitly assume exponential decorrelations in time, whereas both the underlying Navier–Stokes equations (and hence models, GCMs) and empirical analyses respect statistical scaling symmetries [see the review in Lovejoy and Schertzer (2013)]. Due to this lack of solid physical basis, LIM approaches are referred to as "empirical approaches". Nevertheless, its use is justified as a simpler alternative to the difficult task of improving numerical model parameterizations by appealing to physical arguments and first-principle reasoning alone.

Exponential decorrelations assumed by LIM models imply a scale break in time and—ignoring the diurnal and annual cycles—the only strong scale break is at the weather-macroweather transition scale of $\tau_w \approx 5$–15 days (slightly varying according to location (especially latitude and land versus ocean), and also with slight variations from one atmospheric field to another. For the temperature, there is a transition in the spectrum at $\omega \sim \omega_w \approx 1/\tau_w$, with two different asymptotic behaviors for very high and very low frequencies [see Fig. 4 in Lovejoy and Schertzer (2012)]. Empirically we find that $E_T(\omega) \sim \omega^{-\beta}$ with, $\beta_h = 1.8\ (\omega > \omega_w)$ and $\beta_l \approx 0.2$–$0.8\left(\omega < \omega_w\right)$ (depending on the location). The integer order differential equation for the LIM model implies that $\beta_h = 2$ and $\beta_l = 0$ (exactly, everywhere). Note that β_h is the value for a turbulent system, it corresponds to a highly intermittent process, not a process that is close to the integral of white noise (i.e. an Ornstein–Uhlenbeck process). LIM's exactly flat spectral behavior at low frequencies is a consequence of the fact that the highest order differential term is integer ordered, it implies that the low frequencies are (unpredictable) white noise. For times much larger than the decorrelation time, temperature forecasts have no skill. LIM's short memory behavior can be modeled as a Markov process, equivalently as an autoregressive or moving average process.

There are many empirical results that show a non-flat scaling behavior in the temperature spectrum (as well as in many other atmospheric variables) with values for β_l from 0.2 to 0.8 [see the review in Lovejoy and Schertzer (2013), also Lovejoy et al. (2018)]. This power-law behavior in the spectrum (and in the autocorrelation function) reflects the long-range memory that must be modelled. To appreciate the importance of the value of β_l for Gaussian processes, when $\beta_l = 0$, there is no predictability, and when $\beta_l = 1$, there is infinite predictability. The long memory effects mean that the equations become non-Markovian and that also past states need to be considered in order to predict the behavior of the system. The generalization of LIM's integer ordered differential equations to include fractional order derivatives already introduces power-law correlations, the simplest option being to retain the simplest (Gaussian) assumption about the noise forcing. This is the main idea behind the ScaLIng Macroweather Model (SLIMM) (Lovejoy et al. 2015).

In the macroweather regime intermittency is generally low enough that a Gaussian model with long-range statistical dependency is a workable approximation [except perhaps for the extremes; e.g. the review (Lovejoy et al. 2018)]. Some attempts have been made to use Gaussian models for prediction in the mean square prediction framework of autoregressive fractional integrated moving average (ARFIMA) processes (Baillie and Chung 2002; Yuan et al. 2015). The theory behind some of these models only applies to stationary series, while, for example, in the case of globally-averaged temperature time series, there is clearly an increasing trend due to the anthropogenic warming in recent decades. If the trend is not properly removed, the assumption of random equally distributed variables no longer applies, and the skill of the predictions is adversely affected. The ScaLIng Macroweather Model (SLIMM), (Lovejoy et al. 2015) was the first of such models that took all these facts into consideration and offered a complete evaluation of the prediction skill based on hindcasts after the removal of the anthropogenic warming part.

SLIMM is a model for the prediction of stationary series with Gaussian statistics and scaling symmetry of the fluctuations. It proposes a predictor as a linear combination of past data (or past innovations). For the case of Gaussian variables, it has been proven that this kind of linear predictor is optimal in the mean square error sense [see the "Fundamental note" in page 264 of Papoulis and Pillai (2002)]. That is, if any other functional form (i.e. nonlinear) is used to build a predictor based on past data, the mean square error of the predictions will be larger than with the linear combination. This is not necessarily true if the distribution of the variables is not Gaussian, for example, in the case of multifractal processes, where the second moment statistics are not sufficient to describe the process.

Similarly to the spectrum where $E_T(\omega) \sim \omega^{-\beta}$, in the macroweather regime the average of the fluctuations as a function of the time scale also presents a power-law (scaling) behavior with $\langle \Delta T(\Delta t) \rangle \sim \Delta t^H$. Besides the scale-invariance, low intermittency (rough Gaussianity) in time, is another characteristic of the macroweather regime. For Gaussian processes, the spectrum and the fluctuation exponents are related by $H = \left(\beta_l - 1\right)/2$. In Lovejoy et al. (2015) SLIMM was introduced, based on fractional Gaussian noise (fGn), as the simplest stochastic model that includes both characteristics.

For their relevance to the current work, some properties of fGn presented in that paper are summarized here; for an extensive mathematical treatment see Biagini et al. (2008).

Over the range $-1 < H < 0$, an fGn process, $G_H(t)$, is the solution of a fractional order stochastic differential equation of order $H + 1/2$, driven by a unit Gaussian δ-correlated white noise process, $\gamma(t)$, (with $\langle \gamma(t) \rangle = 0$ and $\left\langle \gamma(t)\gamma\left(t'\right) \right\rangle = \delta\left(t - t'\right)$, where $\delta(t)$ is the Dirac function):

$$\frac{d^{H+1/2} G_H(t)}{dt^{H+1/2}} = c_H \gamma(t), \tag{1}$$

where:

$$c_H^2 = \frac{\pi}{2\cos(\pi H)\Gamma(-2-2H)}, \tag{2}$$

and $\Gamma(x)$ is the Euler gamma function. The value for the constant c_H was the standard one chosen to make the expression for the statistics particularly simple, see below. The fractional differential equation (Eq. (1)) was presented in Lovejoy et al. (2015) as a generalization of the LIM integer order equation to account for the power-law behavior observed for the spectrum at frequencies $\omega > \omega_w \approx 1/\tau_w$. Physically it could model a scaling heat storage mechanism.

Integrating Eq. (1), we obtain:

$$G_H(t) = \frac{c_H}{\Gamma(H + 1/2)} \int_{-\infty}^{t} \left(t - t'\right)^{-(1/2-H)} \gamma\left(t'\right) dt'. \tag{3}$$

In other words, $G_H(t)$ is the fractional integral of order $H + 1/2$ of a white noise process, which can also be regarded as a smoothing of a white noise with a power-law filter. The process $\gamma(t)$ is a particular case of $G_H(t)$ for $H = -1/2$. Just as $\gamma(t)$ is a generalized stochastic process (a distribution), the process $G_H(t)$ is also a generalized function without point-wise values. It is the density of the well-known fractional Brownian motion (fBm) measures, $B_{H'}(t)$, with $H' = H + 1$, i.e. $dB_{H'}(t) = G_H(t)dt$ (Wiener integrals for the case $H' = 1/2$). The derivative of a distribution (in this case $B_{H'}(t)$) is formally defined from the following:

$$\int \varphi(t) dB_{H'}(t) = \int \varphi(t) G_H(t) dt = -\int \varphi'(t) B_{H'}(t) dt, \tag{4}$$

where $\varphi(t)$ is any locally integrable function.

From this relation to fBm, the resolution τ (smallest sampling temporal scale) fGn process, $G_{H,\tau}(t)$, can be defined, either as an average of $G_H(t)$, or from the increments of the fBm process, $B_{H'}(t)$, at the same resolution:

$$G_{H,\tau}(t)=\frac{1}{\tau}\int\limits_{t-\tau}^{t}G_H(t')dt'=\frac{1}{\tau}\int\limits_{t-\tau}^{t}dB_{H'}(t')=\frac{1}{\tau}\left[B_{H'}(t)-B_{H'}(t-\tau)\right]. \tag{5}$$

In Lovejoy et al. (2015) it was shown that, for resolution $\tau>\tau_w$, we can model the globally-averaged macroweather temperature as:

$$T_\tau(t)=\sigma_T G_{H,\tau}(t), \tag{6}$$

where $-1<H<0$ and σ_T is the temperature variance (for $\tau=1$). The parameter H, defined in this range, is not the more commonly used Hurst exponent for fBm processes, H', but the fluctuation exponent of the corresponding fractional Gaussian noise process. Fluctuations exponents are used due to their wider generality; they are well defined even for strongly non-Gaussian processes. For a discussion see page 643 in (Lovejoy et al. 2015).

Assuming τ is the smallest scale in our system with the property $\tau>\tau_w$ (e.g. $\tau=1$ month for air temperature), the temperature defined by Eq. (6) has the following properties:

(i) $T_\tau(t)$ is a Gaussian stationary process with continuous paths.

(ii) $\langle T_\tau(t)\rangle=0$ and $\left\langle T_\tau(t)^2\right\rangle=\sigma_T^2\tau^{2H}$; for all t, the notation $\langle\cdot\rangle$ denotes ensemble (infinite realizations) averaging.

(iii) $C_{H,\sigma_T}(\Delta t)=\langle T_\tau(t)T_\tau(t+\Delta t)\rangle$

$$=\frac{\sigma_T^2}{2\tau^2}\left(|\Delta t+\tau|^{2H+2}+|\Delta t-\tau|^{2H+2}-2|\Delta t|^{2H+2}\right);$$

for $\Delta t\geq\tau$. (7)

For more details see Mandelbrot and Van Ness (1968), Gripenberg and Norros (1996) and Biagini et al. (2008).

From Eq. (7.iii), the behavior of the autocovariance function for $\Delta t\gg\tau$ and $-1<H<0$ is:

$$C_{H,\sigma_T}(\Delta t)\approx\sigma_T^2(H+1)(2H+1)\Delta t^{2H} \tag{8}$$

and the corresponding spectrum for low frequencies is:

$$E_T(\omega)\approx\Gamma(3+2H)\sin(\pi H)\omega^{-\beta_l}\Big/\sqrt{2\pi}, \tag{9}$$

where $\beta_l=1+2H$.

Combining Eqs. (3), (5) and (6), we get the following explicit integral expression for the temperature at resolution τ:

$$T_\tau(t)=\frac{1}{\tau}\frac{c_H\sigma_T}{\Gamma(H+3/2)}\left[\int\limits_{-\infty}^{t}(t-t')^{H+1/2}\gamma(t')dt'-\int\limits_{-\infty}^{t-\tau}(t-\tau-t')^{H+1/2}\gamma(t')dt'\right]. \tag{10}$$

Notice that $T_\tau(t)$ is obtained from the difference of fractional integrals of order $H+3/2$ of a white noise process. Our definition of c_H in Eq. (2) implies that $\left\langle T_\tau(t)^2\right\rangle=\sigma_T^2\tau^{2H}$. As $H<0$, it follows that, in the small-scale limit ($\tau\rightarrow 0$), the variance diverges and H is the scaling exponent of the root mean square (RMS) value. This singular small-scale behavior is responsible for the strong power-law resolution effects in fGn. For a detailed discussion on this important resolution effect that leads to a "space–time reduction factor" and its implications for the accuracy of global surface temperature datasets, see Lovejoy (2017).

Using the fact that $T_\tau(t)$ is a Gaussian stationary process, Lovejoy et al. (2015) derived a formula for the predictor of the temperature at some time $t\geq\tau$, given that data are available over the entire past (i.e. from $t=-\infty$ to 0). From Eq. (10), the mean square (MS) estimator for the temperature can be expressed as:

$$\hat{T}_\tau(t)=\frac{1}{\tau}\frac{c_H\sigma_T}{\Gamma(H+3/2)}\int\limits_{-\infty}^{0}\left[(t-t')^{H+1/2}-(t-\tau-t')^{H+1/2}\right]\gamma(t')dt'. \tag{11}$$

As a measure of the skill of the model, we can use the mean square skill score (MSSS), defined as:

$$\mathrm{MSSS}(t,\tau)=1-\frac{\left\langle\left[T_\tau(t)-\hat{T}_\tau(t)\right]^2\right\rangle}{\left\langle T_\tau(t)^2\right\rangle}, \tag{12}$$

i.e. one minus the normalized mean square error (MSE). Here $T_\tau(t)$ represents the verification and $\hat{T}_\tau(t)$ the forecast at time $t\geq\tau$. The reference forecast would be the average of the series $\langle T_\tau(t)\rangle=0$, for which the MSE is the variance $\left\langle T_\tau(t)^2\right\rangle$. Using Eqs. (10) and (11) in (12), an analytical expression for the MSSS can be obtained:

$$\mathrm{MSSS}_H(t/\tau)=\frac{F_H(\infty)-F_H(t/\tau)}{F_H(\infty)+\frac{1}{2H+2}}, \tag{13}$$

where $t\geq\tau$ and

$$F_H(t)=\int\limits_{0}^{t-1}\left((1+u)^{H+1/2}-u^{H+1/2}\right)^2du; \tag{14}$$

in particular,

$$F_H(\infty)=\frac{\Gamma(3/2+H)\Gamma(-2H)}{(2H+2)\Gamma(1/2-H)}-\frac{1}{2H+2}. \tag{15}$$

Although Eq. (11) is the formal expression for the predictor of the temperature, from a practical point of view it has two clear disadvantages: it is expressed as an integral of the unknown past innovations, $\gamma(t)$, and it assumes the knowledge of these innovations for an infinite time in the past. It would be more natural to express the predictor as a function of the observed part of the process. This problem was solved for fBm processes with $1/2 < H' < 1$ (equivalently $-1/2 < H < 0$) by Gripenberg and Norros (1996). The explicit formula they found for the predictor, $\hat{B}_{H',a}(t)$, of the fBm process, $B_{H'}(t)$, known in the interval $(-a, 0)$ for $t > 0$ and $a > 0$, is:

$$\hat{B}_{H,a}(t) = \int_{-a}^{0} g_a\left(t, t'\right) B_H\left(t'\right) dt', \tag{16}$$

where $\mathrm{g}_a\left(t, t'\right)$ is an appropriate weight function given by:

$$g_a\left(t, -t'\right) = \frac{\sin\left[\pi\left(H' - 1/2\right)\right]}{\pi} \left[t'\left(a - t'\right)\right]^{-H'+1/2} \int_{0}^{t} \frac{[x(x+a)]^{H'-1/2}}{x + t'} dx. \tag{17}$$

It is important to note that the weight function goes to infinity both at the origin and at $-a$ [see Fig. 8 in Norros (1995)]. In their words, this divergence when we approach $-a$ is because "the closest witnesses to the unobserved past have special weight".

The results summarized in Eqs. (10–17) are theoretically important, but, from the practical point of view of making predictions, a discrete representation of the process is needed. In the next sections, we present analogous results for the prediction of discrete-in-time, finite past fGn processes and its application to the modelling and prediction of global temperature time series.

2.2 StocSIPS

The theory presented in the previous section and the applicability of SLIMM is restricted to detrended time series with Gaussian statistics and a scaling behavior of the fluctuations. Real-world datasets, in particular raw temperature series, normally include periodic signals corresponding to the diurnal and the seasonal cycles. They are also affected by an increasing trend as a response to anthropogenic forcing and usually combine different scaling regimes depending on the temporal resolution used.

StocSIPS is the general system that includes SLIMM as the core model for the long-term prediction of atmospheric fields. In order to use SLIMM, some of the components of StocSIPS are dedicated to the "cleaning" of the original dataset. In particular, it includes techniques for removing and projecting the seasonality and the anthropogenic trend. It also degrades the temporal series to a scale where only one scaling regime with fluctuation exponent $-1/2 < H < 0$ is present. The initial goal is to produce a temporal series that can be modelled and predicted with the stationary fGn process using the SLIMM theory. Some other aspects of StocSIPS—not discussed in this paper—include the addition of another space–time symmetry [the statistical space–time factorization (Lovejoy and de Lima 2015; Lovejoy et al. 2018)] for the regional prediction, and the combination as copredictors of different atmospheric fields.

One of the objectives of this paper is to show the improvements in the theoretical treatment and in the numerical methods of SLIMM as an essential part of StocSIPS. These recent developments have helped to produce faster and more accurate predictions of global temperature. The improvement in SLIMM and some of the preprocessing techniques are illustrated later on in Sect. 3 through an application to the forecast of globally-averaged temperature series.

2.2.1 Discrete-in-time fGn processes

As we showed in Sect. 2.1, for predicting the stationary component of the temperature with resolution τ at a future time $t > 0$, the linear predictor, $\hat{T}_\tau(t)$, based on past data ($T_\tau(s)$ for $-a < s \le 0$) satisfying the minimum mean square error condition (orthogonality principle between the error and the data) can then be written as:

$$\hat{T}_\tau(t) = \int_{-a}^{0} M_T(t, s) T_\tau(s) ds, \tag{18}$$

or equivalently, based on the past innovations, $\gamma(s)$:

$$\hat{T}_\tau(t) = \int_{-a}^{0} M_\gamma(t, s) \gamma(s) ds, \tag{19}$$

where $M_T(t, s)$ and $M_\gamma(t, s)$ are appropriated weight functions. In SLIMM, the predictor given by Eq. (11) is a particular case of Eq. (19) for $a = \infty$ and $M_\gamma(t, s) = c_H \sigma_T \left[(t-s)^{H+1/2} - (t-\tau-s)^{H+1/2}\right] / \tau \Gamma(H + 3/2)$, while the solution in Gripenberg and Norros (1996) (Eq. (16) here) is the case of Eq. (18) for an fBm process with $M_T(t, s)$ analogous to $\mathrm{g}_a\left(t, t'\right)$ given by Eq. (17).

The mathematical theory presented in Sect. 2.1 is general for a continuous-in-time fGn. Moreover, the integral representation of fGn given by Eq. (10), is based on an infinite past of continuous innovations, $\gamma(t)$. For applications to real-world data, a discrete version of the problem is needed for the case of fGn with finite data in the past ($a < \infty$). In practice, in the case of temperature (and any other atmospheric field) we only have measurements at discrete times with some resolution over a limited period. For modeling these fields, we can consider discrete-in-time fGn process as a more suitable model.

Assuming that we have already removed the low-frequency anthropogenic component of the temperature series (see Sect. 3.2), in the discrete case, we could express the zero mean detrended component by its moving average (MA(∞)) stochastic representation given by the Wold representation theorem (Wold 1938):

$$T_t = \sum_{j=-\infty}^{t} \varphi_{t-j}\gamma_j, \tag{20}$$

where $\{\varphi_t\}$ are weight parameters with units of temperature and $\{\gamma_t\}$ is a white noise sequence with $\gamma_t \sim \mathrm{NID}(0,1)$ and $\langle \gamma_i\gamma_j \rangle = \delta_{ij}$, where δ_{ij} is the Kronecker delta and $\mathrm{NID}(\mu, \sigma^2)$ stands for normally and independently distributed with mean μ and variance σ^2 (the sign ∼ means equal in distribution). This equation is analogous to Eq. (10) for the continuous case.

By inverting Eq. (20) we can obtain the equivalent autoregressive (AR(∞)) representation (Palma 2007):

$$T_t = \sigma_0\gamma_t + \sum_{j=-\infty}^{t-1} \pi_{t-j}T_j, \tag{21}$$

which is more suitable for predictions, as any value of the series is given as a linear combination of the values in the past. In this representation the weights $\{\pi_t\}$ are unitless.

In practice, we only have a finite stretch of data $\{T_{-t},\ldots,T_0\}$. Under this circumstance, the optimal k-steps Wiener predictor for T_k $(k > 0)$, based on the finite past, is given by:

$$\hat{T}_t(k) = \sum_{j=-t}^{0} \phi_{t,j}(k)T_j = \phi_{t,-t}(k)T_{-t} + \cdots + \phi_{t,0}(k)T_0, \tag{22}$$

where the new vector of coefficients, $\boldsymbol{\phi}_t(k) = [\phi_{t,-t}(k),\ldots,\phi_{t,0}(k)]^T$ (the superscript T denotes transpose operation) satisfies the Yule-Walker equations [see page 96 in Hipel and McLeod (1994)]:

$$\mathbf{R}^t_{H,\sigma_T}\boldsymbol{\phi}_t(k) = \mathbf{C}^t_{H,\sigma_T}(k), \tag{23}$$

with $\mathbf{C}^t_{H,\sigma_T}(k) = [C_{H,\sigma_T}(k-i)]^T_{i=-t,\ldots,0} = [C_{H,\sigma_T}(t+k),\ldots,C_{H,\sigma_T}(k)]^T$ and $\mathbf{R}^t_{H,\sigma_T} = [C_{H,\sigma_T}(i-j)]_{i,j=-t,\ldots,0}$ being the autocovariance matrix. The elements $C_{H,\sigma_T}(\Delta t)$ are obtained from Eq. (7.iii) where we assume $\tau = 1$ is the smallest scale in our system with the property $\tau \gg \tau_w$ (e.g. $\tau = 1$ month).

Notice that the coefficients $\{\phi_{t,j}\}$ will only depend on H [σ_T cancels in both sides of Eq. (23)] and further that they are not the same as the coefficients $\{\pi_t\}$, for which the complete knowledge of the infinite past is assumed. The coefficients $\{\pi_t\}$ decrease monotonically as we go further in the past, while this is not the case for the coefficients $\{\phi_{t,j}\}$, as we can see in Fig. 1 for the cases where $H = -0.1, -0.25, -0.4$, and we predict $k = 12$ steps in the future by using a series of $t + 1 = 36$ values. Notice how the memory effect (the weight of the coefficients) increases with the value of H. This behavior of the coefficients is analogous to the one mentioned earlier for the function $g_a(t,t')$ (Eq. (17)). As found in Gripenberg and Norros (1996) for the continuous-in-time case, not only is there a strong weighting of the recent data, but the most ancient available data also have singular weights [compare Fig. 1 here with Fig. 3.1 in (Gripenberg and Norros 1996)].

This behavior of the coefficients for fGn is the main difference (and a clear advantage) over other autoregressive models (AR, ARMA) which do not include fractional integrations accounting for the long-term memory and do not consider the information from the distant past. An additional limitation of these approaches is that for each Δt, the values for $C(\Delta t) = \langle T_\tau(t)T_\tau(t+\Delta t)\rangle$ must be estimated directly from the data. Each $C(\Delta t)$ will have its own error, this effectively introduces a large "noise" in the predictor estimates. In addition, it is computationally expensive if a large number of coefficients are needed. In our fGn model the coefficients have an analytic expression which only depends on the fluctuation exponent, H, obtained directly from the data exploiting the scale-invariance symmetry of the fluctuations; our problem is a statistically highly constrained problem of parametric estimation (H), not an unconstrained one (the entire $C(\Delta t)$ function).

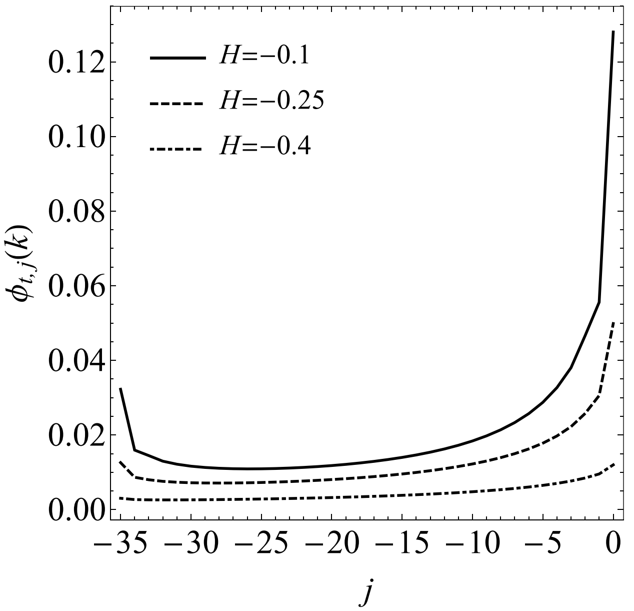

Fig. 1 Optimal coefficients, $\phi_{t,j}$, in Eq. (17) with $H = -0.1, -0.25, -0.4$ (top to bottom) for predicting $k = 12$ steps in the future by using the data for $j = -35,\ldots,0$ in the past. Notice the strong weighting on both the most recent (right) and the most ancient available data (left) and how the memory effect decreases with the value of H. Compare to Fig. 3.1 in Gripenberg and Norros (1996)

In the discrete case, the mean square skill score, defined by Eq. (12), has the following analytical expression:

$$\mathrm{MSSS}_H^t(k) = \tilde{\mathbf{C}}_H^t(k)^T \left(\tilde{\mathbf{R}}_H^t\right)^{-1} \tilde{\mathbf{C}}_H^t(k), \tag{24}$$

where $\tilde{\mathbf{C}}_H^t(k) = \left[\tilde{\mathbf{C}}_H(k-i)\right]_{i=-t,\ldots,0}^T$ is a vector formed by the autocorrelation function $\tilde{C}_H(\Delta t) = C_{H,\sigma_T}(\Delta t)/\sigma_T^2$ (see Eq. (7. iii)) and $\tilde{\mathbf{R}}_H^t = \mathbf{R}_{H,\sigma_T}^t/\sigma_T^2 = \left[\tilde{C}_H(i-j)\right]_{i,j=-t,\ldots,0}$ is the autocorrelation matrix. For a given horizon in the future, k, the MSSS will only depend on the exponent, H, and the extension of our series in the past, t.

In the previous equations, the full length of our known series was $t+1$, but we don't necessarily have to use the complete series to build our predictor. It is enough to use a number $m+1$ of points in the past (memory) with $m < t$. The new predictor and skill score are obtained by just replacing t by m in Eqs. (22–24). By doing this, we can use the remaining $t-m-1$ points for hindcast verifications.

For the case where $H = -0.25$ and $k = 3$, Fig. 2 shows how the MSSS approaches the asymptotic value corresponding to an infinite past as we increase the amount of memory we use. The dashed line represents the MSSS for $m = 500$ and the dotted line is the value we obtain using Eq. (13) for the continuous-in-time case with the infinite past known. The difference between the two is not due to the finite memory ($m = 500$) we have in the discrete case with respect to the infinite past assumed in Eq. (13), but to intrinsic differences due to the discretization and more related to the high-frequency information loss because of the smoothing from a continuous to a discrete process. Note that we do not need to use a large memory to achieve a skill close to the asymptotic value. In this example where $H = -0.25$, we only need to use $m \geq 22$ for $k = 3$ to get more than 95% of the maximum skill.

The amount of memory needed depends on the value of H, as we can see in Fig. 3, where we plot the minimum memory needed, $m_{95\%}$, to get more than 95% of the asymptotic value (corresponding to $m = \infty$) as a function of the horizon, k, for different values of H. The line $m = 15k$ was also included for reference. The larger the value of the exponent, H, (the closer to zero) the less memory we need to approach the maximum possible skill. This fact seems counterintuitive, but the explanation is simple: for larger values of H (closer to zero), the influence of values farther in the past is stronger, but at the same time, more information of those values is included in the recent past, so less memory is needed for forecasting. Following the rule of thumb found by Norros (1995) for the continuous case: "one should predict (…) the next second with the latest second, the next minute with the latest minute, etc." Actually, from Fig. 3 we can conclude than, for predicting k steps into the future, a memory $m = 15k$ would be a safe minimum value for achieving almost the maximum possible skill for any value of H in the range $(-1/2, 0)$, which is the case for temperature and many other atmospheric fields. Of course, if H is close to zero a

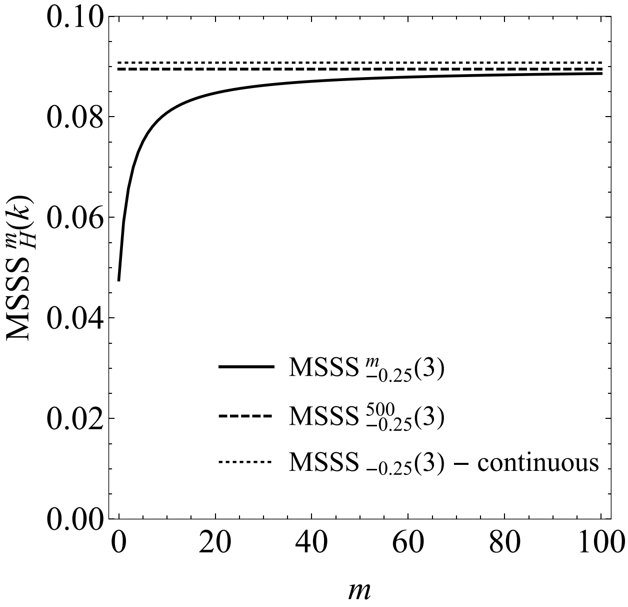

Fig. 2 $\mathrm{MSSS}_H^m(k)$ as a function of the memory, m, for the case where $H = -0.25$ and $k = 3$. The dashed line represents the MSSS for $m = 500$ and the dotted line is the value obtained with Eq. (12) for the continuous-in-time case. For $m = 22$, more than 95% of the asymptotic skill is achieved

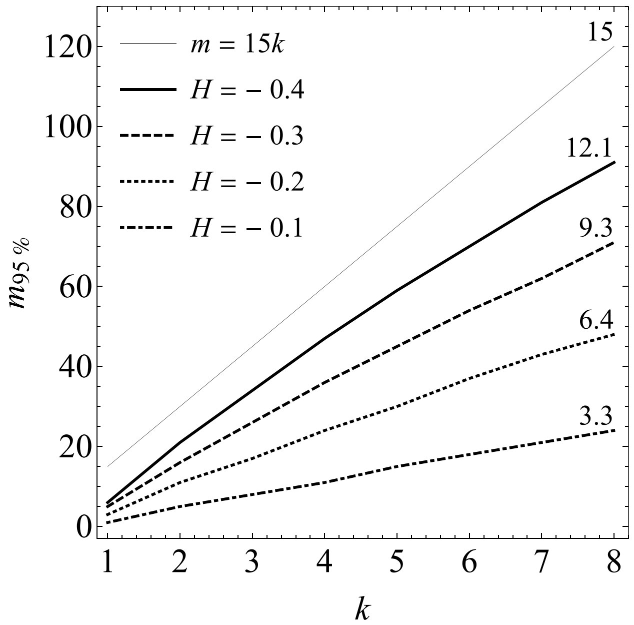

Fig. 3 Minimum memory, m, needed to get more than 95% of the asymptotic value (corresponding to $m = \infty$) as a function of the horizon, k, for different values of H. The larger the value of H (the closer to zero) the less memory is needed for a given horizon. The approximate ratio $m_{95\%}/k$ for each H was included at the top of the respective curve

much smaller value could be taken. The approximate ratio $m_{95\%}/k$ for each H was included at the top of the respective curve. From the point of view of the availability of data for the predictions, this result is important. Once the value for H is estimated, assuming it remains stable in the future, we only need a few of recent datapoints to forecast the future temperature. The information of the unknown data from the distant past is automatically considered by the model.

Previously, we showed that an fGn process is fully characterized by its autocovariance function, which in turn depends only on the covariance, σ_T^2, and the fluctuation exponent H. To extend our description to more general cases, we could allow our series to have a non-zero ensemble mean, μ. This family of three parameters defines our fGn process and represents the link between the mathematical model and real-world historical data.

In Appendix 1 we discuss how to obtain maximum likelihood estimates (MLE) for these parameters on a given time series. For the fluctuation exponent, we show other approximate (and less computationally expensive) methods. We can use Eq. (9) to obtain $\hat{H}_s = \left(\beta_l - 1\right)/2$ from the spectrum exponent at low frequencies. This method, as well as the Haar wavelet analysis to obtain an estimate $\hat{H}_h$ from the exponent of the Haar fluctuations, was used in Lovejoy and Schertzer (2013) and Lovejoy et al. (2015) to obtain estimates of H for average global and Northern Hemisphere anomalies. A Quasi Maximum Likelihood Estimate (QMLE) method is also discussed in Appendix 1. The latter is more accurate than the Haar fluctuations and the spectral analysis methods and is obtained as part of the hindcast verification process. Nevertheless, those two have the advantage of being more general and applicable to any scaling process (even highly nonGaussian ones).

All these methods were applied to fGn simulations and the parameters estimated were summarized in Table 4. The technical details for producing exact simulations are also discussed in Appendix 1. Finally, we show how to to check the adequacy of the fitted fGn model to real-world data and we derive some ergodic properties of fGn processes. Specifically, we show that the temporal average standard deviation squared, $SD_T^2 = \sum_{t=1}^{N} \left(T_t - \bar{T}_N\right)^2/N$, is a strongly biased estimate of the variance of the process, σ_T^2, for values of H close to zero (the overbar denotes temporal averaging: $\bar{T}_N = \sum_{t=1}^{N} T_t/N$). The sample and the ensemble estimates are related by:

$$SD_T^2 = \sigma_T^2\left(1 - N^{2H}\right). \tag{25}$$

When $H = -0.06$, $N = 1656$ (values for the monthly series since 1880) there is a huge difference between the sample and the ensemble estimates ($SD_T^2/\sigma_T^2 = 0.59$). Some skill scores (e.g. the MSSS or the normalized mean squared error NMSE) use the variance for normalization. The implications of the difference in the estimates of the variance on the definition of the MSSS will be discussed in Sect. 3.4.3.

3 Forecasting global temperature anomalies

3.1 The data

The general framework presented here is applicable to forecasting any time series that satisfies, (a) the conditions of stationarity, (b) Gaussianity and (c) long-range dependence given by power-law behavior of the correlation function with fluctuation exponents in the range $(-1/2, 0)$. These three properties are well satisfied for globally-averaged temperature anomaly time series in the macroweather regime, from 10 days to some decades (Lovejoy and Schertzer 2013; Lovejoy et al. 2013, 2015). In the last three decades, there has been a growing literature showing that the temperature (and other atmospheric fields) are scaling in the macroweather regime (Koscielny-Bunde et al. 1998; Blender et al. 2006; Huybers and Curry 2006; Franzke 2012; Rypdal et al. 2013; Yuan et al. 2015) and see the extensive review in Lovejoy and Schertzer (2013). Strictly speaking, in the last century, low frequencies become dominated by anthropogenic effects and after 10–20 years the scaling regime changes from a negative to a positive value of H, as we will show below. As was discussed in detail in Lovejoy (2014, 2017) and Lovejoy et al. (2015), differently from preindustrial epochs, recent temperature time series can be modeled by a trend stationary process, i.e. a stochastic process from which an underlying trend (function solely of time) can be removed, leaving a stationary process. In other words, to first order, variability is unaffected by climate change. The deterministic trend representing the response to external forcings can be removed by using CO_2 radiative forcing as a good linear proxy for all the anthropogenic effects [or equivalent-CO_2 (CO_2eq) radiative forcing as the one used for CMIP5 simulation (Meinshausen et al. 2011)]. There is a nearly linear relation between the actual CO_2 concentration and the estimated equivalent concentration which includes all anthropogenic forcings, including greenhouse gases, aerosols, etc. (Meinshausen et al. 2011).

In this paper, we limit our analysis to globally-averaged temperature anomaly time series at monthly resolution. This is a first step for checking the applicability of the model and at the same time providing an alternative method for obtaining long-term forecasts. The quality of our method can be assessed based on the skill obtained from hindcasts verification and its agreement with the theoretical prediction.

There are five major observation-based global temperature datasets which are in common use. They are (a) the NASA Goddard Institute for Space Studies Surface Temperature Analysis (GISTEMP) series, abbreviated NASA and NASA-L in the following for global and land surface averages respectively (Hansen et al. 2010; GISTEMP Team 2018), (b) the NOAA NCEI series GHCN-M version 3.3.0 plus ERSST dataset (Smith et al. 2008; NOAA-NCEI 2018), updated in Gleason et al. (2015), abbreviated NOAA and NOAA-L (global and land surface averages, as before), (c) the Combined land and sea surface temperature (SST) anomalies from CRUTEM4 and HadSST3, Hadley Centre—Climatic Research Unit Version 4, abbreviated HAD4 and HAD4-L (Morice et al. 2012; Met Office Hadley Centre 2018), (d) the version 2 series of (Cowtan and Way 2014, 2018), abbreviated CowW and CowW-L, and (e) the Berkeley Earth series (Rohde et al. 2013; Berkeley Earth 2018), abbreviated Berk and Berk-L. The average of the global and the land surface series were included in the analysis and we use for the abbreviations Mean-G and Mean-L, respectively.

All these series are of anomalies, i.e. the difference between temperature at a given time and the average during a baseline period. They tend not to be on the same baseline; for NASA and Berk the reference period is 1951–1980, for HAD4 and CowW it is 1961–1990, and for NOAA it is the 20th century (1901–2000). To compare them, we need to use the same zero point. In this case we chose the 20th century average as a common reference period. The average temperature for 1901–2000 is nearly the same as that for 1951–1980, while that of more recent times (1961–1990) is warmer.

Each series spans a somewhat different period: HAD4, CowW and Berk start first, beginning in 1850, NASA and NOAA both start in 1880. When the data were accessed on May 21, 2018, they were all available at monthly resolutions until April 2018. Only the period January 1880–December 2017 was analyzed, i.e. 138 years = 1656 months (the same length that was used in the simulations in Appendix 1). These series (updated until 2012), together with twentieth century reanalysis global average, were used in (Lovejoy 2017) to assess how accurate are the data as functions of their time scale. As it was pointed out in the latter, each data set has its strengths and weaknesses and it is precisely their degree of agreement or disagreement what permits us to evaluate the intrinsic absolute uncertainty in the estimates of the global temperature.

In Fig. 4 we show the global average temperature (bottom) and the land surface average temperature (top). In red are the means of the five global datasets for global and for land, respectively, and in blue are a measure of their level of dispersion given by the standard deviations. The datasets are most dissimilar before 1900, which could be due to the lack of reliable measurements, but otherwise, the overall level of agreement is very good [about ± 0.05 °C and is nearly independent of scale for the global temperature series (Lovejoy 2017)]. Each series shows warming during the last decades, and they all show fluctuations superimposed on the warming trend.

3.2 Removing the anthropogenic component

In the present case of globally-averaged temperatures, the seasonality in the time series is weak. The deterministic annual cycle component was removed first from the original series. It was estimated from the average of every month for the full period of 138 years (1880–2017). Cross-validation effects are weak for such a long reference period and were not considered.

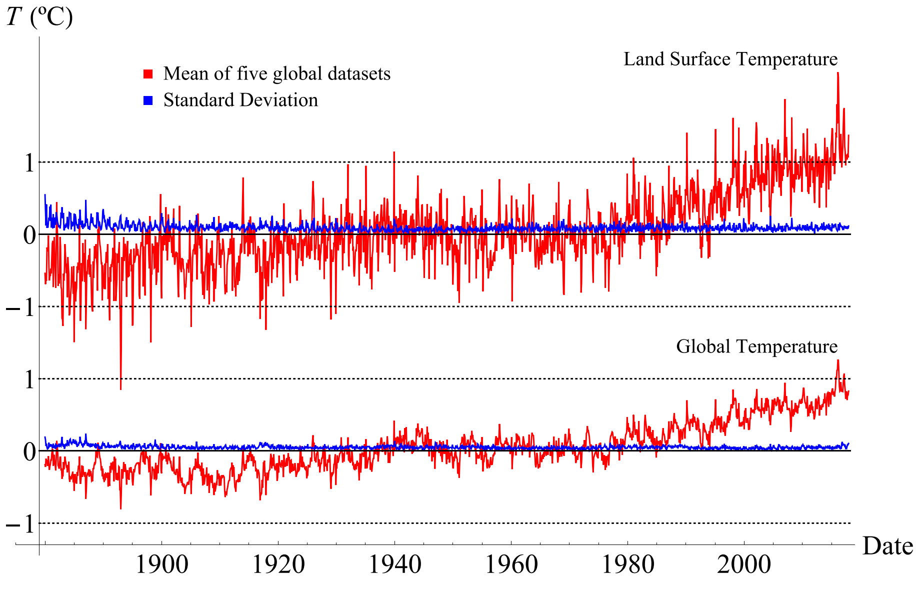

Fig. 4 Monthly surface temperature anomaly series from 1880 to 2017. In red is the mean of the five datasets for global (bottom): NASA, NOAA, HAD4, CowW, and Berk, and for land (top): NASA-L, NOAA-L, HAD4-L, CowW-L, and Berk-L. The dispersion among the series—given by the standard deviations of the five series as a function of time—is shown in blue. Each series represents the anomaly with respect to the mean of the reference period 1901–2000

Because of the anthropogenically induced trends in addition to internal macroweather variability, global temperature time series have low-frequency forced variability. A simple application of the linearity of the climate response to external forcings yields:

$$T(t) = T_{\mathrm{anth}}(t) + T_{\mathrm{nat}}(t), \tag{26}$$

which considers the temperature as a combination of a purely deterministic response to anthropogenic forcings, T_{anth}, plus a strict stationary stochastic component, T_{nat}, with zero mean. The low frequency component can be obtained as:

$$T_{\mathrm{anth}}(t) = \lambda_{2\times\mathrm{CO_2eq}} \log_2 \left[\rho_{\mathrm{CO_2eq}}(t) / \rho_{\mathrm{CO_2eq,pre}}\right] + T_0, \tag{27}$$

where $\rho_{\mathrm{CO_2eq}}$ is the observed globally-averaged equivalent-CO$_2$ concentration with preindustrial value $\rho_{\mathrm{CO_2eq,pre}} = 277$ ppm and $\lambda_{2\times\mathrm{CO_2eq}}$ is the transient climate sensitivity (that excludes delayed responses) related to the doubling of atmospheric equivalent-CO$_2$ concentrations. For $\rho_{\mathrm{CO_2eq}}$ we used the CMIP5 simulation values (Meinshausen et al. 2011). The definition of CO$_2$eq here includes not only greenhouse gases, but also aerosols, with their corresponding cooling effect. The reference value T_0 is chosen so that $\bar{T}_{\mathrm{nat}} = 0$, (the overbar indicates temporal averaging). The parameters $\lambda_{2\times\mathrm{CO_2eq}}$ and T_0 are estimated from the linear regression of $T(t)$ vs. $\log_2 \left[\rho_{\mathrm{CO_2eq}}(t) / \rho_{\mathrm{CO_2eq,pre}}\right]$. The residuals are the stochastic natural variability component, T_{nat}.

The natural variability includes "internal" variability and the response of the system to natural forcings: solar and volcanic. There is no gain in trying to model the responses to these two natural forcings independently. They would represent unpredictable signals while the ensemble of T_{nat} can be directly modelled using the techniques discussed in Sect. 2 for fGn processes. We made some experiments trying to predict the internal variability and the solar and the volcanic responses independently, and the combined error was larger than if we try to forecast the natural variability component as a whole. On the other hand, the relatively smooth dependence of the anthropogenic component makes it easy to project it a few years into the future with good accuracy.

As an example, the temperature anomalies for the global average dataset (Mean-G) is shown in Fig. 5 (red in the online version) together with the CO$_2$eq response to anthropogenic forcings (dashed, black) and the residual natural variability component (blue). To use CO$_2$ instead of CO$_2$eq forcings leads to almost the same residuals due to the nearly linear relation between the two, but it avoids the uncertainties due to the estimation of the cooling effects of the aerosols as well as other radiative assumptions. The CO$_2$ forcing is taken as a surrogate for all the anthropogenic forcings. The focus of this work is to model and forecast the residuals (natural variability), and for that purpose, either of the two concentrations would lead to the same residuals (they differ by a factor of 1.12 over the last century). From a direct inspection of Fig. 5, it is clear that a CO$_2$eq response does a much better job of reproducing the actual trend of the temperature series than a simple regression linear in time, which is often used for estimating the warming trend.

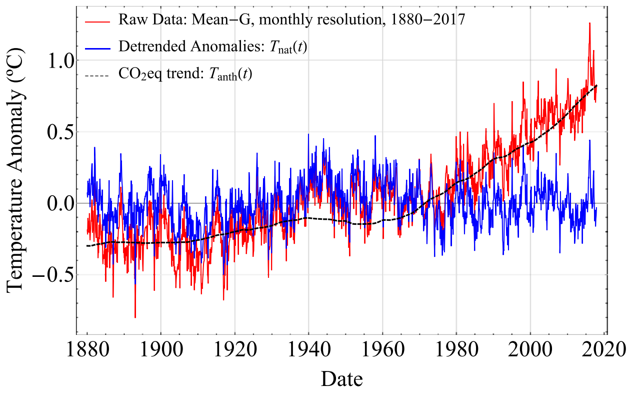

Fig. 5 Temperature anomalies for the Mean-G dataset (red in the online version) together with the CO$_2$eq trend (dashed, black) and the residual natural variability component (blue)

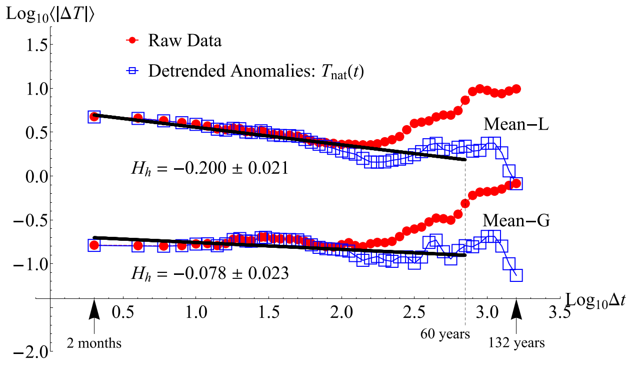

Fig. 6 Haar fluctuation analysis for the Mean-G (bottom) and Mean-L (top) datasets before (red) and after (blue) removing the trends. The reference lines with slopes $H_h = -0.064 \pm 0.020$ for the global series and $H_h = -0.241 \pm 0.017$ for the land surface series were obtained from regression of the residuals between 2 months and 60 years. The last points were dropped to get better statistics. The units for Δt and ΔT are months and °C, respectively

Before making predictions, we need to verify the adequacy of the model and verify the hypothesis that the residual natural variability component has scaling fluctuations with exponent in the range $(-1/2, 0)$. The Haar fluctuation analysis for the Mean-G (bottom) and Mean-L (top) datasets before and after removing the anthropogenic trends are shown in Fig. 6 (red for the raw dataset fluctuations and blue for the detrended series in the online version). The reference lines with slopes $H_h = -0.078 \pm 0.023$ for the global series and $H_h = -0.200 \pm 0.021$ for the land surface series were obtained from regression of the residuals' fluctuations between 2 months and 60 years. The points corresponding to scales of more than 60 years were not considered for estimating the parameters as there were not many fluctuations to average at those time scales. In addition, some of the low frequency natural variability was presumably removed with the forced variability. The units for Δt and ΔT are months and °C, respectively. Notice that the anthropogenic warming breaks the scaling of the fluctuations at a time scale of around 10 years (the red and blue curves diverge at ~ 100 months). The residual natural variability, on the other hand, shows reasonably good scaling for the whole period analyzed (138 years). The same range of scaling with decreasing fluctuations has been obtained in temperature records from preindustrial multiproxies and GCMs preindustrial control runs (Lovejoy 2014).

The global series are a composition of land surface data and sea surface temperature data. The average temperature over the ocean shows fluctuations increasing with the time scale (positive H) up to 2 years. This corresponds to the ocean weather regime as discussed in Lovejoy and Schertzer (2013). The same break in the scaling is found in the global temperature fluctuations, but this break is subtle, and an overall unique scaling regime can be assumed for the global data. The influence of the ocean on the global temperature also brings its fluctuation exponent towards higher values (closer to zero) compared to the land surface fluctuations. This makes the global data more predictable than the land-only series.

In the frequency domain, the corresponding spectra for the Mean-G dataset are shown in Fig. 7. The raw spectrum for the natural variability series is represented in grey. It shows scaling, but with large fluctuations, as expected. To get better estimates of the exponent we can average the raw spectra using logarithmically spaced bins. These "cleaner" spectra for the series before and after removing the anthropogenic trend are shown in red and blue in the online version, respectively. Notice that they only differ appreciably for the low-frequency range, corresponding to the removed deterministic trend. The frequency, ω, is given in units of $(138\,\text{years})^{-1}$. The particularly low variabilities at frequencies corresponding to $(30\,\text{years})^{-1}$ is an artefact of the 30-year detrending period used in most of the datasets. The solid black line was obtained from a linear regression on the residues. The exponent obtained from the absolute value of the slope was $\beta = 0.81 \pm 0.13$. Using the monofractal relation $\beta = 1 + 2H$, we obtain the estimate for the fluctuation exponent: $H_s = -0.096 \pm 0.063$. The dashed reference line with slope corresponding to $\beta_h = 1 + 2H_h = 0.84 \pm 0.05$ was included in the figure for comparison (using the value obtained from the Haar fluctuation analysis in Fig. 6).

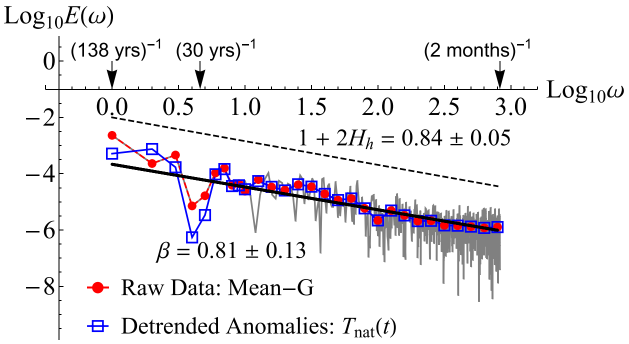

Fig. 7 Spectra for the Mean-G dataset. In grey is the raw spectrum of the residuals. Averages with logarithmically spaced bins are shown for the series before (dashed, red) and after (blue) removing the trend. The solid black line, with slope $-\beta$, was obtained from a linear regression on the residues. The reference dashed line with absolute value of the slope $1 + 2H_h = 0.84$ was included for comparison (using the value obtained from the Haar fluctuation analysis in Fig. 6). The frequency, ω, is given in units of $(138\,\text{years})^{-1}$

It is worth mentioning that this very simple approach to removing the warming trend is a special (low memory) case of the much more general model of linear response theory with a scaling response function proposed by Hébert et al. (2019). In this work, the authors directly exploit the stochasticity of the internal variability and the linearity and scaling of the forced response to make projections based on historical data and a scaling step Climate Response Function that has a long memory. They not only include anthropogenic effects, but also solar and volcanic forcings. Consequently, the residuals they obtain once these forced components are removed, do not represent the forced natural variability response, but the internal variability of the system. The authors based their analysis on the assumption that this internal stochastic component can be approximated by an fGn process. This hypothesis has been confirmed on GCMs preindustrial control runs outputs where the forcings are not present.

3.3 Fitting fGn to global data

Having obtained the stationary natural variability component, T_{nat}, for the Mean-G dataset from the residuals of the linear regression of $T(t)$ vs. $\log_2\left[\rho_{\mathrm{CO_2eq}}(t)/\rho_{\mathrm{CO_2eq,pre}}\right]$ (Eqs. (26) and (27)), we can now model this series using the theory presented in Sect. 2 and Appendix 1. The first step is to obtain the parameters μ, σ_T^2 and H. We would like to underline that these parameters describe the—infinite ensemble—fGn stochastic process, but we can only obtain estimates for them based on a single realization (our globally-averaged temperature time series). In Appendix 1 we show how to obtain the MLE for μ and σ_T^2. In the case of the fluctuation exponent, we can repeat the methods presented in Sect. 3.2 and obtain estimates from the slopes in the Haar fluctuations and the spectrum curves. However, as we mentioned before, it is clear in Figs. 6 and 7 that the error in the estimates is much higher for these methods than by using the MLE or QMLE due to the high variability of the fluctuations. Nevertheless, their advantage over the latter is that they are general and apply not only to Gaussian processes (such as fGn), but also to multifractal or other intermittent processes with different statistics. The MLE and QMLE methods make the extra assumption of adequacy of the fGn model, which ultimately must be verified.

To get an idea of how well the stochastic model describes the observational dataset, we created completely synthetic time series by superimposing fGn simulations on the low-frequency anthropogenic trend. Four randomly chosen simulations are shown in Fig. 8 together with the Mean-G dataset (top). The synthetic series were created using $\lambda_{2\times\mathrm{CO_2eq}} = 2.03$ °C and $T_0 = -0.379$ °C for the anthropogenic trend, T_{anth}, and following the procedure described in Appendix 1-i with parameters $\mu = 0$ °C, $\sigma_T = 0.195$ °C and $H = -0.060$ for simulating T_{nat} (see Eqs. (26) and (27)). All these parameters were obtained by fitting the Mean-G observations in the period 1880–2017 ($N = 1656$ months). In Appendix 2 (Table 5), we summarize the parameters obtained for the ten datasets and the corresponding mean series for global and for land.

Although a visual inspection of Fig. 8 is not a convincing proof of the applicability of the model, it is clear that if we eyeball the completely synthetic time series with the observational Mean-G dataset, you cannot tell which is which. A simple verification of the fGn behavior of the detrended data can be done by checking that the biased temporal estimate of the variance, SD_T^2, and the value obtained using maximum likelihood, $\hat{\sigma}_T^2$, satisfy Eq. (25) (derived in Appendix 1-iii.).

Following Eq. (25), the temporal estimate of the variance should depend on the number of months, n, that is used for the estimates: $SD_T^2(n) = \sigma_T^2\left(1 - n^{2H}\right)$. For only one time series, the estimate of $SD_T^2(n)$ is noisy. To reduce the noise, this value can be estimated using k-segments of the series from $t = k$ to $t = k + n - 1$ (each of length n), and then averaged over the total ensemble of segments (in this case $N_{\mathrm{segments}} = N - n_{\max}$, where $N = 1656$ months is the full length of the series and $n_{\max} = 120$ months is the maximum length of the segments used):

$$\left\langle SD_T^2(n)\right\rangle = \frac{n-1}{n} SD_T^2(n) = \frac{1}{N-n_{\max}} \sum_{k=1}^{N-n_{\max}} \left[\frac{1}{n}\sum_{t=k}^{k+n-1}\left(T_t-\bar{T}_n\right)^2\right], \tag{28}$$

where $\bar{T}_n = \sum_{t=1}^{n} T_t/n$, the values T_t are for the natural variability component of the Mean-G dataset and the factor $(n-1)/n$ accounts for the bias of the length-n sample estimate, $SD_T^2(n)$, with respect to the length-n population variance, $\left\langle SD_T^2(n)\right\rangle$.

In Fig. 9 we show in red line with circles the empirical values of the standard deviation $\left\langle SD_T^2(n)\right\rangle^{1/2}$ as a function of n (obtained using Eq. (28) for the ensemble of $N - n_{\max}$ segments). The function $f_{\sigma_T,H}(n) = \sigma_T\sqrt{\left(1-n^{2H}\right)\left(1-n^{-1}\right)}$ (obtained by replacing the expression for $SD_T^2(n)$ in Eq. (28) and taking the square root) is plotted using $\sigma_T = \hat{\sigma}_T = 0.195$ °C and the following values of H: $H_f = -0.069$ (solid black line), obtained from the fit of the red curve; $H_l = -0.060$ (dashed line), obtained using MLE, and $H_q = -0.080$ (dotted line), from the QMLE. The empirical curve for a synthetic realization of Gaussian white noise with standard deviation $\sigma_{\mathrm{wn}} = 0.141$ °C was also included for comparison (blue line with squares).

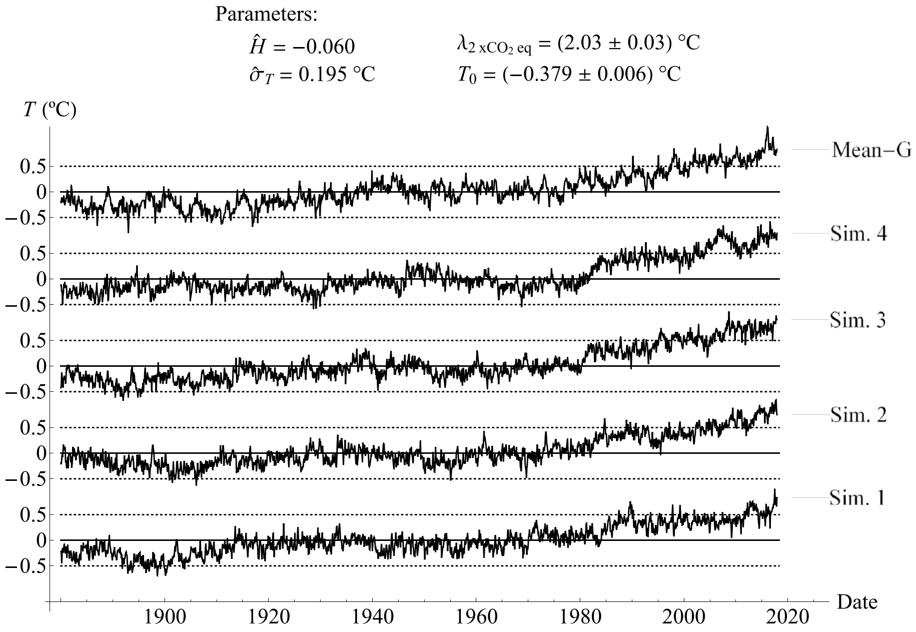

Fig. 8 Four randomly chosen synthetic time series together with the Mean-G dataset (top). The simulations were created by superimposing fGn simulations for T_{nat} to the low-frequency anthropogenic trend, T_{anth} (see Appendix 1 and Eqs. (26) and (27)). The parameters used for the simulation (shown in the figure) were obtained by fitting the Mean-G series in the period 1880–2017

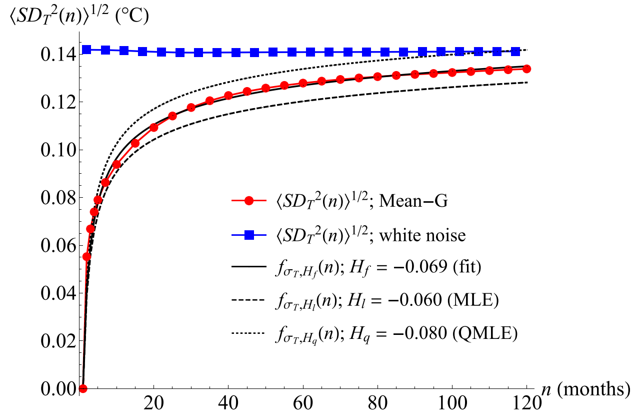

Fig. 9 Empirical values of $\left\langle SD_T^2(n)\right\rangle^{1/2}$ as a function of n, obtained using Eq. (28) (red line with circles). The function $f_{\sigma_T,H}(n) = \sigma_T\sqrt{\left(1-n^{2H}\right)\left(1-n^{-1}\right)}$, with $\sigma_T = \hat{\sigma}_T = 0.195$ °C, is plotted for three values of H: $H_f = -0.069$ (solid black line), obtained from the fit of the red curve; $H_l = -0.060$ (dashed line), obtained using MLE and $H_q = -0.080$ (dotted line), from QMLE. The empirical curve for a synthetic realization of Gaussian white noise with variance $\sigma_{\mathrm{wn}}^2 = 0.02$ °C was also included for comparison (blue line with squares). The agreement between the red line with circles and the solid black line is an evidence of the fGn behavior of the natural variability

The difference between the red curve for the observational time series and the blue curve for the uncorrelated synthetic series illustrates the effects of the long-range correlations in the natural variability of the globally-averaged temperature time series. This strong dependence of the estimates of the variance with the length of the estimation period for H close to zero could have an influence on statistical methods that depend on the covariance matrix [e.g. empirical orthogonal function (EOF) and empirical mode decompositions (EMD)].

The agreement between the $\left\langle SD_T^2(n)\right\rangle^{1/2}$ curve estimated from the data and the function $f_{\sigma_T,H}(n)$—that only depends on the two parameters σ_T and H—is an evidence of the good fit of the fGn stochastic model to the natural variability. At the same time, it could be used as an alternative method for obtaining the parameters σ_T and H by fitting the curve $\left\langle SD_T^2(n)\right\rangle^{1/2}$ based on observations using the function $f_{\sigma_T,H}(n)$.

More detailed statistical tests to check the fit of the model to the data are shown in Appendix 2 using the theory presented at the end of Appendix 1. The main conclusion is that the global average temperature series can be considered Gaussian as well as their innovations, while for the case of land average temperature, there are some deviations from Gaussianity. Nevertheless, the residual autocorrelation functions (RACF) satisfy the normality condition with good enough accuracy for all datasets, corroborating the whiteness of the innovations and hence that an fGn model can be considered a good approximation in all cases.

3.4 Forecast and validation

3.4.1 The low-frequency anthropogenic component

Ultimately, as a final step to confirm the adequacy of the model to simulating and forecasting global temperature data, we present the skill scores obtained from hindcast verifications and compare their values with the theoretical predictions. First, we should point out that for predicting the global temperature we need to forecast both the anthropogenic component and the natural variability. Our final estimator for k steps into the future, following Eq. (26), is given by:

$$\hat{T}(t+k) = \hat{T}_{\mathrm{anth}}(t+k) + \hat{T}_{\mathrm{nat}}(t+k), \tag{29}$$

where $\hat{T}_{\mathrm{nat}}$ is obtained from Eq. (22) using the theory presented in Sect. 2.2.1. The anthropogenic component, which we model with a separate low-frequency process must also be forecast. Nevertheless, even if we use persistence of the $\mathrm{CO_2eq}$ increments, the error on predicting the low-frequency component is small compared to the error on forecasting the natural variability (for lead times up to a year or so). For this reason, for obtaining $\hat{T}_{\mathrm{anth}}(t+k)$ based on the previous values of the trend, we just assume persistence of the increments $\Delta T_{\mathrm{anth}}(t,k) = T_{\mathrm{anth}}(t) - T_{\mathrm{anth}}(t-k)$, that is:

$$\begin{aligned}\hat{T}_{\mathrm{anth}}(t+k) &= T_{\mathrm{anth}}(k) + \Delta T_{\mathrm{anth}}(t,k)\\ \hat{T}_{\mathrm{anth}}(t+k) &= 2T_{\mathrm{anth}}(t) - T_{\mathrm{anth}}(t-k).\end{aligned} \tag{30}$$

For a linear trend, the absolute error $\left\langle \left| T_{\mathrm{anth}}(t+k) - \hat{T}_{\mathrm{anth}}(t+k)\right|\right\rangle = \left\langle \left|\Delta T_{\mathrm{anth}}(t+k,k) - \Delta T_{\mathrm{anth}}(t,k)\right|\right\rangle = 0$. In the case of the $\mathrm{CO_2eq}$ trend shown in black in Fig. 5, for small k, the function is almost linear in a k-vecinity of any t. This justifies the rejection of this error compared to the error on forecasting the natural variability. For reference, the root mean square error (RMSE) using this method for the anthropogenic component, in the 1044-months hindcast period January 1931–December 2017, performed with $k = 24$ months in advance for every month, was of 0.01 °C for all global datasets.

3.4.2 The natural variability component

For the natural variability, the expectation of the RMSE—taking the infinite ensemble average using the theory for fGn—for a prediction k steps into the future is defined by:

$$\mathrm{RMSE}_{\mathrm{nat}}^{\mathrm{theory}}(k) = \sqrt{\left\langle \left[T_{\mathrm{nat}}(t+k) - \hat{T}_{\mathrm{nat}}(t+k)\right]^2\right\rangle}. \tag{31}$$

According to the definition of MSSS, given by Eq. (12), and the analytical expression, Eq. (24), a theoretical ensemble estimate of $\mathrm{RMSE}_{\mathrm{nat}}(k)$, for prediction using a memory of m steps, is given by:

$$\mathrm{RMSE}_{\mathrm{nat}}^{\mathrm{theory}}(k)=\mathrm{RMSE}_{H,\sigma_T}^{m}(k)=\sigma_T\sqrt{1-\tilde{\mathbf{C}}_H^m(k)^T\left(\tilde{\mathbf{R}}_H^m\right)^{-1}\tilde{\mathbf{C}}_H^m(k)}. \tag{32}$$

Notice that, unlike the MSSS, this is not only a function of the horizon, k, the memory, m, and the exponent, H, but also of the specific series we are forecasting due to the presence of the parameter σ_T, which must be estimated using Eq. (50) in Appendix 1. As expected, for given values of k, m and H, the RMSE is proportional to the amplitude of the series we want to predict.

3.4.3 Validation

To validate our model, we produced series of hindcasts at monthly resolution, each for a different horizon from 1 to 12 months, in the verification period January 1931–December 2017. For this hindcast series each subsequent point plotted on the graph was independently predicted using the information available k months before. What changes from month to month is the initialization date while the forecast horizon is kept fixed. Such hindcast series are useful because they show how close the predictions are to the observations for a given value of k. The dependence with the horizon of many scores (e.g., the RMSE), are obtained from the difference between hindcasts series at a fixed k and the corresponding series of observations.

StocSIPS assumes an additive fixed annual cycle independent of the low-frequency trend; it does not make distinctions from month to month from the point of view of the statistics of the anomalies. In fact, is this month-to-month correlation that is exploited as a source of predictability in the stochastic model. Nevertheless, there is always an intrinsic multiplicative seasonality in the data that is impossible to completely remove without affecting the scaling behavior of the spectrum. To account for the effects of this seasonality, we can stratify the observations and the forecasts series to show dependences with the initialization date.

For each horizon, k, we used a memory $m = 20k$. For example, to predict the average temperature for January 1931 with $k = 1$ month, we used the previous 21 months, including December 1930, and the same was done for each month up to December 2017. For $k = 2$ months, we used the previous 41 months, including December 1930, to produce the first forecast for February 1931, and so on.

Examples of the hindcasts series initialized every month, each for a different horizon, are shown in Fig. 10 for the Mean-G natural variability. In blue, we show the hindcasts series for $k = 1$, 3 and 6 months (bottom to top). In red we show the verification curve of observations for the natural variability starting in January 1931. The vertical gridlines correspond to the forecast and verification for each January; that is, initializing the first day of each January with data up to every December in the bottom panel, up to every October in the middle panel and up to every July in the top one. This shows how the stratification is done for obtaining dependences of the skill with the initialization date (shown later).

As can be seen in Fig. 10, there is a reduction of the amplitude and an increasing lag between the observed and forecast time series as the horizon increases (more noticeable in the top panel). This is due to the model tendency to predict the return rate towards the mean as a function of H. Extremes can therefore only be predicted as a consequence of the anthropogenic increase. However, the general behavior of the temperature is well predicted.

Equation (31) is the definition of the infinite ensemble expectation of the RMSE, for which we get an analytical expression (Eq. (32)). The all-months verification RMSE can then be computed from the series shown in Fig. 10 as:

$$\mathrm{RMSE}_{\mathrm{nat}}(k)=\sqrt{\frac{1}{N-k+1}\sum_{t=0}^{N-k+1}\left(T_{\mathrm{nat}}(t+k)-\hat{T}_{\mathrm{nat}}(t+k)\right)^2} \tag{33}$$

where $N = 1044$ months (from January 1931 to December 2017) and the number of terms in the sum is reduced in $k - 1$ because the last verification date (December 2017) is the same for every k while the first verification date is k months after December 1931 ($t = 0$) for each horizon. This equation can be adapted to get the RMSE for each horizon and for each initialization month.

In Fig. 11a, we show a comparison between the RMSE obtained from the hindcasts of all the months in the verification period 1931–2017 using Eq. (33) and the theoretical

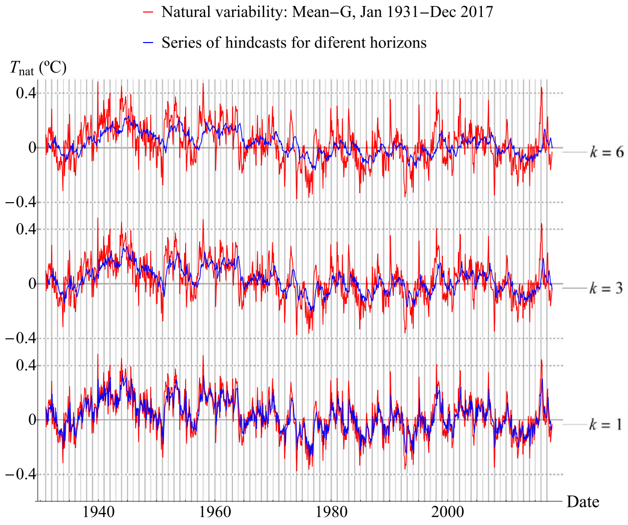

Fig. 10 In blue, series of hindcasts for the Mean-G natural variability initialized every month for horizons $k = 1$, 3 and 6 months (bottom to top). In red, the verification curve of observations for the natural variability starting in January 1931. The vertical gridlines correspond to the forecast and verification for each January; that is, initializing with data up to every December in the bottom panel, every October in the middle and every July in the top

expected RMSE, which is only a function of $\hat{\sigma}_T$, H and m (Eq. (32)). The agreement between the theory (solid black) and the actual errors (red curve) is another confirmation of the model for the simulation and prediction of global temperature. In the figure, we also included the values $\hat{\sigma}_T = 0.195$ °C and $SD_T = 0.147$ °C for the Mean-G natural variability (dotted and dashed lines respectively). The value of the former is the same as shown in Table 5, while the value of the latter is slightly different from the value reported there because now it was computed for the verification series in the period 1931–2017 (red curve in Fig. 10). Notice that, for $N = 1044$ months and $H = -0.060$ (see Table 5), $SD_T/\sqrt{1 - N^{2H}} = 0.195$ °C, in perfect agreement with the value of $\hat{\sigma}_T$ for that dataset.

The error for the anthropogenic trend forecast calculated using Eq. (30) is always less than 7% of the $\mathrm{RMSE}_{\mathrm{nat}}$ shown in Fig. 11a (see the final paragraph of Sect. 3.4.1). Because of this, its contribution to the overall error, $\mathrm{RMSE}_{\mathrm{raw}}$, on forecasting the raw temperature (natural plus anthropogenic) is lower than 0.4% for all horizons (compare the red-circles and the blue-squares curves in Fig. 11a). For all practical purposes, $\mathrm{RMSE}_{\mathrm{raw}} \approx \mathrm{RMSE}_{\mathrm{nat}}$ with a high degree of accuracy.

In Fig. 11b, we show a density plot with the RMSE as a function of the forecast horizon and the initialization month. The diagonal pattern from the top-left corner to the

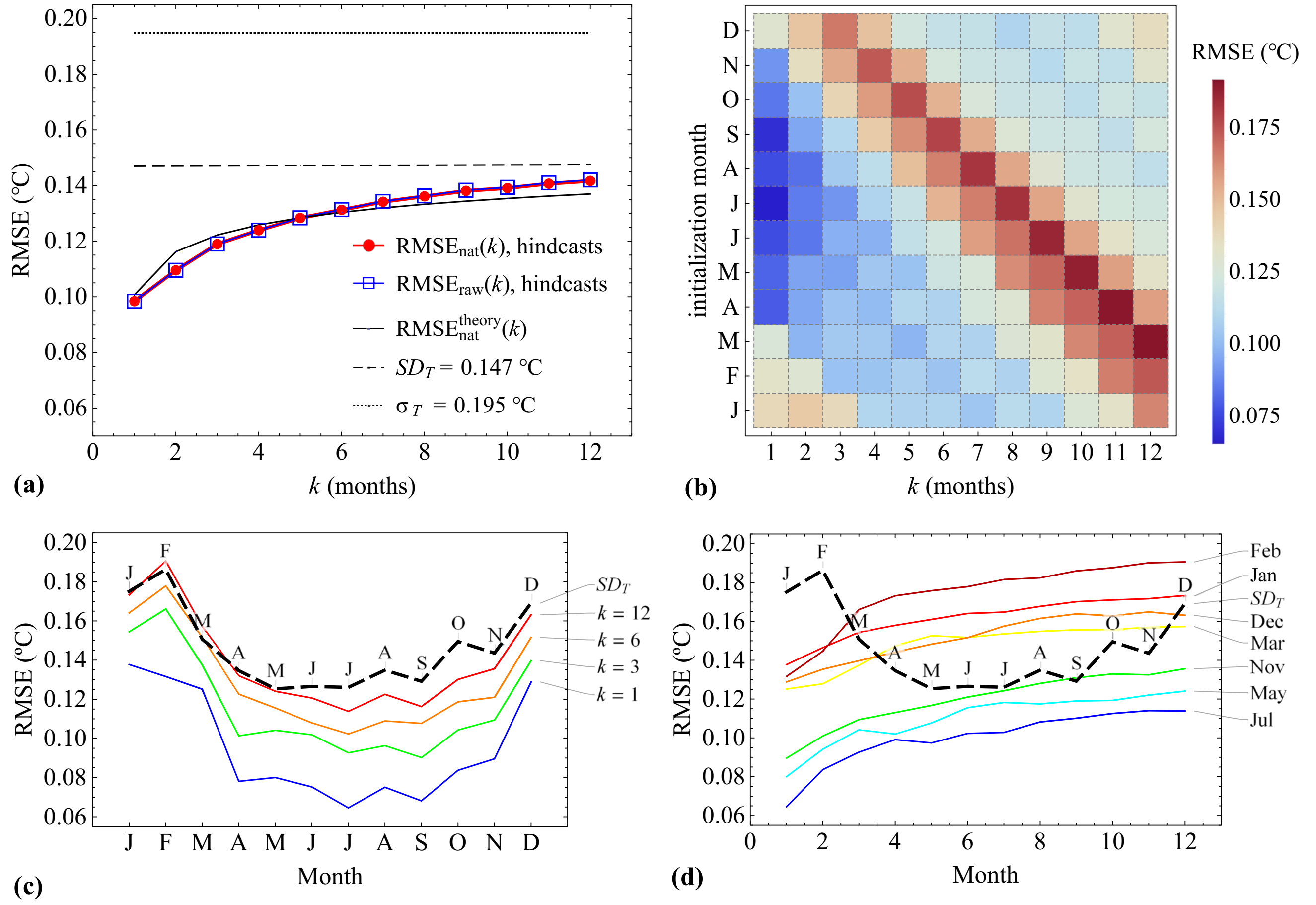

Fig. 11 RMSE of StocSIPS forecasts for the Mean-G dataset. **a** Curves of $\mathrm{RMSE}_{\mathrm{nat}}(k)$ (red circles) and $\mathrm{RMSE}_{\mathrm{raw}}(k)$ (blue squares), for the natural variability component and for the raw series, respectively. The curves were obtained using Eq. (33) from the hindcasts of the Mean-G dataset including all the months in the verification period 1931–2017. The difference between the two is negligible. The theoretical expected $\mathrm{RMSE}_{\mathrm{nat}}^{\mathrm{theory}}(k)$ (solid black), given by Eq. (32), is also shown for comparison. The values of $\hat{\sigma}_T$ (Table 5) and SD_T for the Mean-G natural variability were included for reference (dotted and dashed lines, respectively). **b** Density plot with the RMSE as a function of the forecast horizon and the initialization month. The diagonal pattern from the top-left corner to the bottom-right is an indication of the intrinsic multiplicative seasonality in the time-series. **c** Graphs of RMSE vs. initialization month for different forecast horizons ($k = 1$, 3, 6 and 12 months). There is an increase in the RMSE for the forecast of the Boreal winter months associated to the increase in the standard deviation, SD_T, of the globally-averaged temperature for those months (shown in dashed black line in the bottom panels figures). **d** Graphs of RMSE vs. k for different initialization months. For large values of k the skill of the model is small and the value of the RMSE is close to the standard deviation for that specific month (dashed black line). The RMSE graph in **a** is close to the average of the RMSE graphs in **d**

bottom-right is an indication of the intrinsic seasonality in the time-series. This is shown in detail in the bottom panels figures.

In Fig. 11c, we show graphs of RMSE vs. initialization month for different forecast horizons ($k = 1$, 3, 6 and 12 months). There is an increase in the RMSE for the forecast of the Boreal winter months associated to the increase in the variability (standard deviation, SD_T) of the globally-averaged temperature for those months (shown in dashed black line in the bottom panels figures). In Fig. 11d, we show graphs of RMSE vs. k for different initialization months. As expected, there is an increase in the RMSE with k. For large values of k the skill of the model is small and the value of the RMSE is close to the standard deviation for that specific month (dashed black line). The RMSE graph in panel (a) is close to the average of the RMSE graphs in panel (d). It is actually the all-month MSE the one that is the average of the MSEs for each month (as long as the number of years used for the average is the same for every month).

Related to the RMSE score, the mean square skill score (MSSS) is a commonly used metric:

$$\mathrm{MSSS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE_{ref}}}, \tag{34}$$

where $\mathrm{MSE} = \mathrm{RMSE}^2$ is computed using Eq. (33) and $\mathrm{MSE_{ref}}$ is the mean square error of some reference forecast.

The climatology—constant annual cycle taken from the average in a given reference period of at least 30 years—is commonly used as reference forecast. In this case, $\mathrm{MSE_{ref}} = SD^2_{\mathrm{raw}}$, is the variance of the raw series:

$$SD^2_{\mathrm{raw}} = \overline{\left(T_{\mathrm{anth}} + T_{\mathrm{nat}}\right)^2} = \overline{T^2_{\mathrm{anth}}} + SD^2_T \tag{35}$$

(assuming that the natural and anthropogenic variabilities are independent) and we call $\mathrm{MSSS} = \mathrm{MSSS_{raw}}$.

If we take as reference the anthropogenic trend forecast, then $\mathrm{MSE_{ref}} = SD^2_T$, is the variance of the natural variability component (detrended series, T_{nat}) and we name $\mathrm{MSSS} = \mathrm{MSSS_{nat}}$. This would be the same as the skill on forecasting the detrended series taking as reference forecast its mean value. Using the theoretical expressions for SD^2_T and for $\mathrm{RMSE} = \mathrm{RMSE}^{\mathrm{theory}}_{\mathrm{nat}}(k)$ (Eqs. (25) and (32), respectively) we can obtain an analytical expression for $\mathrm{MSSS_{nat}}$:

$$\mathrm{MSSS}^{\mathrm{theory}}_{\mathrm{nat}}(k) = \frac{\mathrm{MSSS}^m_H(k) - N^{2H}}{1 - N^{2H}}, \tag{36}$$

where $\mathrm{MSSS}^m_H(k)$ was defined for the infinite ensemble average in Eq. (24) [Eq. (13) for the continuous-time case]. Notice that $\mathrm{MSSS}^{\mathrm{theory}}_{\mathrm{nat}}(k)$ is not only a function of the fluctuation exponent, H, and the memory used for the forecasts, m, but also of the length of the verification period, N. For an infinite series, the ergodicity of the system is verified; i.e. the temporal average is equal to the ensemble average: $\mathrm{MSSS}^{\mathrm{theory}}_{\mathrm{nat}}(k) = \mathrm{MSSS}^m_H(k)$ (recall $H < 0$). We can check the agreement between the theoretical result (Eq. (36)) and the $\mathrm{MSSS_{nat}}$ obtained from hindcast to verify the validity of the model.

The anomaly correlation coefficient (ACC) is another commonly used verification score. In this case, we can also obtain the ACC for the raw or for the detrended series:

$$\mathrm{ACC}_{\mathrm{nat/\,raw}}(k) = \frac{\overline{T_{\mathrm{nat/\,raw}}(t+k)\hat{T}_{\mathrm{nat/\,raw}}(t+k)}}{SD_{T/\,\mathrm{raw}}\sqrt{\overline{\hat{T}_{\mathrm{nat/\,raw}}(t)^2}}}, \tag{37}$$

where we assume that $T(t)$ and the predictor $\hat{T}(t)$ are zero mean anomalies, the overbars indicate temporal average for a constant forecast horizon, k, and either all the subscripts are "nat" or all are "raw" depending on whether we forecast the detrended or the raw anomalies, respectively. In the latter case, spurious high values of the ACC (similarly for the MSSS) are found due to the presence of the deterministic trend. This is a very common flaw found throughout the literature, where this score is routinely reported for undetrended anomalies.

It is useful to note the relationship between the ACC and MSSS obtained from minimum mean square predictions. It can be easily seen from the orthogonality principle, $\left\langle \hat{T}\left(T - \hat{T}\right)\right\rangle = 0$, that the stochastic predictions satisfy

$$\mathrm{ACC_{nat}}(k) = \sqrt{\mathrm{MSSS_{nat}}(k)} \tag{38}$$

for any horizon k. This relation can also be used to check the agreement between the theoretical predictions of the model and the actual results obtained from hindcasts verification.

In Fig. 12 we summarize the results for the MSSS (top) and the ACC (bottom). In Fig. 12a, we show curves of MSSS vs. k for the Mean-G dataset considering all months in the verification period 1931–2017. In red line with circles, the curve for $\mathrm{MSSS_{nat}}$ taking as reference the anthropogenic trend forecast, for which $\mathrm{MSE_{ref}} = SD^2_T$ ($SD_T = 0.147$ °C). In green line with triangles, the values for $\mathrm{MSSS_{raw}}$ taking as reference the climatology forecast with $\mathrm{MSE_{ref}} = SD^2_{\mathrm{raw}}$ ($SD_{\mathrm{raw}} = 0.293$ °C). The theoretical expected $\mathrm{MSSS}^{\mathrm{theory}}_{\mathrm{nat}}(k)$ (solid black), given by Eq. (36), is also shown for comparison. There is relatively good agreement between this theoretical prediction of the model and the MSSS values obtained from the verification. The asymptotic value of $\mathrm{MSSS}^{\mathrm{theory}}_{\mathrm{nat}}(k)$ for $N \to \infty$ (given by Eq. (24)) is shown in dotted line with squares (dashed line for the continuous-time case, Eq. (13)). The longer the verification period the closer the MSSS will be to that asymptotic value. For the discrete theoretical curves (solid black line and dotted black with squares), we used a memory $m = 20k$. The small difference for $k = 1$ month, between this curve and the one for the continuous case (solid black) is due to the high-frequency information loss in the discretization process.

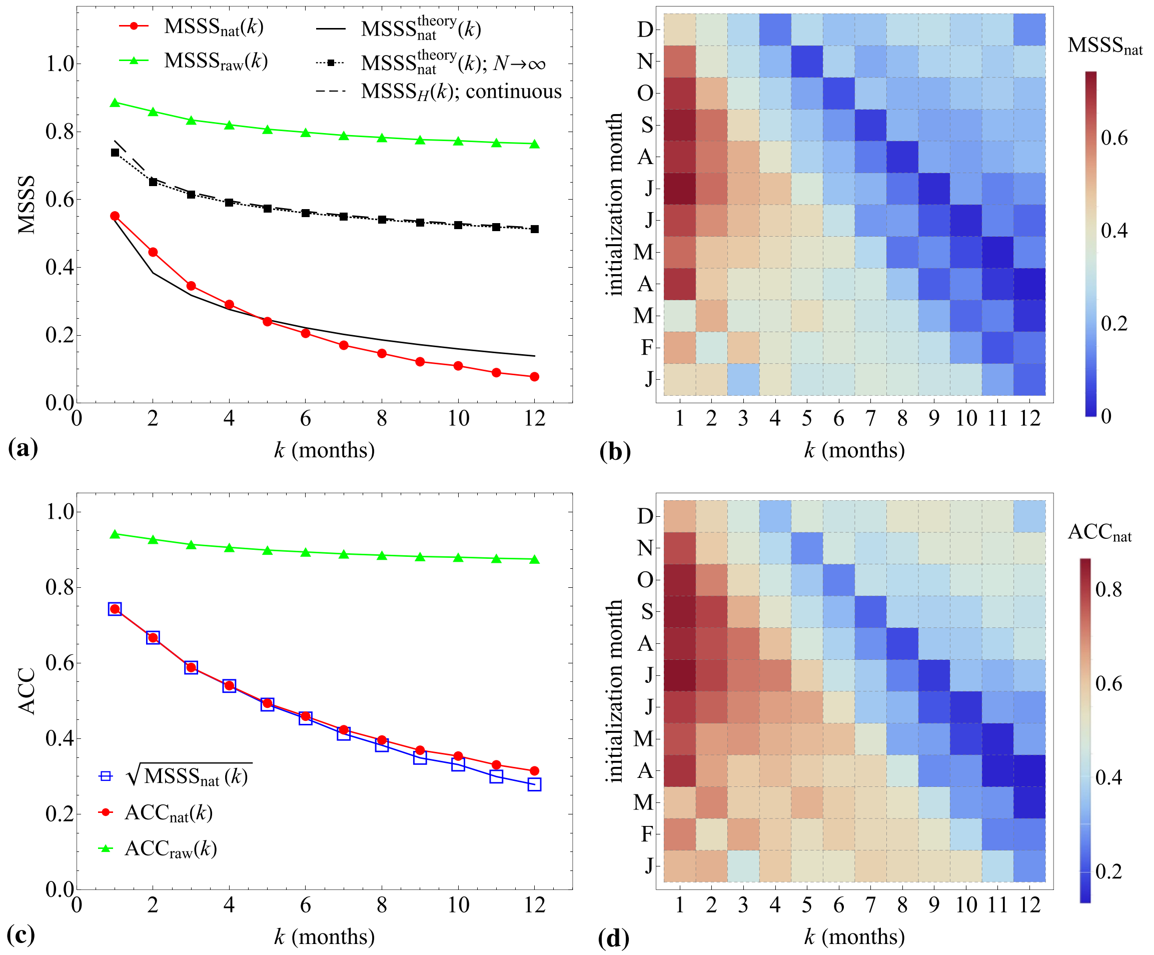

Fig. 12 MSSS and ACC of StocSIPS forecasts for the Mean-G dataset. **a** Curves of MSSS vs. k for the Mean-G dataset considering all months in the verification period 1931–2017. In red line with circles, the curve for $\mathrm{MSSS_{nat}}$ taking as reference the anthropogenic trend forecast. In green line with triangles, the values for $\mathrm{MSSS_{raw}}$ taking as reference the climatology forecast. The theoretical expected $\mathrm{MSSS}_{\mathrm{nat}}^{\mathrm{theory}}(k)$ (solid black), given by Eq. (36), is also shown for comparison. The asymptotic value for $N \to \infty$ (given by Eq. (24)) is shown in dotted line with squares (dashed line for the continuous-time case, Eq. (13)). The longer the verification period the closer will be the MSSS to that asymptotic value. **b** Density plot showing the MSSS as a function of the forecast horizon and the initialization month. **c** Curves of $\mathrm{ACC_{nat}}$ (red circles) and $\mathrm{ACC_{raw}}$ (green triangles) as a function of the forecast horizon obtained from Eq. (37). The values of $\sqrt{\mathrm{MSSS_{nat}}}$ (blue squares) were included to check the consistency of the theoretical relationship given by Eq. (38). **d** Density plot of the ACC as a function of the forecast horizon and the initialization month. The diagonal patterns from the top-left corner to the bottom-right in **b**, **d** are consequences of the intrinsic seasonality in the time-series

In Fig. 12c, we show curves of $\mathrm{ACC_{nat}}$ (red circles) and $\mathrm{ACC_{raw}}$ (green triangles) obtained from Eq. (37). Here, we can appreciate the spuriously high correlation values of $\mathrm{ACC_{raw}}$ compared to the $\mathrm{ACC_{nat}}$ due to the presence of the anthropogenic trend. The values of $\sqrt{\mathrm{MSSS_{nat}}}$ (blue squares) were included to check the consistency of the theoretical relationship given by Eq. (38); we see that it is relatively well satisfied, confirming the validity of the model.

In the right panels of Fig. 12, we show density plots with the MSSS and the ACC [panels (b) and (d), respectively] as a function of the forecast horizon and the initialization month. As we already showed for the RMSE, there are diagonal patterns from the top-left corner to the bottom-right as a consequence of the seasonality in the globally-averaged temperature anomalies. Nevertheless, for the MSSS and the ACC, these patterns are relatively less significative compared to the ones in the RMSE because—roughly speaking—both scores are functions of the ratio $\mathrm{RMSE_{nat}}/SD_T$, reducing the impact of the variation of the standard deviation of each individual month (see Fig. 11c). Some results of the hindcast validation are summarized in Table 7 for the twelve datasets, including the mean series for the global and the land surface.

3.4.4 Parametric probability forecast

Probability forecasts from long-term prediction dynamical models are usually obtained by fitting probability distributions to the ensemble forecast for each month and deriving probabilities of three climatologically equiprobable categories: below normal, near normal and above normal conditions. In general, the form of the distribution and the skill of the forecast is affected by the size of the ensemble. One of the main advantages of StocSIPS over conventional numerical models is that, by its inherent stochastic nature, the infinite ensemble parametric probability forecast can be obtained analytically without the need of simulating any individual realization. Following the results presented in Sect. 2, the theoretical probability distribution forecast at horizon k, taking data up to time t, is a Gaussian with mean $\mu_f = \hat{T}(t+k)$ given by Eq. (29) and standard deviation $\sigma_f(k) = \mathrm{RMSE}^m_{H,\sigma_T}(k)$ given by Eq. (32) (we neglected the error in the projection of the anthropogenic trend). In this section we only consider results for the full time series without stratification of the data. The theoretical expression for $\sigma_f(k)$, obtained from the results for an infinite ensemble, only applies in this case.

The "reliability" is defined as the consistency or repeatability of the probabilistic forecast. In order to evaluate the reliability of the probabilistic forecast of an ensemble model, the ensemble spread score (ESS) is commonly used as a summarizing metric. The ensemble spread score (ESS) is defined as the ratio between the temporal mean of the intra-ensemble variance, $\sigma^2_{\text{ensemble}}$, and the mean square error between the ensemble mean and the observations (Palmer et al. 2006; Keller and Hense 2011; Pasternack et al. 2018):

$$\mathrm{ESS} = \frac{\overline{\sigma^2_{\text{ensemble}}}}{\mathrm{MSE}}. \tag{39}$$

In the case of StocSIPS, $\overline{\sigma^2_{\text{ensemble}}} = \sigma_f^2$ is obtained analytically using Eq. (32) and $\mathrm{MSE} = \mathrm{RMSE}^2$ is obtained from the hindcasts using Eq. (33).

Following Palmer et al. (2006), an ESS of 1 indicates perfect reliability. The forecast is "overconfident" when ESS < 1; i.e. the ensemble spread underestimates forecast error. If the ensemble spread is greater than the model error (ESS > 1), the forecast is "overdispersive" and the forecast spread overestimates forecast error. In Fig. 11a, we showed that there is good agreement between the theoretical estimate $\mathrm{RMSE}^m_{H,\sigma_T}(k) = \sigma_f(k)$ and the hindcast error $\overline{\mathrm{RMSE}_{\text{nat}}}(k)$ for all horizons k, or—what is the same—between $\overline{\sigma^2_{\text{ensemble}}}$ and MSE in Eq. (39). This gives a value of $\mathrm{ESS} \approx 1$, so that StocSIPS is a nearly perfectly reliable system without needing a recalibration of the forecast probability distribution.

Examples of probability forecasts for July 1984 for the natural variability component of the Mean-G dataset are shown in Fig. 13 for horizons $k = 1$ and 3 months (left and right panels, respectively). That is, using data up to June 1984 for the $k = 1$ month forecast and up to April 1984 for $k = 3$ months. The normal probability density function (PDF) in grey represents the climatological distribution of the monthly temperatures for the detrended anomalies of the Mean-G dataset for the full period 1931–2017, for which $\sigma_{\text{clim}} = SD_T = 0.147$ °C. The terciles of the climatological distribution are indicated by vertical dashed lines. These vertical lines define three equiprobable categories of above normal, near normal, and below normal monthly temperatures observed in the verification period. The forecast distribution is indicated by the black curve with the forecast mean $\mu_f = \hat{T}(\text{Jul } 1984) = -0.118$ °C and standard deviation $\sigma_f = \mathrm{RMSE}^m_{H,\sigma_T}(k) = 0.101$ °C for $k = 1$ month (left panel) and $\mu_f = -0.063$ °C, $\sigma_f = 0.122$ °C for $k = 3$ months (right panel). The areas under the forecast PDF in different colors indicate probabilities of below normal (blue), near normal (yellow), and above normal (pink) temperatures. These probabilities are summarized in the top-left corner as bar plots. The climatological probability of 33% is indicated by the horizontal dashed line. The observed temperature for that specific date, $T_{\text{obs}} = -0.191$ °C, is represented by the vertical green line. The forecast distributions for $k = 1$ month are sharper than for $k = 3$ months. As expected, the confidence of the probabilistic forecast decreases as the lead time increases and they become more conservative.

The verification of the probabilistic forecast in categories (above, near and below normal) is done using 3×3 contingency tables (Stanski et al. 1989). The forecast and observed categories are simply classified in a table of three rows and three columns. There is a row for each observed category and a column for each forecast category. For each month forecast, one is added to the grid element of the contingency table according to the intersection of the forecast category and the observed category. In Table 1 we show the contingency table for the $k = 1$ month forecast of the natural variability anomalies, T_{nat}, of the Mean-G dataset (red curves in Fig. 10). The 1044 month period (Jan 1931–Dec 2017) was used for verification. The climatological distribution was defined using the mean and standard deviation of the detrended series over that period.

There are many scores that can be obtained from the contingency table (Stanski et al. 1989). In this paper we used the percent correct (PC) obtained from the elements in the main diagonal (shown in bold in Table 3). This score, often called accuracy, is very intuitive and it counts, overall, the percentage of the category forecasts that were correct. From Table 1, we obtain the values $\mathrm{PC}_{\text{nat}} = 100\,(272 + 160 + 250)/1044 \approx 65\%$. We can obtain contingency tables for all k. The dependence of the PC with k, is shown in Fig. 14 for the forecasts of the detrended anomalies, T_{nat} (blue line with squares in the figure). The dashed line at 33.3% is a reference showing the skill of the climatological forecast.

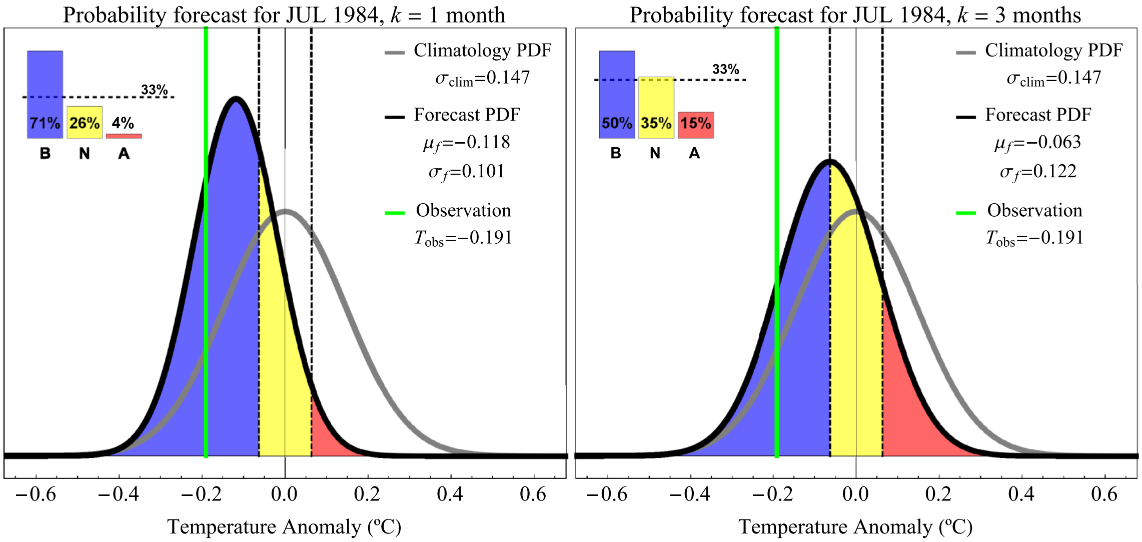

Fig. 13 Example of parametric probability forecasts for July 1984 for the natural variability component of the Mean-G dataset for horizons $k = 1$ and 3 months (left and right panels, respectively). That is, using data up to June 1984 for the $k = 1$ month forecast and up to April 1984 for $k = 3$ months. The normal probability density function in grey represents the climatological distribution of the monthly temperatures for the detrended anomalies of the Mean-G dataset for the full period 1931–2017. The terciles of the climatological distribution are indicated by vertical dashed lines. The colored areas under the forecast density function are proportional to the forecast probabilities for each category: below normal (blue), near normal (yellow) and above normal (pink). These probabilities are summarized in the top-left corner as bar plots. The climatological probability of 33% is indicated by the horizontal dashed line. The observed temperature for that specific date, T_{obs} = −0.198 °C, is represented by the vertical green line. The parameters for all the distributions are included in the legends

Table 1 Contingency table for the $k = 1$ month forecast of the natural variability anomalies, T_{nat}, of the Mean-G dataset (red curves in Fig. 10)

Contingency table for the detrended anomalies, T_{nat}	Forecasts			Total
	Below	Normal	Above	
Observations				
Below	**272**	77	9	358
Normal	102	**160**	90	352
Above	15	69	**250**	334
Total	389	306	349	**1044**

The 1044 months period (Jan 1931–Dec 2017) was used for verification. The climatological distribution was defined using the mean and standard deviation of the detrended series over that period ($\sigma_{\mathrm{clim}} = SD_T = 0.147$ °C)

The thresholds for the three equiprobable categories, above normal, near normal and below normal, will depend on the base-line of zero temperature and the standard deviation of the reference climatological distribution used. This will affect the distribution of events in the contingency table and consequently, the PC score obtained even though the forecast system has not changed. In that sense, the PC is a relative score. To avoid this dependence we could use absolute scores (independent of the climatology used), such as the ignorance score or the continuous ranked probability score (CRPS) (Hersbach 2000; Gneiting et al. 2005). The latter is the one we used in this paper for evaluating the quality of the probability forecasts of StocSIPS.

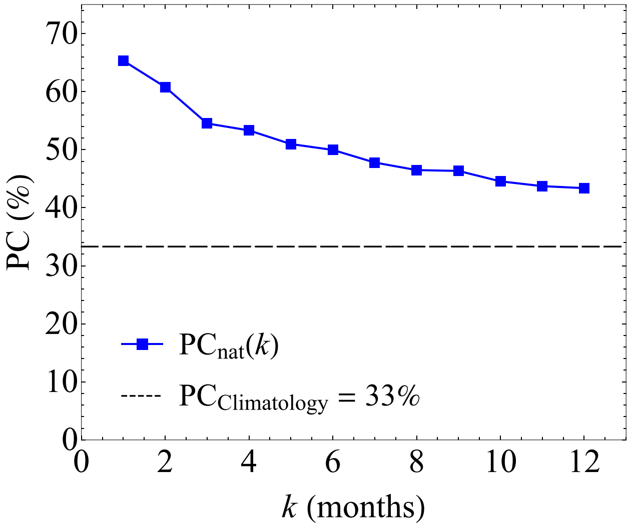

Fig. 14 PC as a function of k for the forecasts of the detrended anomalies, T_{nat} (blue line with squares in the figure). The dashed line at 33.3% is a reference showing the skill of the climatological forecast

The CRPS for a forecast initialized at time t with horizon k is defined as:

$$\mathrm{crps}(t+k) = \int_{-\infty}^{\infty} \left[P_f(t+k,x) - P_o(t+k,x)\right]^2 dx, \tag{40}$$

where $P_f(t,x)$ is the cumulative forecast distribution with mean $\mu_f = \hat{T}(t+k)$ given by Eq. (29) and standard deviation $\sigma_f(k) = \mathrm{RMSE}^m_{H,\sigma_\mathrm{T}}(k)$ and $P_o(t+k,x) = H\left[x - T_\mathrm{obs}(t+k)\right]$ is the cumulative observed distribution defined in terms of the Heaviside function $H(x)$. The CRPS can be determined for a single forecast, but a more accurate value is determined from a temporal average of many forecasts. The time mean CRPS as a function of horizon k is:

$$\mathrm{CRPS}(k) = \frac{1}{N-k+1}\sum_{t=0}^{N-k}\mathrm{crps}(t+k). \tag{41}$$

The CRPS is a negatively oriented measure of forecast accuracy, similar to the RMSE for deterministic ensemble mean forecasts; that is, smaller values indicate better skill. In fact, for deterministic forecasts, where $\sigma_f \to 0$, the crps in Eq. (40) reduces to the absolute error: $\mathrm{AE} = \left|T_\mathrm{obs} - \hat{T}\right|$. If we assume that P_f is the cumulative distribution function (CDF) of a normal distribution with mean μ_f and standard deviation σ_f, a closed form for crps can be derived by repeatedly integrating by parts in Eq. (40) (Gneiting et al. 2005):

$$\mathrm{crps}(t+k) = \sigma_f\left\{\frac{T_\mathrm{obs}-\mu_f}{\sigma_f}\left[2\Phi\left(\frac{T_\mathrm{obs}-\mu_f}{\sigma_f}\right)-1\right] + 2\varphi\left(\frac{T_\mathrm{obs}-\mu_f}{\sigma_f}\right) - \frac{1}{\sqrt{\pi}}\right\}, \tag{42}$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the PDF and the CDF, respectively, of the normal distribution with mean 0 and variance 1 evaluated at the normalized prediction error, $\varepsilon_n = \left(T_\mathrm{obs} - \mu_f\right)/\sigma_f$. This expression is very useful for obtaining the CRPS of large or many verification series and for calibrating ensemble forecasts from its optimization. In this paper, we will use it for deriving a general result that relates the CRPS with the RMSE of the ensemble mean of Gaussian probability forecasts.

Let us assume that the ensemble-mean forecast error, $\varepsilon = T_\mathrm{obs} - \mu_f$, follows a Gaussian distribution with zero mean and standard deviation σ_ε. Notice that $\sigma_f \neq \sigma_\varepsilon$; the former is given by the intra-ensemble spread, $\sigma_f = \sigma_\mathrm{ensemble}$, and the latter can be estimated from the RMSE between ensemble mean and observation. The CRPS and the RMSE can be related by averaging Eq. (42) for all possible values of the error, ε:

$$\langle \mathrm{crps}(t+k)\rangle_\varepsilon = \int_{-\infty}^{\infty}\varphi\left(\frac{\varepsilon}{\sigma_\varepsilon}\right)\mathrm{crps}(t+k)\,d\left(\frac{\varepsilon}{\sigma_\varepsilon}\right), \tag{43}$$

where $\varphi(\cdot)$ is defined as in Eq. (42). If we now replace Eq. (42) in Eq. (43) and integrate by parts, we obtain:

$$\langle \mathrm{crps}(t+k)\rangle_\varepsilon = \frac{\sigma_\varepsilon}{\sqrt{\pi}}\left[\sqrt{2\left(1+\sigma_f^2\big/\sigma_\varepsilon^2\right)} - \sigma_f\big/\sigma_\varepsilon\right]. \tag{44}$$

The average for all possible values of the error, $\langle\cdot\rangle_\varepsilon$, can be approximated by the time average, Eq. (41), for long enough verification periods. Moreover, we can approximate σ_f and σ_ε by their corresponding time-average estimates: $\sigma_f^2 \approx \overline{\sigma^2_\mathrm{ensemble}}$ and $\sigma_\varepsilon = \mathrm{RMSE}$. Using the definition of $\mathrm{ESS} = \overline{\sigma^2_\mathrm{ensemble}}/\mathrm{MSE}$ (Eq. (41)), we can finally rewrite Eq. (44) as:

$$\mathrm{CRPS}(k) = \frac{\mathrm{RMSE}(k)}{\sqrt{\pi}}\lambda(\mathrm{ESS}), \tag{45}$$

where $\lambda(\mathrm{ESS}) = \sqrt{2(1+\mathrm{ESS})} - \sqrt{\mathrm{ESS}}$. The function $\lambda(\mathrm{ESS})$ takes the minimum value $\lambda_\mathrm{min} = 1$ for a system with perfect reliability where $\mathrm{ESS} = 1$. For any other value of ESS, $\mathrm{CRPS} > \mathrm{RMSE}/\sqrt{\pi}$. This result shows that, for ensemble prediction systems, the optimal way of producing parametric probabilistic forecasts, assuming a Gaussian distribution, is by calculating the standard deviation of the forecast distribution from the hindcast period rather than just from the current forecast ensemble. This result agrees with previous studies (Kharin and Zwiers 2003; Kharin et al. 2009, 2017), which reach the same conclusion from the optimization of other standard probabilistic skill measures (e.g., the Brier skill score).

As we mentioned before, StocSIPS is a system with nearly perfect reliability and it assumes, by hypothesis, the Gaussianity of the errors. In that sense, the analytical expression for $\mathrm{RMSE}^m_{H,\sigma_\mathrm{T}}(k)$ (Eq. (32)) can be used to obtain a theoretical expression for $\mathrm{CRPS}(k)$ in Eq. (45). At the same time, the verification of this expression through a comparison between the values of $\mathrm{RMSE}(k)$ and $\mathrm{CRPS}(k)$ obtained from hindcasts can be used to check the validity of the model.

In Fig. 15 we show the time mean CRPS as a function of k, calculated in the verification period 1931–2017 for the probabilistic forecast of the monthly temperature anomalies of the Mean-G dataset. In the figure we show the results of the forecast of the raw anomalies (red circles), for which both the natural variability and the anthropogenic trend have to be forecast. Similarly to the previous results for the RMSE, the difference with the score of the forecast of the detrended anomalies is negligible ($\mathrm{CRPS}_\mathrm{raw} \approx \mathrm{CRPS}_\mathrm{nat}$), corresponding to the very small error on the projection of the trend compared to the error on the prediction of the detrended anomalies. The line in blue with empty squares, almost coincident with the red line, shows the function

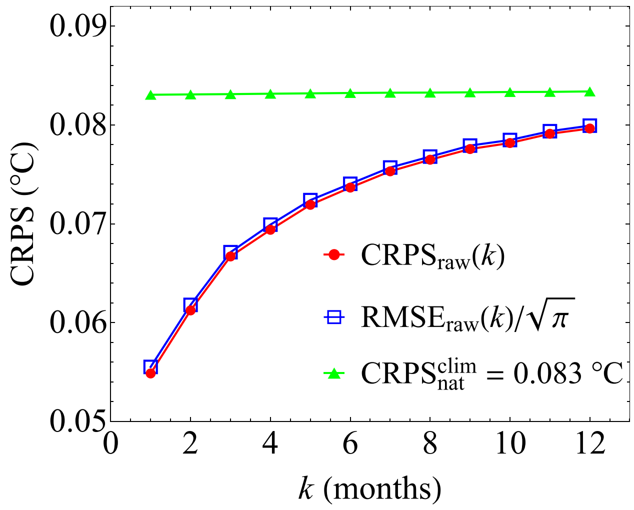

Fig. 15 CRPS as a function of the forecast horizon, k, calculated in the verification period 1931–2017 for the probabilistic forecast of the monthly temperature anomalies of the Mean-G dataset. In red circles we show the CRPS for the forecast of the raw anomalies, for which both the natural variability and the anthropogenic trend have to be forecast. The line in blue with squares, almost coincident with the red line, shows the function $\mathrm{RMSE}_{\mathrm{raw}}(k)/\sqrt{\pi}$, in perfect agreement with the theoretical prediction for the optimal value $\lambda_{\min} = 1$ in Eq. (45). In green triangles we included the CRPS of the reference climatology forecast of the detrended anomalies, $\mathrm{CRPS}_{\mathrm{nat}}^{\mathrm{clim}} = 0.083$ °C

$\mathrm{RMSE}_{\mathrm{raw}}(k)/\sqrt{\pi}$, in perfect agreement with the theoretical prediction for the optimal value $\lambda_{\min} = 1$ in Eq. (45), corresponding to perfect reliability. In the green triangles we included the CRPS of the reference climatology forecast of the natural variability component ($\mathrm{CRPS}_{\mathrm{nat}}^{\mathrm{clim}} = 0.083$ °C). That is, using the fixed climatological probability distribution (shown in grey in Fig. 13), with zero mean and standard deviation $\sigma_{\mathrm{clim}} = 0.147$ °C, to forecast the detrended anomalies. If we use the same climatological distribution for forecasting the raw anomalies, we obtain the much larger value $\mathrm{CRPS}_{\mathrm{raw}}^{\mathrm{clim}} = 0.181$ °C.

3.5 Comparison with GCMs

According to the World Meteorological Organization (WMO) (http://www.wmo.int/pages/prog/wcp/wcasp/gpc/gpc.php), there are currently fifteen major centers providing global seasonal forecasts. Thirteen of them have been officially designated by the WMO as Global Producing Centres for Long-Range Forecasts (GPCLRFs). The Meteorological Service of Canada (MSC) contributes with the Canadian Seasonal to Interannual Prediction System (CanSIPS) (Merryfield et al. 2011, 2013).

CanSIPS is a multi-model ensemble (MME) system using 10 members from each of two climate models (CanCM3 and CanCM4) developed by the Canadian Centre for Climate Modelling and Analysis (CCCma) for a total ensemble size of 20 realizations. It is a fully coupled atmosphere–ocean–ice–land prediction system relying on operational data assimilation for the initial state of the atmosphere, sea surface temperature and sea ice.

To evaluate forecasts and compare StocSIPS with CanSIPS, we accessed the publicly available series of hindcasts of CanSIPS covering the period 1981–2010 (CanSIPS 2016). The fields, available on 145 × 73 latitude–longitude grids at resolutions of 2.5° × 2.5° for each of the 20 ensemble members, were area-weight averaged to obtain global mean series of hindcasts at monthly resolution. CanSIPS produces forecast at the beginning of every month for the average value of that month and the next 11 months; i.e. for lead times from 0 to 11 months for each initialization date. In our case, that would correspond to forecast horizons (number of periods ahead that are forecasted) from 1 to 12 months. In the verification for $k = 1$ month (lead zero), the hindcast period is January 1981–December 2010; for $k = 2$ months (lead one), the hindcast period is February 1981–January 2011, and so on. This way, all the 12 series of hindcasts (one for each horizon) have a length of 360 months.

An optimal use of the dynamical model can be obtained after advanced postprocessing and calibration to reduce the bias of the model (Crochemore et al. 2016; Kharin et al. 2017; Van Schaeybroeck and Vannitsem 2018; Pasternack et al. 2018). We do not pretend here to make an exhaustive use of these calibration techniques. To keep the comparison simple, we followed the postprocessing for CanSIPS described in Sects. 3.a and 3.b of (Kharin et al. 2017) for deterministic and parametric probability forecasts, respectively. The statistical adjustment used by the authors is based on a linear rescaling of the ensemble mean and standard deviation of the anomaly forecast. The regression coefficients are obtained by minimizing the MSE and CRPS of the ensemble forecast in some verification period.

It can be easily shown that, after the recalibration, their method will lead to the optimal expression for CRPS given by Eq. (45) when ESS = 1: $\mathrm{CRPS} = \mathrm{RMSE}/\sqrt{\pi}$. The recalibration method can be reduced to using—as optimal deterministic predictor—the projection of the ensemble mean that minimizes the MSE in some verification period. Then, for the probability distribution forecast, the standard deviation is made equal to the RMSE of the adjusted deterministic forecast instead of calculating it from the intra-ensemble spread. In that sense, the ensemble members are only useful for obtaining the ensemble mean. They do not contribute further to the forecast as the optimal probabilistic scores are obtained from the condition ESS = 1.

In their paper, (Kharin et al. 2017) also show that the optimal average skill scores are obtained when time-invariant (independent of the season) coefficients are used. We will use this result here and, instead of using only 30 years for estimating individual coefficients for each month, we use the monthly series to estimate constant coefficients based on 360 months that only depend on the lead time. These coefficients are more stable and do not significantly degrade the accuracy of the forecast due to sampling errors as would season-dependent coefficients.

In Fig. 16, we show an example of a forecast for the 12 months following April 1982 for both StocSIPS and CanSIPS. In red we show the verification curve of observations for the Mean-G dataset. In blue, the median hindcasts for StocSIPS, with the corresponding 95% confidence interval based on the RMSE for the verification period. The ensemble mean for CanSIPS is shown in black, with each of the 20 members shown in dashed light colors and the 95% confidence interval based on the RMSE of the hindcasts represented in grey. The CO_2eq trend for the Mean-G dataset (green line) was added as a reference of the long-term equilibrium of the temperature fluctuations.

As expected, the dispersion of the different ensemble members for the dynamical model increases as the horizon increases, which shows the stochastic-like character of GCMs for long-term predictions with the consequent loss in skill. Despite this increase in the spread of the ensemble, the dynamical model is underdispersive for all horizons. The ESS (see Eq. (39) in Sect. 3.4.4) is in the range 0.57–0.74 for all lead times, except for zero months lead time where ESS = 0.40. (Kharin et al. 2017) show that inflating the ensemble spread to satisfy the condition ESS = 1, results in more conservative estimates for the forecast probabilities of the three categories and improved reliability of the probability forecast and overall probabilistic skill scores.

3.5.1 Deterministic forecast comparison and seasonality

In this section we present scores for the deterministic forecast (ensemble mean forecast) for both models using for verification the Mean-G dataset in the period 1981–2010. In all cases we used the calibrated ensemble mean for CanSIPS, unless stated otherwise. In Fig. 17, we show density plots of the RMSE as a function of the forecast horizon and the initialization month for StocSIPS and CanSIPS [panels (a) and (b), respectively]. For both models, there is a seasonality pattern with large errors during the Boreal winter months. In the case of StocSIPS, the largest values of the RMSE are found for February, January and March, in that order, while CanSIPS has the largest errors for the forecasts of November and February. In Fig. 17c, we show the difference between CanSIPS RMSE and StocSIPS RMSE; positive values indicate that StocSIPS has better skill. StocSIPS outperforms CanSIPS for most of the horizons and initialization months, except for the forecasts of January and February and some other initialization dates for $k = 1$ month. The overall values of RMSE vs. k—averaging for all the months in the verification period independently of the initialization date—are shown in Fig. 17d. The curve for StocSIPS is represented in red line with solid squares. For CanSIPS, we show in solid blue line with empty squares the RMSE for the calibrated

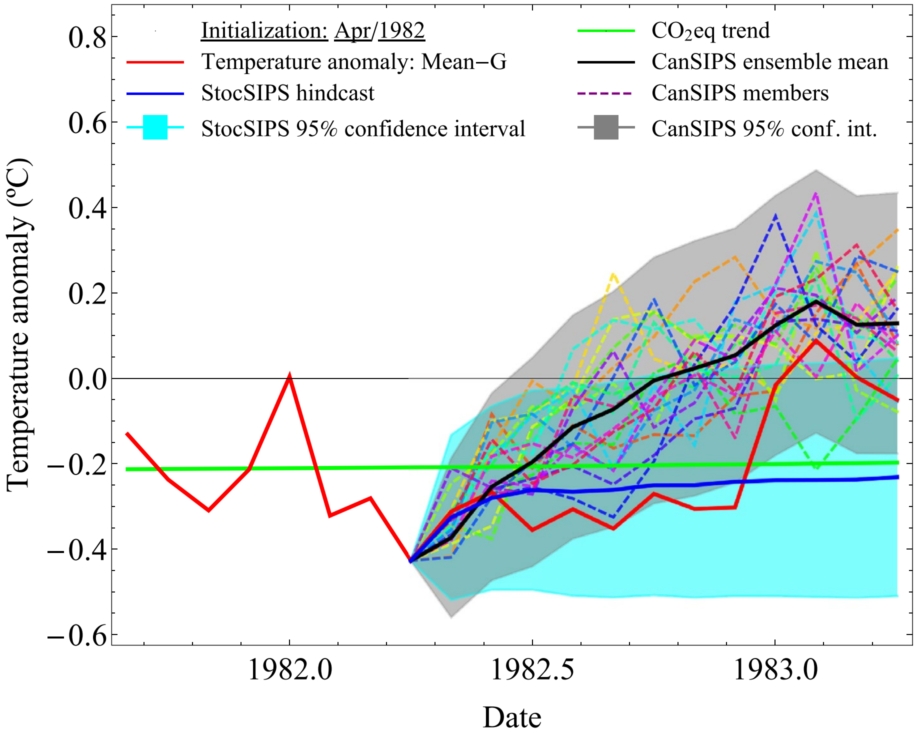

Fig. 16 One example of forecast for the 12 months following April 1982 for both StocSIPS and CanSIPS. In red we show the verification curve of observations for the Mean-G dataset. In blue, the median hindcasts for StocSIPS, with the corresponding 95% confidence interval based on the RMSE for the verification period. The ensemble mean for CanSIPS is shown in black, with each of the 20 members shown in dashed light colors and the 95% confidence interval based on the RMSE of the hindcasts represented in grey. The CO_2eq trend for the Mean-G dataset (green line) was added as a reference of the long-term equilibrium of the temperature fluctuations

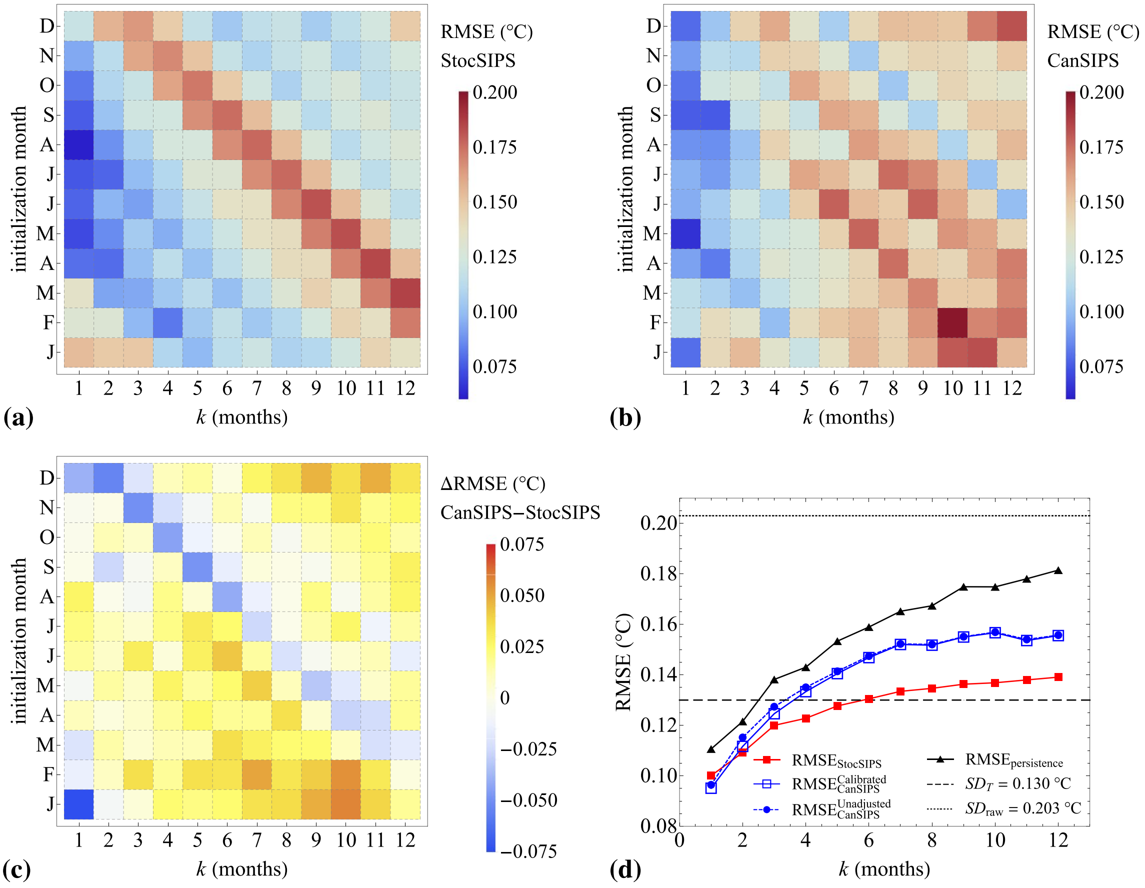

Fig. 17 Density plots of the RMSE as a function of the forecast horizon, k, and the initialization month for StocSIPS and CanSIPS (**a**, **b**, respectively). For both models, there is a seasonality pattern with large errors during the Boreal winter months. In **c**, we show the difference between CanSIPS and StocSIPS RMSE; positive values indicate that StocSIPS has better skill. StocSIPS outperforms CanSIPS for most of the horizons and initialization months, except for the forecasts of January and February and some other initialization dates for $k = 1$ month. The overall values of RMSE vs. k—averaging for all the months in the verification period independently of the initialization date—are shown in **d**. The curve for StocSIPS is represented in red squares. For CanSIPS, we show in solid blue line with empty squares the scores for the calibrated ensemble mean and in dashed blue line with solid circles the RMSE for the unadjusted model. We can see that the improvement in the RMSE due to the recalibration is very small. We included, for comparison, the curve obtained from hindcasts using persistence (black-triangles). The standard deviations for the detrended and for the raw series in the verification period were also included for reference (SD_T and SD_{raw}, respectively)

ensemble mean and in dashed blue line with solid circles the values for the unadjusted model. We can see that the improvement in the RMSE due to the recalibration is very small. We included, for comparison, the curves obtained from hindcasts using persistence (black-triangles). That is, for horizon k, assuming that the temperature k months into the future is predicted by the present value. The standard deviations for the detrended and for the raw series in the verification period were also included for reference (SD_T and SD_{raw}, respectively).

Similar results are reported in Fig. 18 for the MSSS and for the ACC. From the density plots [panels (a) and (c)] we can reach the same conclusion based on these scores: StocSIPS is better than CanSIPS for most of the horizons and initialization months, except for the forecasts of January and February. In panels (b) and (d), we show the all-months average scores without considering the initialization dates. The results for StocSIPS are shown in red line with solid squares and for CanSIPS in blue line with circles. In the MSSS graphs, we only show the results for the calibrated model. For the ACC, as the calibration for CanSIPS is just a rescaling of the ensemble mean, the correlations with or without the calibration are the same. The curves obtained from hindcasts using persistence were also included for comparison (black-triangles).

For the MSSS, we choose the climatology as reference forecast with $\mathrm{MSE}_{\mathrm{ref}} = SD_{\mathrm{raw}}^2$ being the variance of the raw series. We use accordingly the notation $\mathrm{MSSS} = \mathrm{MSSS}_{\mathrm{raw}}$. The horizontal line (green empty squares) included in the graph represents the value of skill obtained by

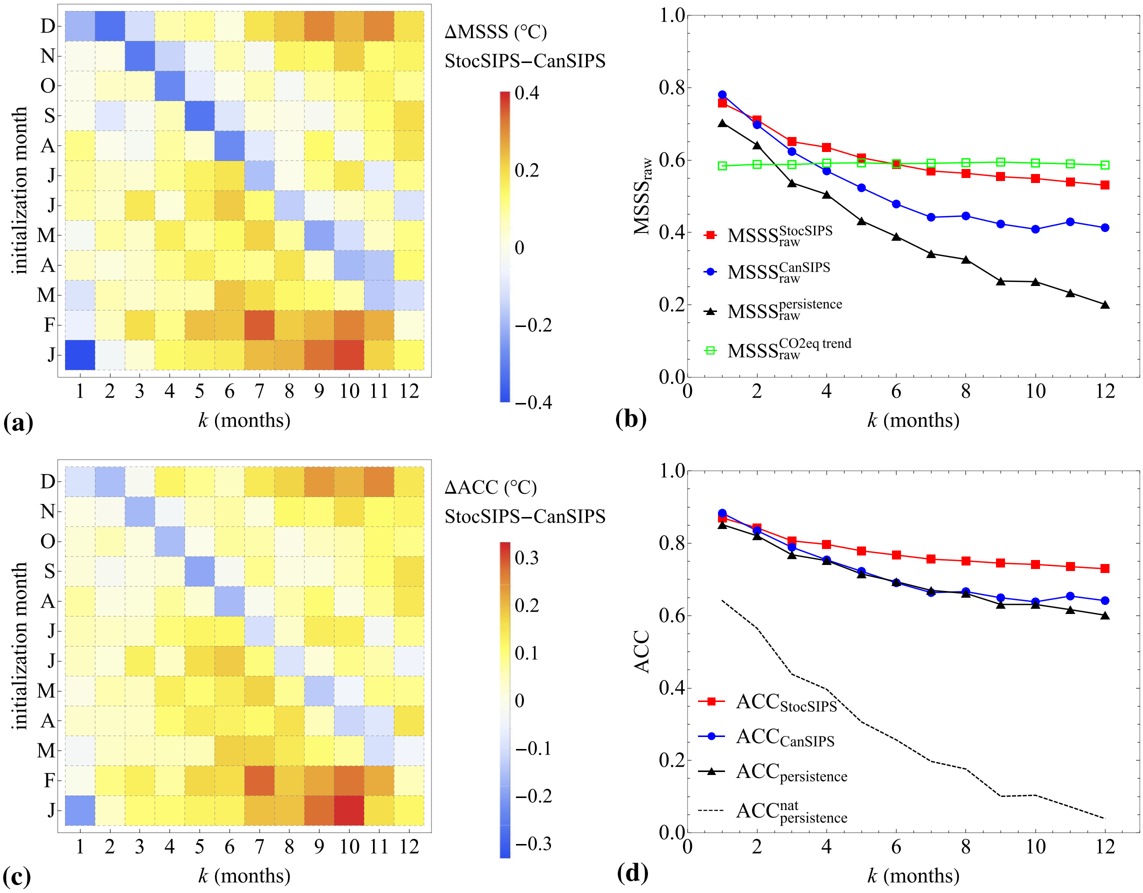

Fig. 18 Density plots for the MSSS and for the ACC (**a**, **c**, respectively) as a function of the forecast horizon and the initialization date. The positive values indicate that StocSIPS is better than CanSIPS for most of the horizons and initialization months, except for the forecasts of January and February. In **b**, **d**, we show the all-months average scores without considering the initialization dates. The results for StocSIPS are shown in red line with solid squares and for CanSIPS in blue line with circles. In the MSSS graphs, we only show the results for the calibrated model. The horizontal line (green line with empty squares) included in the graph represents the value of skill obtained by projecting the CO_2eq trend with respect to the climatological forecast. For the ACC, as the calibration for CanSIPS is just a rescaling of the ensemble mean, the correlations with or without the calibration are the same. The curves obtained from hindcasts using persistence were also included for comparison (black-triangles). The autocorrelation function for the detrended series (natural variability component), which is the same as the ACC for the forecast of that series using persistence, was included for comparison as a dashed black curve ($ACC_{persistence}^{nat}$ in the figure)

projecting the CO_2eq trend with respect to the climatological forecast. The MSSS can be easily computed as $MSSS_{raw}^{CO2eq\ trend} = 1 - SD_T^2/SD_{raw}^2$ (≈ 0.59 for the Mean-G dataset) because the errors of the forecast would be the amplitude of the detrended anomalies. The values obtained using this equation do not vary significantly for different horizons in the period analyzed. The extra contribution in the skill for StocSIPS comes from the forecast of the natural variability component.

The ACC, in the case of persistence, is the same as the autocorrelation function with lag k of the reference series. As mentioned before, the values obtained for the ACC (even for the poor persistence forecasts), are spuriously high due to the anthropogenic trends superimposed on the series. Many authors report similarly high values without taking this fact into consideration. More realistic values would be obtained for the forecast of the detrended series, but there is no impartial way of removing the anthropogenic component for CanSIPS. The anthropogenic forcing is an intrinsic part of the GCM and to have a prediction of the natural variability only, we would have to remove its contribution before running the dynamical model. The autocorrelation function for the detrended series (natural variability component), which is the same as the ACC for the forecast of that series using persistence, was included for comparison as a dashed black curve ($ACC_{persistence}^{nat}$ in the figure).

With respect to the comparison of the two models for the deterministic forecast, the conclusion is clear: StocSIPS presents better skill than CanSIPS on average for all the measures used and for all horizons except for $k = 1$ month, where CanSIPS is slightly better. This was expected as, for the case of GCMs, 1 month is still close to the deterministic predictability limit imposed by the chaotic behavior of the system (~ 10 days for the atmosphere and 1–2 years for the ocean). After 1 month, the relative advantage of StocSIPS increases as the horizon increases. The reduced skill of StocSIPS for January and February are related to the intrinsic seasonality of the globally-averaged temperature. In future work, this seasonality in the variability could be removed by pre-processing, presumably resulting in further error reduction.

3.5.2 Probabilistic forecast comparison

In the previous section we showed how the two systems (CanSIPS and StocSIPS) compare for deterministic forecasts where the scores only depend on the ensemble mean. In Fig. 17d, we showed that the reduction in the RMSE of CanSIPS due to the recalibration is very small. In this section we show how this improvement is more noticeable if probabilistic scoring rules are used, as they are influenced not only by the ensemble mean, but also by the ensemble spread which is readjusted to maximize the CRPS using the condition ESS = 1 mentioned before.

Examples of probabilistic forecasts for July 1994 are shown in Fig. 19 for StocSIPS (left) and for CanSIPS (right) for horizon $k = 2$ months (one month lead time; i.e. using data up to May 1994). The normal PDF in grey represents the climatological distribution of the monthly temperatures for the Mean-G dataset for the verification period 1981–2010. The terciles of the climatological distribution are indicated by vertical dashed lines. These vertical lines define three equiprobable categories of above normal, near normal, and below normal monthly temperatures observed in the verification period. In the left, the forecast distribution for StocSIPS is indicated by the black curve with the forecast mean $\mu_f = \hat{T}(\text{July 1994}) = -0.105$ °C and standard deviation $\sigma_f = \text{RMSE}_{\text{StocSIPS}} = 0.109$ °C for $k = 2$ months. In the right, the distribution in dashed black line represents the unadjusted forecast of CanSIPS for $k = 2$ months with parameters $\mu_f = -0.051$ °C (ensemble mean) and $\sigma_f = \sigma_{\text{ensemble}} = 0.084$ °C (intra-ensemble standard deviation). The calibrated forecast PDF for CanSIPS is shown in solid black in the right panel. The adjusted mean for this distribution is $\mu_f = -0.062$ °C and the inflated standard deviation $\sigma_f = \text{RMSE}_{\text{CanSIPS}}^{\text{Calibrated}} = 0.112$ °C. The areas under the forecast PDF's in different colors indicate probabilities of below normal (blue), near normal (yellow), and above normal (pink) temperatures. These probabilities are summarized in the top-left corner as bar plots. The climatological probability of 33% is indicated by the horizontal dashed line. The observed temperature for that specific date, $T_{\text{obs}} = -0.127$ °C, is represented by the vertical green line.

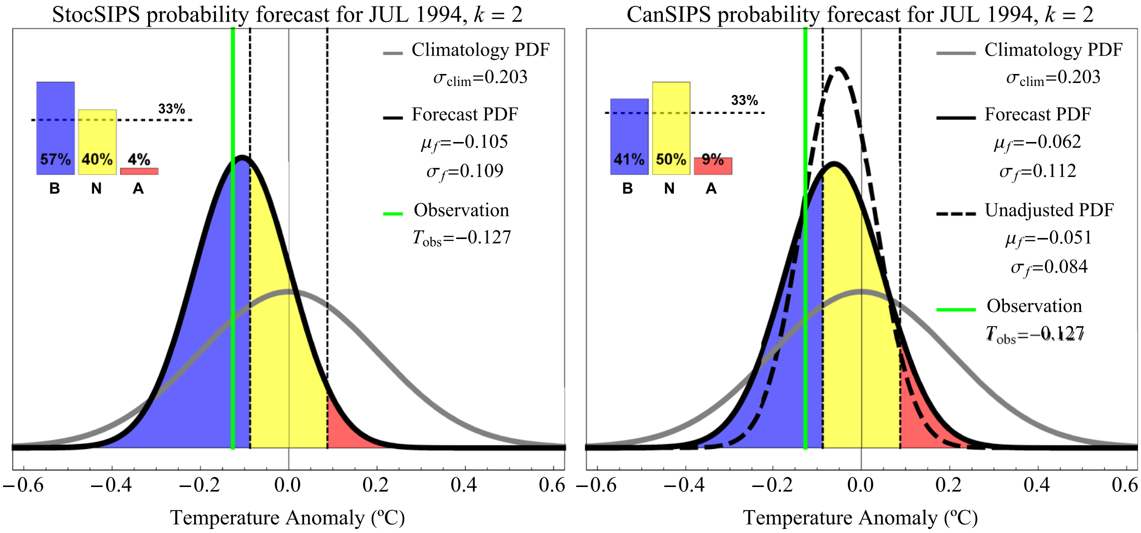

Fig. 19 Examples of probabilistic forecasts for July 1994 are shown for StocSIPS (left) and for CanSIPS (right) for horizon $k = 2$ months (1 month lead time; i.e. using data up to May 1994). The normal probability density function in grey represents the climatological distribution of the monthly temperatures for the Mean-G dataset for the verification period 1981–2010. The terciles of the climatological distribution are indicated by vertical dashed lines. The colored areas under the forecast density function are proportional to the forecast probabilities for each category: below normal (blue), near normal (yellow) and above normal (pink). These probabilities are summarized in the top-left corner as bar plots. The climatological probability of 33% is indicated by the horizontal dashed line. The observed temperature for that specific date, $T_{\text{obs}} = -0.127$ °C, is represented by the vertical green line. In the right, the distribution in dashed black line represent the unadjusted forecast of CanSIPS for $k = 2$ months and the calibrated forecast PDF is shown in solid black. The parameters for all the distributions are included in the legends

Table 3 Contingency table for StocSIPS 3 category probabilistic forecasts (below normal, near normal and above normal) for the detrended series (T_{nat}, red curves in Fig. 10) of the Mean-G dataset with zero months lead time (k = 1 month)

Contingency table for the detrended anomalies, T_{nat}	Forecasts			Total
	Below	Normal	Above	
Observations				
Below	**83**	25	12	120
Normal	39	**39**	40	118
Above	12	27	**83**	122
Total	134	91	135	**360**

The verification period is January 1981–December 2010 (360 months). The number of hits and total number of events are shown in bold in the main diagonal. Here we use the climatology obtained from the detrended anomalies with $\sigma_{\mathrm{clim}} = SD_T = 0.130$ °C

For the unadjusted distribution of CanSIPS, the standard deviation for each specific month and lead time is estimated from the intra-ensemble spread and, as the model is underdispersive, it is generally lower than the standard deviation of the calibrated forecast distribution, which is estimated from the whole verification period and is constant for all months for a particular lead time.

The combined contingency table for the forecasts of StocSIPS (grey rows) and CanSIPS (white rows with the values of the unadjusted forecast in parenthesis) for $k = 1$ month is shown in Table 2. For observational reference we used the Mean-G dataset for verification in the period January 1981–December 2010 (360 months). The number of hits and total number of events are shown in bold in the main diagonal.

The reduced number of observation events in the near-normal category is a consequence of the deviation from Gaussianity of the undetrended anomalies in the verification period 1981–2010. Specifically, there is a reduced kurtosis caused by the presence of the anthropogenic trend, as can be clearly seen in Fig. 5. The distribution of the detrended anomalies, T_{nat}, is much closer to a Gaussian (see Appendix 2). In Table 3, we show the contingency table for the forecast of this series using StocSIPS. Now the total number of observations are almost equally distributed among the three categories obtained using the climatological distribution based on the detrended series.

From the diagonal elements in Table 2 we get the following PC scores for $k = 1$ month: for StocSIPS, $\mathrm{PC}_{\mathrm{StocSIPS}} \approx 78\%$ and for CanSIPS we get $\mathrm{PC}_{\mathrm{CanSIPS}}^{\mathrm{Calibrated}} \approx 76\%$ and $\mathrm{PC}_{\mathrm{CanSIPS}}^{\mathrm{Unadjusted}} \approx 74\%$ for the calibrated and the unadjusted forecasts, respectively. These values are spuriously high due to the presence of the trend in the raw series. Just from direct inspection of the reference series (red curve in Fig. 5), by projecting the trend we could predict that most of the temperature values in the decade 2001–2010 would fall in the above normal category, while most of the events in the decade 1981–2000 would fall in the below normal category. The PC score obtained from Table 3 for the forecast of the natural variability component with $k = 1$ month using StocSIPS is more realistic: $\mathrm{PC}_{\mathrm{StocSIPS}}^{\mathrm{Nat}} \approx 57\%$. As we mentioned before, we cannot perform a similar forecast using CanSIPS since the anthropogenic forcing is an intrinsic part of the GCM. To obtain a prediction of the natural variability only, we would have to remove its contribution before running the dynamical model.

The PC scores for all horizons from $k = 1$ to 12 months are shown in Fig. 20. In blue squares we show the PC scores for StocSIPS and in red circles and green triangles for CanSIPS, calibrated and unadjusted forecasts, respectively. The solid black line shows the skill of StocSIPS for the forecast of the detrended series. The values obtained in this case are lower than those obtained for the raw anomalies. Those

Table 2 Contingency table for 3 category probabilistic forecasts (below normal, near normal and above normal) for the raw (undetrended) Mean-G dataset with zero months lead time (k = 1 month)

Combined contingency table for the forecasts of StocSIPS (grey rows) and CanSIPS (white rows) for $k = 1$ month.			Forecast			Total
			Below	Normal	Above	
Observations	Below	StocSIPS	**102**	25	1	128
		CanSIPS (Unadjusted)	**99 (95)**	28 (33)	1 (0)	
	Normal	StocSIPS	23	**61**	18	102
		CanSIPS (Unadjusted)	20 (15)	**69 (74)**	13 (13)	
	Above	StocSIPS	2	11	**117**	130
		CanSIPS (Unadjusted)	0 (0)	24 (31)	**106 (99)**	
Total		StocSIPS	127	97	136	**360**
		CanSIPS (Unadjusted)	119 (110)	121 (138)	120 (112)	

The verification period is January 1981–December 2010 (360 months). The number of hits and total number of events are shown in bold in the main diagonal. Here we compacted in one table the results for the forecasts of StocSIPS (grey rows) and CanSIPS (white rows with the values of the unadjusted forecast in parentheses)

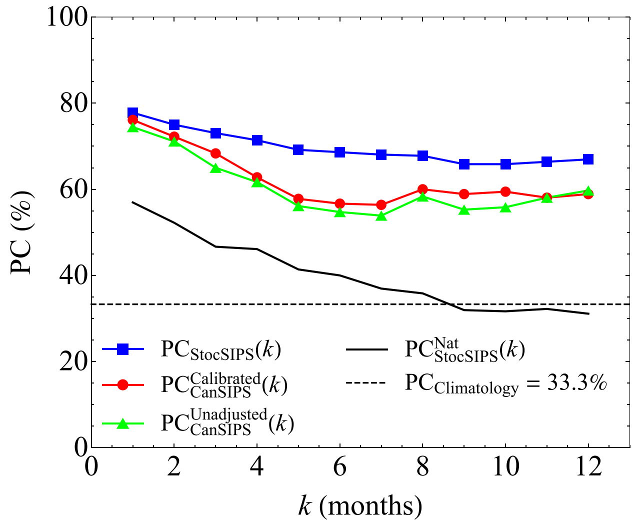

Fig. 20 PC as a function of k for StocSIPS (blue squares) and for CanSIPS, calibrated and unadjusted forecasts in red circles and green triangles, respectively. The solid black line shows the skill of StocSIPS for the forecast of the detrended series. The dashed line at 33.3% is a reference showing the skill of the climatological forecast

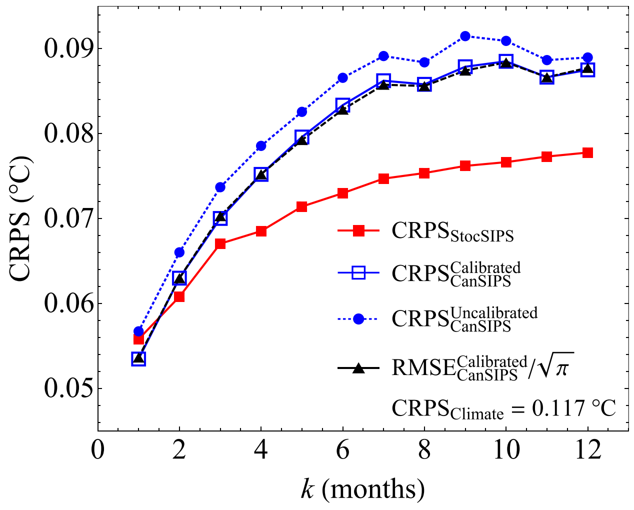

Fig. 21 CRPS vs. k for both models in the verification period 1981–2010 for the Mean-G dataset. In red, we show the CRPS for StocSIPS and in blue for CanSIPS with dotted line and solid circles for the unadjusted forecast and solid line with open squares for the calibrated forecast. The function $\mathrm{RMSE}^{\mathrm{Calibrated}}_{\mathrm{CanSIPS}}(k)/\sqrt{\pi}$ is shown in dashed black line with triangles. There is perfect agreement between these optimal values and the CRPS of CanSIPS after the calibration, in correspondence with Eq. (45). The score for the climatological forecast was included in the legend ($\mathrm{CRPS}_{\mathrm{Climate}} = 0.117$ °C)

values are a better measure of the actual quality of the forecasting system since the spurious effects of the trend are removed. The dashed line at 33.3% is a reference showing the skill of the climatological forecast.

Three main conclusions can be obtained from the analysis of Fig. 20. First, there is an improvement on the probabilistic forecast skill of CanSIPS thanks to the recalibration. This improvement is small but is more noticeable than the one obtained for the deterministic scores (e.g. RMSE, MSSS). Second, StocSIPS performs better than CanSIPS for all lead times and the relative advantage increases with the forecast horizon up to $k = 7$ months. Finally, from the comparison of the blue and the solid black curves for the StocSIPS forecasts of the raw and the detrended series, respectively, we can notice that most of the skill comes from the projection of the trend and for $k > 8$ months this is the only source of skill.

Although the PC score for StocSIPS is larger for all horizons, it is difficult to evaluate the relative advantage over the probabilistic CanSIPS forecasts based on that score alone. The PC is influenced by the climatological distribution used for defining the categories and mainly by the presence of the trend. A more realistic comparison should be based in absolute scores that only depend on the forecast system and are independent of the base-line or the climatology chosen. The dependence of the CRPS with the forecast horizon is shown in Fig. 21 for both models in the verification period 1981–2010 for the Mean-G dataset. In red, we show the CRPS for StocSIPS and in blue for CanSIPS with dotted line and solid circles for the unadjusted forecast and solid line with open squares for the calibrated forecast. The function $\mathrm{RMSE}^{\mathrm{Calibrated}}_{\mathrm{CanSIPS}}(k)/\sqrt{\pi}$ is shown in dashed black line with triangles. There is perfect agreement between these optimal values and the CRPS of CanSIPS after the calibration, in correspondence with Eq. (45). The score for the climatological forecast was included in the legend for reference ($\mathrm{CRPS}_{\mathrm{Climate}} = 0.117$ °C).

If we compare Fig. 21 with Fig. 17d, we can see that the effect of the calibration of the CanSIPS output is more noticeable for the CRPS than for the RMSE. The probabilistic forecast gains from both the inflation of the standard deviation and the scaling of the ensemble mean, while only the latter influences the deterministic forecast. After the adjustment, CanSIPS forecast is better for zero months lead time, but for the rest of the forecast horizons StocSIPS shows more skill. The relative advantage of the stochastic model over the GCM increases the further we forecast into the future. For the first month, the numerical model forecast still falls in the deterministic predictability limit.

4 Discussion

Over the last decades, conventional numerical approaches have developed to the point where they are now skillful at lead times that approach their theoretical (deterministic) predictability limits—itself close to the lifetimes of planetary structures (about 10 days). This threshold is due to the nonlinearity and complexity of the equations of atmospheric dynamics and their sensitive dependence on initial conditions (butterfly effect) (Lorenz 1963, 1972), and it cannot be overcome using purely deterministic models, not even by using combinations of deterministic-stochastic approaches such as recent stochastic parameterization models (Berner et al. 2017). In the macroweather regime (from 10 days to decades), GCMs become stochastic: the model integrations are extended far beyond their predictability limits producing "random" outputs that are finally averaged to obtain the forecast as the model ensemble mean.

The convergence of the dynamical models to their own climate follows from the macroweather property of internal fluctuations to decrease with time scale (see Fig. 6 for the case of natural variability—including volcanic and solar forcings). This scaling behavior with negative fluctuation exponent is present in real data and in GCM control runs, so the statistics of conventional numerical models' variability is of similar type to that found in the real-world temperature series. The main problem is that each GCM converges to its own model climate, which is different from the actual climate. Also, the models cannot fully reproduce the actual high frequency weather noise even if the statistics of the noise they generate is similar to the real-world one.

In that sense, the SLIMM model, developed in Lovejoy et al. (2015), uses real data to generate the high-frequency noise with the correct statistical symmetries for the fluctuations and with a realistic climate. The main characteristics of SLIMM were summarized in Sect. 2.1. In this paper we presented the Stochastic Seasonal to Interannual Prediction System (StocSIPS), which includes SLIMM as the core model to forecast the natural variability component of the temperature field. StocSIPS also represents a more general framework for modelling the seasonality and the anthropogenic trend and the possible inclusion of other atmospheric fields at different temporal and spatial resolutions. In this sense, StocSIPS is the general system and SLIMM is the main part of it dedicated to the modelling of the stationary scaling series.

StocSIPS is based on some statistical properties of the macroweather regime such as: the Gaussianity of temperature fluctuations (as justified in Appendix 2) and the temporal scaling symmetry of the natural variability with negative fluctuation exponents, as shown in Sect. 3.2. It also assumes the independence between the high frequency natural variability of the temperature field and the low frequency component dominated by anthropogenic effects. The anthropogenic component is represented as a short memory linear response to equivalent CO_2 forcing. The natural variability component is modeled and predicted using the stochastic approach originally proposed in SLIMM.

The scaling of the fluctuations implies that there are power-law decorrelations in the system and hence a large memory effect that can be exploited. The simplest stochastic model that includes both the Gaussianity and the scaling of the fluctuations is the fGn process. The Gaussian statistics of the temperature natural variability fluctuations allowed us to use the mean square prediction framework to build an optimal conditional expectation predictor based on a linear combination of past data.

In Sects. 2 and 2.1 we discuss how fGn can be obtained in SLIMM as the solution of a fractional order differential equation, which in turn is a generalization of the integer order stochastic differential equation in LIM models. The fractional derivative is introduced to account for the large memory effect given by the power-law behavior of the correlation function, in contrast, integer order derivatives imply short memory autoregressive moving average processes with asymptotic exponential decorrelations. The fractional differential equation can be obtained as the high frequency limit of a fractional energy balance equation in which the usual (exponential) temperature relaxation to equilibrium is replaced by power-law relaxation (work in progress). The main characteristics of SLIMM are summarized in Sect. 2.1, including the formal expression for the predictor as an integral of innovations going an infinite time into the past. Physically, the source of the long-range memory is energy stored in ocean gyres, eddies, at depth, or over land, in ice, soil moisture, etc.

The original technique that was used to make the SLIMM forecasts was basically correct, but it made several approximations (such as that the amount of data available for the forecast was infinite) and it was numerically cumbersome. Most of this work was dedicated to improving the mathematical treatment and the numerical techniques of SLIMM and validate them on ten different global temperature series since 1880 (five globally averaged and five over land).

The main improvement included in StocSIPS for the prediction of temperature series is the application of discrete-in-time fGn to obtain an optimal predictor based on a finite amount of past data. In Sect. 2.2.1 we give the theoretical expressions for the predictor coefficients and the skill as functions of the fluctuation exponent alone. This represents an advantage over other autoregressive models (AR, ARMA) which do not include fractional integrations that account for the long-term memory and hence do not consider the information from the distant past. An additional limitation of these approaches is that, in order to predict, the autocorrelation function for each time lag, $C(\Delta t)$, must be estimated directly from the data. Each $C(\Delta t)$ will have its own sampling error, this effectively introduces a large "noise" in the predictor estimates and a large computational cost if many coefficients are needed. In our fGn model the coefficients have an analytic expression which only depends on the fluctuation exponent, H, obtained directly from the data exploiting the scale-invariance symmetry of the fluctuations;

our problem is a statistically highly constrained problem of parametric estimation (H), not an unconstrained one (the entire $C(\Delta t)$ function).

Other technical details of discrete-in-time fGn models are given in Appendix 1. We discuss how to produce exact realizations of fGn processes with a given length, N and family of parameters σ, μ and H. The inverse process of obtaining those parameters for a given time series is also discussed. Other important results shown in Appendix 1 are an algorithm called quasi maximum likelihood estimation (QMLE) for obtaining the parameter H, and the derivation of some ergodic properties of fGn processes. The QMLE method is slightly less accurate—but much more efficient computationally—than the usual maximum likelihood method. It has the advantage of being part of the verification process as it minimizes the mean square error of the hindcasts. The ergodicity of the variance of the process, expressed in Eq. (62), besides proving the convergence of the temporal average estimate of the variance to the ensemble variance, also shows that this convergence is ultra slow for values of H close to zero. This fact implies a strong dependence of the value of the resulting skill score with the length of the hindcast series used for verification. It could potentially impact statistical methods that depend on the covariance matrix, e.g. empirical orthogonal functions (EOF) and empirical mode decomposition (EMD).

The main result of this work is the application of StocSIPS to the modeling and forecasting of global temperature series. With that purpose, we selected the five major observation-based global temperature data series which are in common use (see Sect. 3.1).

Over the last century, low frequencies are dominated by anthropogenic effects and after 10–20 years the scaling regime changes from a negative to a positive value of H (see Fig. 6). The anthropogenic component was modelled as a linear response to equivalent CO_2 forcing and removed. The residual natural variability component was then modeled and predicted using the theory presented in Sect. 2 and Appendix 1. The quality of the fit of the fGn model to the real data was evaluated in detail in Appendix 2.

To validate our model, we produced series of hindcasts for the period 1931–2017 with forecast horizons from 1 to 12 months. These series were stratified to obtain the dependence of the forecast skill with the forecast horizon and the initialization time. The RMSE of the hindcasts was lower than the standard deviation of the verification series for all horizons, showing positive skill. The values obtained for the all-month average results were in good agreement with the theoretical predictions. Other skill scores, such as the MSSS and the ACC were obtained.

StocSIPS source of predictability is based on the strong long-range correlations present in the temperature time series. In that sense, there is no source of skill coming from interannual variations since the model assumes that the seasonality, as well as the low frequency trend in the raw data, are deterministic. Theoretically, we should not expect a dependence of the skill on the initialization time. However, the stratification of the data shows that there is a multiplicative seasonality effect that makes the variability different for each individual month (see Fig. 11). The standard deviation of the temperature for the Boreal winter months is considerably larger than for the rest. This affects the skill of StocSIPS for those months and is a discrepancy with respect to the stationarity hypothesis. In future work, we could compensate for this effect through preprocessing of the time series and study the implications in StocSIPS forecast skill.

In Sect. 3.4.4 we showed how to make parametric probability forecasts using StocSIPS. For a prediction system with Gaussian errors, we derived a theoretical relation between the deterministic score RMSE and the probabilistic CRPS. We also showed that StocSIPS is—by definition—a nearly perfectly reliable system and that this theoretical relation is satisfied by the verification results.

Finally, in Sect. 3.5 we compared StocSIPS with the Canadian Seasonal to Interannual Prediction System (CanSIPS), which is one of the GCMs contributing to the Long-Range Forecast project of the World Meteorological Organization. Deterministic and probabilistic forecast skill scores for StocSIPS and for the CanSIPS were compared for the verification period 1981–2010.

The main conclusion is that, for the overall forecast including all the months in the verification period and without considering different initialization times, StocSIPS has higher skill than CanSIPS for all the metrics used and for all horizons except for $k = 1$ month, where CanSIPS is slightly better. This was not surprising since for GCMs, 1 month is still close to the deterministic predictability threshold imposed by the chaotic behavior of the system (~ 10 days for the atmosphere and 1–2 years for the ocean). Beyond 1 month, the relative advantage of StocSIPS increases as the horizon increases. The seasonal stratification of the verification shows that, due to the interannual variability, CanSIPS performs better than StocSIPS for the forecasts of January and February. For other months (beyond zero months lead times) StocSIPS has better skill.

5 Conclusions

In this paper we presented the Stochastic Seasonal to Interannual Prediction System (StocSIPS), which is based on some statistical properties of the macroweather regime such as: the Gaussianity of temperature fluctuations and the temporal scaling symmetry of the natural variability. StocSIPS includes SLIMM as the core model to forecast the natural variability component of the temperature field. Here we improved the theory and numerical methods of SLIMM for its direct application to macroweather forecast.

In summary, StocSIPS models the temperature series as a superposition of a periodic signal corresponding to the annual cycle, a low frequency deterministic trend from anthropogenic forcings and a high frequency stochastic natural variability component. The annual cycle can be estimated directly from the data and is assumed constant in the future, at least for horizons of a few years. The anthropogenic component is represented as a linear response to equivalent CO_2 forcing and can be projected very accurately 1 year into the future by using two parameters, the climate sensitivity and an offset, which can be obtained from linear regression given historical emissions. Finally, the natural variability is modeled as a discrete-in-time fGn process which is completely determined by the variance and the fluctuation exponent. That gives a total of only four parameters for modeling and predicting the temperature series. Those parameters are quite stable and can be estimated with good accuracy from past data.

The comparison with CanSIPS validates StocSIPS as a good alternative and a complementary approach to conventional numerical models. The reason is that whereas CanSIPS and StocSIPS have the same type of statistical variability around the climate state, the CanSIPS model climate is different from the real-world climate. In comparison, StocSIPS uses historical data to force the forecast to the real-world climate. From a forecast point of view, in general, GCMs can be seen as an initial value problem for generating many "stochastic" realizations of the state of the atmosphere, while StocSIPS is effectively a past value problem that directly estimates the most probable future state.

The prediction of global average temperature series presented in this paper is based on some symmetries of the macroweather regime: scale-invariance and low intermittency (rough Gaussianity). In a future paper (currently in preparation), we show how another macroweather symmetry, the statistical space time factorization (Lovejoy and de Lima 2015), can be included to extend the application of StocSIPS to temperature forecasts at a regional level with any arbitrary spatial resolution without need for downscaling. Another future application of StocSIPS that can be derived from this work is that, due to its qualitatively different approach with respect to traditional GCMs, it is possible to combine CanSIPS and StocSIPS into a single hybrid forecasting system that improves on both, especially at zero lead times. We have already obtained some predictions with the combined model, "CanStoc", and we are currently working on a future publication on these results. We are also working on the application of StocSIPS to the forecast of GCMs preindustrial control runs to show that they satisfy the same macroweather symmetries as real-world data and hence, together with their deterministic predictability limits, there are also stochastic predictability limits applicable to GCMs. These limits correspond to the maximum possible skill that can be achieved by a stochastic Gaussian scaling system with a given scaling exponent (measure of the memory and the predictability in the data).

In May 2016, we created the website: http://www.physics.mcgill.ca/StocSIPS/, where global average and regional temperature forecasts at monthly, seasonal and annual resolutions using StocSIPS are published on a regular basis.

Appendix 1: Simulation, parameters estimation, ergodicity and model adequacy

Simulation

When modeling real time series and testing numerical algorithms, it is often useful to obtain synthetic realizations of fGn processes. There are many methods for simulating approximate samples of fGn, e.g.: (1) type 1 (Mandelbrot and Wallis 1969), (2) type 2 (Mandelbrot and Wallis 1969), (3) fast fGn (Mandelbrot 1971), (4) filtered fGn (Matalas and Wallis 1971), (5) ARMA(1,1) (O'Connell 1974), (6) broken line (Garcia et al. 1972; Mejia et al. 1972; Rodriguez-Iturbe et al. 1972; Mandelbrot 1972), (7) ARMA-Markov models (Lettenmaier and Burges 1977) and some approximate, more efficient, recent methods (Paxson 1997; Jeong et al. 2003). We can choose among these methods based on their strengths and weaknesses, depending on the specific application we need.

Nevertheless, instead of using short memory approximations for simulating fGn, it is possible to generate exact realizations by applying the following procedure (Hipel and McLeod 1994; Palma 2007). In Eq. (20) we gave the MA representation of our series for any time, t, based on the knowledge of an infinite past of innovations, $\left\{\gamma_{t-j}\right\}_{j=1,\ldots,\infty}$ with $\gamma_t \sim \mathrm{NID}(0,1)$ and $\langle\gamma_i\gamma_j\rangle = \delta_{ij}$. If we want a series with specific length, N, mean μ, variance σ_T^2 and fluctuation exponent H, we can work in a similar way as we did with the AR representation for obtaining the predictor. By replacing the coefficients, φ_j, we could write instead the finite sum:

$$T_t = \mu + \sum_{j=1}^{t} m_{tj}\gamma_{t+1-j} = \mu + m_{t1}\gamma_t + \cdots + m_{tt}\gamma_t, \tag{46}$$

for $t = 1, \ldots, N$, where the optimal coefficients m_{ij} are the elements of the lower triangular matrix $\mathbf{M}_{H,\sigma_T}^N$ given by the Cholesky decomposition of the autocovariance matrix, $\mathbf{R}_{H,\sigma_T}^N = \left[C_{H,\sigma_T}(i-j)\right]_{i,j=1,\ldots,N}$; that is:

$$\mathbf{R}_{H,\sigma_T}^N = \mathbf{M}_{H,\sigma_T}^N \left(\mathbf{M}_{H,\sigma_T}^N\right)^T, \tag{47}$$

with $m_{ij} = 0$ for $j > i$. In summary, for obtaining an fGn realization of length N, we need to generate a white-noise process $\left\{\gamma_t\right\}_{t=1,\dots,N}$ with an appropriate method, obtain the autocovariance matrix $\mathbf{R}^N_{H,\sigma_T}$ using Eq. (7.iii), then get $\mathbf{M}^N_{H,\sigma_T}$ from the Cholesky decomposition of $\mathbf{R}^N_{H,\sigma_T}$, and finally apply Eq. (46) for every t to obtain our $\left\{T_t\right\}$ series. The variables T_t will be $\mathrm{NID}\left(\mu, \sigma_T^2\right)$ and the process will have fluctuation exponent H in the interval $(-1, 0)$.

Maximum likelihood estimation

If instead of simulating an fGn process, we are interested in the opposite operation of finding the parameters that best fit a given time series, the most accurate method to use is based on maximizing the log-likelihood function (Hipel and McLeod 1994). Suppose that we have our vector $\mathbf{T}_N = \left[T_1, \dots, T_N\right]^T$ that represents a stationary Gaussian process. Then the log-likelihood function of this process is given by:

$$\mathfrak{L}\left(\mu, \sigma_T, H\right) = -\frac{1}{2}\log\left[\det\left(\mathbf{R}^N_{H,\sigma_T}\right)\right] - \frac{1}{2}\tilde{\mathbf{T}}^T_{N,\mu}\left(\mathbf{R}^N_{H,\sigma_T}\right)^{-1}\tilde{\mathbf{T}}_{N,\mu} \tag{48}$$

where $\tilde{\mathbf{T}}_{N,\mu} = \left[T_1 - \mu, \dots, T_N - \mu\right]^T$ is a vector formed by our original series after removing the mean.

For fixed H, the maximum likelihood estimators (MLE) of μ and σ_T are:

$$\hat{\mu} = \frac{\mathbf{1}^T_N\left(\tilde{\mathbf{R}}^N_H\right)^{-1}\mathbf{T}_N}{\mathbf{1}^T_N\left(\tilde{\mathbf{R}}^N_H\right)^{-1}\mathbf{1}_N} \tag{49}$$

and

$$\hat{\sigma}_T^2 = \frac{1}{N}\tilde{\mathbf{T}}^T_{N,\hat{\mu}}\left(\tilde{\mathbf{R}}^N_H\right)^{-1}\tilde{\mathbf{T}}_{N,\hat{\mu}}, \tag{50}$$

where $\mathbf{1}_N = [1, 1, \dots, 1]^T$ is an $N \times 1$ vector with all the elements equal to 1 and $\tilde{\mathbf{R}}^N_H = \mathbf{R}^N_{H,\sigma_T}/\sigma_T^2$ is the autocorrelation matrix, which only depends on H.

Substituting these values into Eq. (48), we obtain the maximized log-likelihood function of H:

$$\mathfrak{L}_{\max}(H) = -\frac{1}{2}\log\left[\det\left(\tilde{\mathbf{R}}^N_H\right)\right] - \frac{N}{2}\log\left[\frac{1}{N}\tilde{\mathbf{T}}^T_{N,\hat{\mu}}\left(\tilde{\mathbf{R}}^N_H\right)^{-1}\tilde{\mathbf{T}}_{N,\hat{\mu}}\right]. \tag{51}$$

The estimate for the fluctuation exponent, $\hat{H}_l$, is obtained by maximizing $\mathcal{L}_{\max}(H)$ and can be used then to obtain $\hat{\mu}$ and $\hat{\sigma}_T^2$ using Eqs. (49) and (50).

Ergodicity

It is worth noticing here that $\hat{\mu}$ and $\hat{\sigma}_T^2$ are estimates of the ensemble mean $\mu = \langle T_t\rangle$ and variance $\sigma_T^2 = \left\langle\left(T_t - \mu\right)^2\right\rangle$ of the fGn process, respectively (see Sect. 2.1). If we try to estimate these parameters based on temporal averages of a single realization, some differences may arise with the values obtained using Eqs. (49) and (50). To explain these differences, we briefly discuss some ergodic properties of fGn processes.

Let

$$\overline{T}_N = \frac{\sum_{t=1}^N T_t}{N} \tag{52}$$

and

$$SD_T^2 = \frac{\sum_{t=1}^N\left(T_t - \overline{T}_N\right)^2}{N} = \overline{\left(T_N - \mu\right)^2} - \left(\overline{T}_N - \mu\right)^2 \tag{53}$$

be the temporal average estimates of the mean and the variance of our process, respectively (the overbar indicates temporal averaging, N is considered large here), SD indicates "standard deviation".

Using the relationship between fBm and fGn (Eq. (5)), we can write the temperature as:

$$T_t = \sigma_T\left[B_{H'}(t) - B_{H'}(t-1)\right]. \tag{54}$$

The fBm process has the following properties:

$$\begin{aligned}
&\text{(i) } B_{H'}(t) \text{ is Gaussian with stationary increments;}\\
&\text{(ii) } \langle B_{H'}(t)\rangle = \mu t/\sigma_T \text{ for all } t\\
&\qquad\text{(the notation } \langle\cdot\rangle \text{ denotes ensemble averaging)}\\
&\text{(iii) } C_{B_{H'}}(t,s) = \left\langle\left[B_{H'}(t) - \mu t/\sigma_T\right]\left[B_{H'}(s) - \mu s/\sigma_T\right]\right\rangle\\
&\qquad = \left(|t|^{2H'} + |s|^{2H'} - |t-s|^{2H'}\right)/2
\end{aligned} \tag{55}$$

Usually, the condition $B_{H'}(0) = 0$ is added to this definition. Using this and Eq. (54), by telescopic sum all addends cancel except for the last one and we obtain:

$$\overline{T}_N = \frac{1}{N}\sigma_T B_{H'}(N). \tag{56}$$

Taking ensemble averages and using Eqs. (55) (ii) and (iii) we get:

$$\left\langle\overline{T}_N\right\rangle = \mu \tag{57}$$

and

$$\left\langle\left(\overline{T}_N - \mu\right)^2\right\rangle = \frac{1}{N^2}\sigma_T^2\left\langle\left[B_{H'}(N) - \mu N/\sigma_T\right]^2\right\rangle = \sigma_T^2 N^{2H}, \tag{58}$$

where we replaced $H' = H + 1$.

Table 4 Average estimates of H for 200 realizations of simulated fGn with length $N = 1656$ and parameters $\mu = 0$, $\sigma_T = 1$ and H corresponding to the values in the first column

H	$\hat{H}_q$	$\hat{H}_l$	$\hat{H}_{l,\mathrm{ens}}$	$\hat{H}_h$	$\hat{H}_{h,\mathrm{ens}}$	$\hat{H}_s$	$\hat{H}_{s,ens}$	$\hat{\sigma}_T$	SD_T	$\frac{SD_T}{\sqrt{1-N^{2H}}}$
– 0.45	– 0.45 (0.02)	– 0.45 (0.02)	– 0.45	– 0.48 (0.07)	– 0.45	– 0.51 (0.06)	– 0.44	1.00	1.00	1.00
– 0.40	– 0.40 (0.01)	– 0.40 (0.01)	– 0.40	– 0.42 (0.07)	– 0.40	– 0.45 (0.05)	– 0.39	1.00	1.00	1.00
– 0.35	– 0.35 (0.02)	– 0.35 (0.02)	– 0.35	– 0.37 (0.07)	– 0.35	– 0.40 (0.06)	– 0.33	1.00	1.00	1.00
– 0.30	– 0.30 (0.02)	– 0.30 (0.02)	– 0.30	– 0.34 (0.08)	– 0.30	– 0.35 (0.06)	– 0.28	1.00	0.99	1.00
– 0.25	– 0.26 (0.02)	– 0.25 (0.02)	– 0.25	– 0.28 (0.08)	– 0.25	– 0.29 (0.05)	– 0.24	1.00	0.99	1.00
– 0.20	– 0.21 (0.02)	– 0.20 (0.02)	– 0.20	– 0.24 (0.08)	– 0.20	– 0.24 (0.06)	– 0.18	1.00	0.97	1.00
– 0.15	– 0.17 (0.02)	– 0.15 (0.02)	– 0.15	– 0.18 (0.09)	– 0.15	– 0.19 (0.06)	– 0.12	0.99	0.94	1.00
– 0.10	– 0.12 (0.02)	– 0.10 (0.02)	– 0.10	– 0.12 (0.07)	– 0.10	– 0.13 (0.05)	– 0.07	1.00	0.88	1.00
– 0.05	– 0.08 (0.01)	– 0.06 (0.02)	– 0.05	– 0.08 (0.08)	– 0.05	– 0.09 (0.06)	– 0.02	0.98	0.71	0.99

The values in parentheses represent the standard deviation for each ensemble. The following methods were used: QMLE ($\hat{H}_q$), MLE ($\hat{H}_l$), Haar fluctuations ($\hat{H}_h$) and spectral analysis ($\hat{H}_s$). For these last three methods, direct ensemble estimates were also obtained ($\hat{H}_{\_,\mathrm{ens}}$); $\hat{H}_\_$ could be seen as the mean of all the $\hat{H}_\_$'s while $\hat{H}_{\_,\mathrm{ens}}$ is the $\hat{H}_\_$ of the mean. The last three columns show the average estimates $\hat{\sigma}_T$, SD_T and the confirmation of their relationship given by Eq. (62)

Table 5 Values of the parameters obtained for the ten datasets and the corresponding mean series for global and for land

Dataset	$\hat{H}_l$	$\hat{H}_q$	$\hat{H}_h$	$\hat{H}_s$	$\hat{\sigma}_T$	SD_T	$\frac{SD_T}{\sqrt{1-N^{2H}}}$	$\lambda_{2\times\mathrm{CO}_2\mathrm{eq}}$	T_0
NASA	– 0.08	– 0.10	– 0.11 (0.02)	– 0.08 (0.04)	0.183	0.155	0.184	2.10 (0.03)	– 0.391 (0.006)
NOAA	– 0.06	– 0.09	– 0.06 (0.02)	– 0.03 (0.04)	0.183	0.144	0.187	2.00 (0.02)	– 0.372 (0.006)
HAD4	– 0.07	– 0.08	– 0.06 (0.02)	– 0.10 (0.06)	0.194	0.159	0.201	1.89 (0.03)	– 0.353 (0.006)
CowW	– 0.09	– 0.10	– 0.09 (0.03)	– 0.10 (0.05)	0.183	0.163	0.193	1.98 (0.03)	– 0.369 (0.006)
Berk	– 0.08	– 0.09	– 0.07 (0.02)	– 0.12 (0.07)	0.197	0.174	0.209	2.20 (0.03)	– 0.410 (0.007)
Mean-G	– 0.06	– 0.08	– 0.08 (0.02)	– 0.10 (0.06)	0.195	0.153	0.199	2.03 (0.03)	– 0.379 (0.006)
NASA-L	– 0.25	– 0.24	– 0.21 (0.02)	– 0.29 (0.04)	0.373	0.371	0.376	2.96 (0.06)	– 0.551 (0.015)
NOAA-L	– 0.25	– 0.25	– 0.24 (0.02)	– 0.27 (0.03)	0.331	0.325	0.329	2.95 (0.05)	– 0.550 (0.013)
HAD4-L	– 0.18	– 0.19	– 0.19 (0.02)	– 0.24 (0.04)	0.297	0.285	0.295	2.70 (0.05)	– 0.503 (0.011)
CowW-L	– 0.22	– 0.22	– 0.18 (0.03)	– 0.27 (0.04)	0.337	0.333	0.339	2.84 (0.06)	– 0.529 (0.013)
Berk-L	– 0.23	– 0.23	– 0.21 (0.02)	– 0.25 (0.03)	0.348	0.342	0.349	2.81 (0.06)	– 0.523 (0.014)
Mean-L	– 0.22	– 0.22	– 0.20 (0.02)	– 0.26 (0.04)	0.327	0.321	0.327	2.85 (0.05)	– 0.531 (0.013)

From left to right we have estimates of H using the following methods: MLE ($\hat{H}_l$), QMLE ($\hat{H}_q$), Haar fluctuations ($\hat{H}_h$) and spectral analysis ($\hat{H}_s$); estimate of the standard deviation of the ensemble using MLE ($\hat{\sigma}_T$); amplitude of each series ignoring the correlations (SD_T); confirmation of the relationship between $\hat{\sigma}_T$ and SD_T given by Eq. (25); the climate sensitivity and offset used to remove the anthropogenic trend, $\lambda_{2\times\mathrm{CO}_2\mathrm{eq}}$ and T_0, respectively (Eq. (27)). Uncertainty estimates are given in parentheses

Consequently, since the process $B_{H'}(t)$ is Gaussian, we conclude that, the temporal average estimate of the mean satisfies:

$$\overline{T}_N \sim \mathrm{N}\left(\mu, \sigma_T^2 N^{2H}\right). \tag{59}$$

Now, taking the ensemble average of Eq. (53), we get:

$$\left\langle SD_T^2 \right\rangle = \left\langle \overline{\left(T_N - \mu\right)^2} \right\rangle - \left\langle \left(\overline{T}_N - \mu\right)^2 \right\rangle. \tag{60}$$

The ensemble and the time averaging operations commute in the first term of the right-hand side of Eq. (60):

$$\left\langle \overline{\left(T_N - \mu\right)^2} \right\rangle = \overline{\left\langle \left(T_N - \mu\right)^2 \right\rangle} = \sigma_T^2. \tag{61}$$

Using this and Eq. (58) for the last term in Eq. (60), we finally get:

$$SD_T^2 = \sigma_T^2\left(1 - N^{2H}\right), \tag{62}$$

meaning that the temporal average SD_T is a biased estimate of the variance of the process, σ_T^2. An unbiased estimate would then be $SD_T^2/\left(1 - N^{2H}\right)$. The variance of this estimator is more difficult to obtain. Its derivation, together with potential applications for treating climate series, will be presented in a future paper (currently in preparation).

In the limit $N \to \infty$, as $-1 < H < 0$, we have $SD_T^2 \to \sigma_T^2$, meaning that the process is ergodic (the temporal average and the ensemble average coincide for infinitely long series). Nevertheless, for $H \to 0$ this convergence is very slow, and a very long series would be needed in order to estimate the variance of the process from the sample variance without any correction. For example, for $H = -0.1$ and $N = 1656$ months = 138 years (realistic values for globally-averaged temperatures, see Sect. 3), we have $SD_T^2/\sigma_T^2 = \left(1 - N^{2H}\right) = 0.772$, i.e. a 23% difference between both estimates. In the same sense, if we want to estimate σ_T^2 from the sample variance with 95% accuracy, we would need a series with $N = 3.2 \cdot 10^6$ (if N is in months that would be $N = 266{,}667$ years!). The last three columns of Table 4 show the average estimates $\hat{\sigma}_T = \sqrt{\hat{\sigma}_T^2}$ (Eq. (50)), SD_T (Eq. (53)) and the confirmation of their relationship (Eq. (62)), for simulations of fGn with length $N = 1656$ and parameters $\mu = 0, \sigma_T = 1$ and values of H in the range $(-1/2, 0)$. In each case, 200 realizations were analyzed, but only the average values of the estimates are shown. The standard deviations are always 2–7% of the respective mean values and were not reported. Notice that the difference between $\hat{\sigma}_T$ and SD_T increases as H goes close to zero and the memory effects become more important.

Let us return now to the estimates $\hat{\mu}$ and $\hat{\sigma}_T^2$ given by Eqs. (49) and (50), respectively. These ensemble estimates are still obtained from the information of only one finite series, $\mathbf{T}_N = \left[T_1, \ldots, T_N\right]^T$, but the presence of the correlation matrix, $\tilde{\mathbf{R}}_H^N$, automatically includes all the information from the infinite unknown past. If we make $\tilde{\mathbf{R}}_H^N = \mathbf{I}_N$ ($\mathbf{I}_N$ is the $N \times N$ identity matrix) in Eqs. (49) and (50) (or equivalently $H = -1/2$), we obtain:

$$\hat{\mu} = \frac{\mathbf{1}_N^T \mathbf{T}_N}{\mathbf{1}_N^T \mathbf{1}_N} = \frac{\sum_{t=1}^N T_t}{N} = \overline{T}_N \tag{63}$$

and

$$\hat{\sigma}_T^2 = \frac{1}{N}\tilde{\mathbf{T}}_{N,\hat{\mu}}^T \tilde{\mathbf{T}}_{N,\hat{\mu}} = \frac{\sum_{t=1}^N \left(T_t - \hat{\mu}\right)^2}{N} = SD_T^2. \tag{64}$$

This means that the temporal average estimates based on one realization of the process are only valid for uncorrelated process, for which the ensemble and the sample averages are equal. When both correlations and memory effects are present, this information must be considered. In the case of fGn processes, the memory effects are introduced by including the correlation matrix which only depends on the fluctuation exponent H. The value of this parameter for the process can also be obtained from only one realization of the same as shown below.

Quasi-maximum-likelihood estimation for *H*

As we mentioned before, the MLE for the fluctuation exponent, $\hat{H}_l$, is obtained by maximizing $\mathcal{L}_{\max}(H)$ (Eq. (51)). The process of optimization of $\mathcal{L}_{\max}(H)$ could easily be computationally expensive for large values of N. To avoid this, many approximate methods have been developed. We can use Eq. (9) to obtain $\hat{H}_s = \left(\beta_l - 1\right)/2$ from the spectral exponent at low frequencies. This method, as well as the Haar wavelet analysis to obtain an estimate $\hat{H}_h$ from the exponent of the Haar fluctuations, was used in Lovejoy and Schertzer (2013) and Lovejoy et al. (2015) to obtain estimates of H for average global and Northern Hemisphere anomalies. These two methods depend on the range selected for the linear regression and, when the graphs are noisy, it could result in poor estimates of the exponents. They, nevertheless, have the advantage of being more general; they yield H estimates even for highly nonGaussian processes. In the present case, a more accurate approximation is based on quasi-maximum-likelihood estimates (QMLE) from autoregressive approximations (Palma 2007).

Suppose we have a series of N observations, $\left\{T_t\right\}_{t=1,\ldots,N}$, we can build the one-step predictor for T_t, $\hat{T}_t^p(1)$ from Eq. (22) using a memory of p steps in the past with $p + 1 < t \leq N$:

$$\hat{T}_t^p(1) = \sum_{j=-p}^{0} \phi_{p,j}(k) T_{t+j-1} = \phi_{p,-p}(k) T_{t-p-1} + \cdots + \phi_{p,0}(k) T_{t-1}. \tag{65}$$

Then, the approximate QMLE, $\hat{H}_q$, is obtained by minimizing the function

$$\begin{aligned}\mathfrak{L}_1(H) &= \sum_{t=p+2}^{N} \left[T_t - \hat{T}_t^p(1)\right]^2 \\ &= \sum_{t=p+2}^{N} \left[T_t - \phi_{p,-p}(1) T_{t-p-1} - \cdots - \phi_{p,0}(1) T_{t-1}\right]^2.\end{aligned} \tag{66}$$

Remember that the coefficients $\phi_{p,j}$ only depend on H. An added advantage of this method is that, by construction, it is done as part of the verification process based on hindcasts. The actual mean square error (MSE) of our one-step predictor with memory p is $\mathcal{L}_1(H)/(N - p - 1)$, so in practice, we perform the one-step hindcasts for different values of H in the specified range and select the value that gives the minimum MSE. The computation of the coefficients ϕ_{pj} is fast, since we do not need to take very large values of p to achieve nearly the asymptotic skill, as we showed in Sect. 2.2.1.

In order to compare these different estimation methods, we performed some numerical experiments. By using Eq. (46) for the exact method with parameters $\mu = 0$ and $\sigma_T = 1$, we generated fGn ensembles of one hundred members of length $N = 1656$ (see Sect. 3) for each value of $H \in \{-0.45, -0.40, -0.35, -0.30, -0.25, -0.20, -0.15, -0.10, -0.05\}$. Then, we estimated H from the four previously mentioned methods for each realization. The results are summarized in Table 4. The values in parentheses represent the standard deviations for each ensemble. The maximum likelihood, the Haar fluctuation and the spectral methods allow for direct estimates of the ensemble values (shown with the subscript "ens" in Table 4) by considering the maximum likelihood of the vector process, the ensemble of all the fluctuations or the average of all the spectra, respectively from all the paths instead of from each of the series independently. We could say, for example that $\hat{H}_s$ is the mean of all the $\hat{H}_s$'s obtained from each realization spectrum, while $\hat{H}_{s,\text{ens}}$ is the value obtained from the mean of all the spectra. This ensemble estimate reduces the error due to dispersion of each of the ensemble members. For the QMLE, a memory $p = 20$ was used.

As we can see from Table 4, for the MLE method, there is good agreement between the average of the estimates for each realization and the direct ensemble estimate. This is not the case for the less accurate methods of Haar fluctuation and spectral analysis in the member-by-member cases. Comparatively, the standard deviation of these two methods (without considering the estimation error for each specific realization) is much larger than for the MLE. Nevertheless, the ensemble estimates for the Haar are very accurate because the dispersion for the ensemble is much lower than for each individual graph. In practice, it is almost always the case that we only have a given time series to analyze instead of multiple realizations of an ensemble. In that sense, unless we have more theoretical or empirical justifications for the scaling, estimates based on these graphical methods should be considered cautiously.

A direct comparison of the second and third columns in Table 4 shows the accuracy of the QMLE method if we take the MLE as reference. The average values and the standard deviations for the two methods are very close for small values of H, but as we move to values close to zero there is a systematic bias in the QMLE method towards slightly smaller values than those obtained with the MLE. Nevertheless, the presence of this bias is of little consequence from the point of view of forecasting and can be reduced by increasing the memory used. As we mentioned before, the QMLE method is based on minimizing the MSE—or what is the same—maximizing the MSSS obtained from hindcasts. Near the extreme, a small variation of the value of H used to perform the forecast will produce almost no change on the MSSS obtained.

Model adequacy

The final step after finding the parameters μ, σ_T^2 and H, is to check the adequacy of the fitted model to the data. Imagine we have a time series $\left\{T_t\right\}_{t=1,\ldots,N}$. The residuals of our fGn model are obtained from inverting Eq. (46) and calculating the vector

$$\mathbf{e}_N = \left(\mathbf{M}_{H,\sigma_T}^N\right)^{-1}\tilde{\mathbf{T}}_{N,\mu}. \tag{67}$$

If the model provides a good description of the data, the elements of the residual vector $\mathbf{e}_N = \left[e_1, \ldots, e_N\right]^T$ should be white noise, i.e. they should be NID(0, 1) with autocorrelation function $\left\langle e_i e_j \right\rangle = \delta_{ij}$. Many statistical tests for whiteness of $\left\{e_i\right\}$ could be performed, the more descriptive one being based on the examination of the graph of the residual autocorrelation function (RACF). The RACF at lag l is calculated as:

$$r_l\left(\mathbf{e}_N\right) = \frac{\sum_{t=1}^{N-l} e_i e_{i+l}}{\sum_{t=1}^{N} e_i^2}. \tag{68}$$

Asymptotically, $r_l\left(\mathbf{e}_N\right) \sim \text{NID}(0, 1/N)$ for any lag $l \geq 1$ and $r_0\left(\mathbf{e}_N\right) = 1$. In the graph of $r_l\left(\mathbf{e}_N\right)$ vs. l, there should not be any point significantly far outside the 95% confidence interval given by the horizontal lines $\pm 1.96/\sqrt{N}$, and the number of points outside this range, should represent around 5% of the total number of points. As additional tests, we could verify that the estimates of the fluctuation exponent of $\left\{e_i\right\}$, using the previous graphical methods, are $\hat{H}_s \approx \hat{H}_h \approx -0.5$, which is the value for white noise as a particular case of fGn. The less important Gaussianity assumption could also be verified by visualizing the empirical probability distribution against a normal distribution and checking for the presence of extremes.

Appendix 2: Checking the fGn model fit to global temperature data

In Table 5 we show the values of the parameters obtained for the ten datasets and the corresponding mean series for the globe and for land:

As we can see in Table 5, there is relatively good agreement between the more robust estimates of the fluctuation exponent, $\hat{H}_l$ and $\hat{H}_q$ (see Appendix 1 for the notation), with the small bias of $\hat{H}_q$ towards smaller values (we used a memory $p = 20$ months for estimating $\hat{H}_q$). The estimates $\hat{H}_h$ and $\hat{H}_s$, obtained using the general methods, also roughly agree with the MLE and QMLE considering their relatively wide one-standard deviation confidence interval (given in parentheses in Table 5). Notice the difference between the parameter $\hat{\sigma}_T$ and the amplitude of each series, SD_T. The former is an unbiased estimate of the standard deviation for the ensemble process using maximum likelihood, while the latter is a biased estimate, where the bias is because of the limited time series and autocorrelated samples (see Ergodicity in Appendix 1). We also include the values of $SD_T/\sqrt{1 - N^{2H}}$ for confirmation of Eq. (25) ($N = 1656$

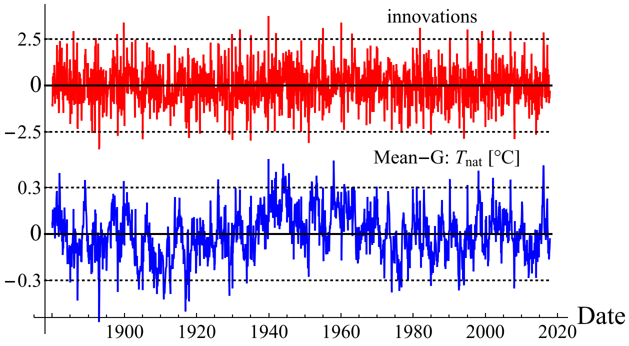

Fig. 22 Natural variability component for the Mean-G dataset, together with its corresponding series of residual innovations, $\{e_i\}$, obtained using Eq. (67). The units for the T_{nat} series are °C, while the innovations are unitless

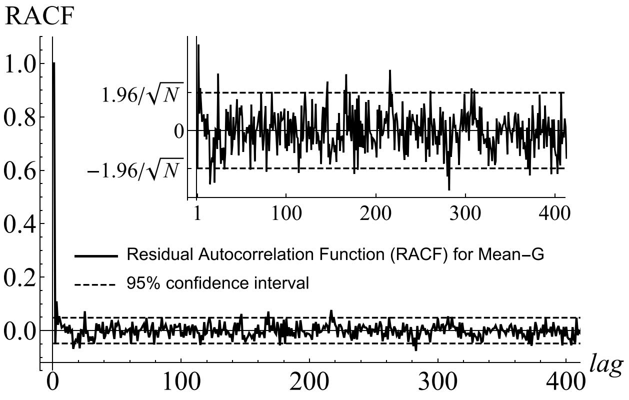

Fig. 23 RACF for the innovations of the Mean-G dataset. The theoretical 95% confidence interval, given by the values $\pm 1.96/\sqrt{N}$, is shown in dashed lines ($N = 1656$ is the total number of points)

months). The last two columns show the climate sensitivity, $\lambda_{2\times \text{CO}_2\text{eq}}$, and the parameter T_0 (Eq. (27)) used to remove the anthropogenic trend in each global series. The value T_0 was chosen to obtain $\bar{T}_{\text{nat}} = 0$ for each dataset, but this condition does not imply that $\hat{\mu} = 0$ in Eq. (49), as this last one is an estimate for the ensemble mean. Nevertheless, the values obtained for $\hat{\mu}$ were too small compared to $\hat{\sigma}_T$ and they were not included in Table 5.

With the parameters shown in Table 5 for global temperature series, we can check the fit of the model to the data as described at the end of Appendix 1. As an example, in Fig. 22 we show the natural variability component for the Mean-G dataset, together with its corresponding series of residual innovations, $\{e_i\}$, obtained using Eq. (67). The first series should be Gaussian with standard deviations SD_T while the residuals should be white noises, i.e. they should be NID(0, 1) with autocorrelation function $\langle e_i e_j \rangle = \delta_{ij}$. To verify the whiteness of the innovations, we should check that the residual autocorrelation function (RACF, (Eq. (68)) satisfies $r_l(\mathbf{e}_N) \sim \text{NID}(0, 1/N)$ for any lag $l \geq 1$ (for $l = 0, r_0(\mathbf{e}_N) = 1$).

The graph of the RACF for the innovations of the Mean-G dataset is shown in Fig. 23 for $0 \leq l \leq N/4$, where $N = 1656$ is the total number of points. The inset was obtained by dropping the point for zero lag and zooming in the y-axis. The theoretical 95% confidence interval, given by the values $\pm 1.96/\sqrt{N}$, is shown in dashed lines. From a direct inspection, we can see that there are not too many points that fall outside the band considered and the extreme values are not too far from these thresholds.

With the purpose of checking the Gaussianity hypothesis of the series represented in Figs. 22 and 23, a detailed statistical analysis was performed. Extremes in temperature natural variability are an important issue for the prediction of catastrophic events. Its presence would show as large tails in the distributions of temperature anomalies and their corresponding innovations. If this were the case, the model could be fixed by assuming white noise with a different distribution for the innovations (e.g. a Levy distribution or one from a multifractal process). On the other hand, deviations from Gaussianity in the RACF distributions would imply a different correlation structure and would automatically invalidate the applicability of the fGn model.

As an example, in Fig. 24, we show, from top to bottom, the results of this analysis for the natural variability component of the Mean-G dataset, for its corresponding series of residual innovations and for the RACF. In the left, there is a visual comparison of the empirical cumulative distribution functions, CDF, (blue) to that of the respective fitted Gaussian distributions (red) and in the right the more enlightening probability graphs where the empirical probabilities obtained from the graphs in the left are plotted against the theoretical probabilities (blue curve). The reference line shown in red corresponds to a perfect fit. The Kolmogorov–Smirnov (K–S) test can be used to create a measure that quantifies the behavior in probability graphs. The K–S test statistic is equivalent to the maximum vertical distance between a point in the plot and the reference line. The closer the points are to the reference line, the more probable is the data satisfies the fitted theoretical distribution.

In Table 6 we summarize the standard deviations of the normal distributions obtained for the series of anomalies (SD_T), the series of residual innovations (SD_{innov}) and the RACF (SD_{RACF}) for each dataset. The mean values of the distributions were very small compared to the respective standard deviations and they were omitted. The K–S test statistics with the corresponding p-values are also shown. More powerful statistical tests for normality could be performed, like the Shapiro–Wilk or the Anderson–Darling tests. However, these other tests have their own disadvantages, and, for the purpose of this work, the conclusions obtained from the K–S test to check the Gaussianity hypothesis of the original anomalies and the adequacy of the fGn process fit, are good enough.

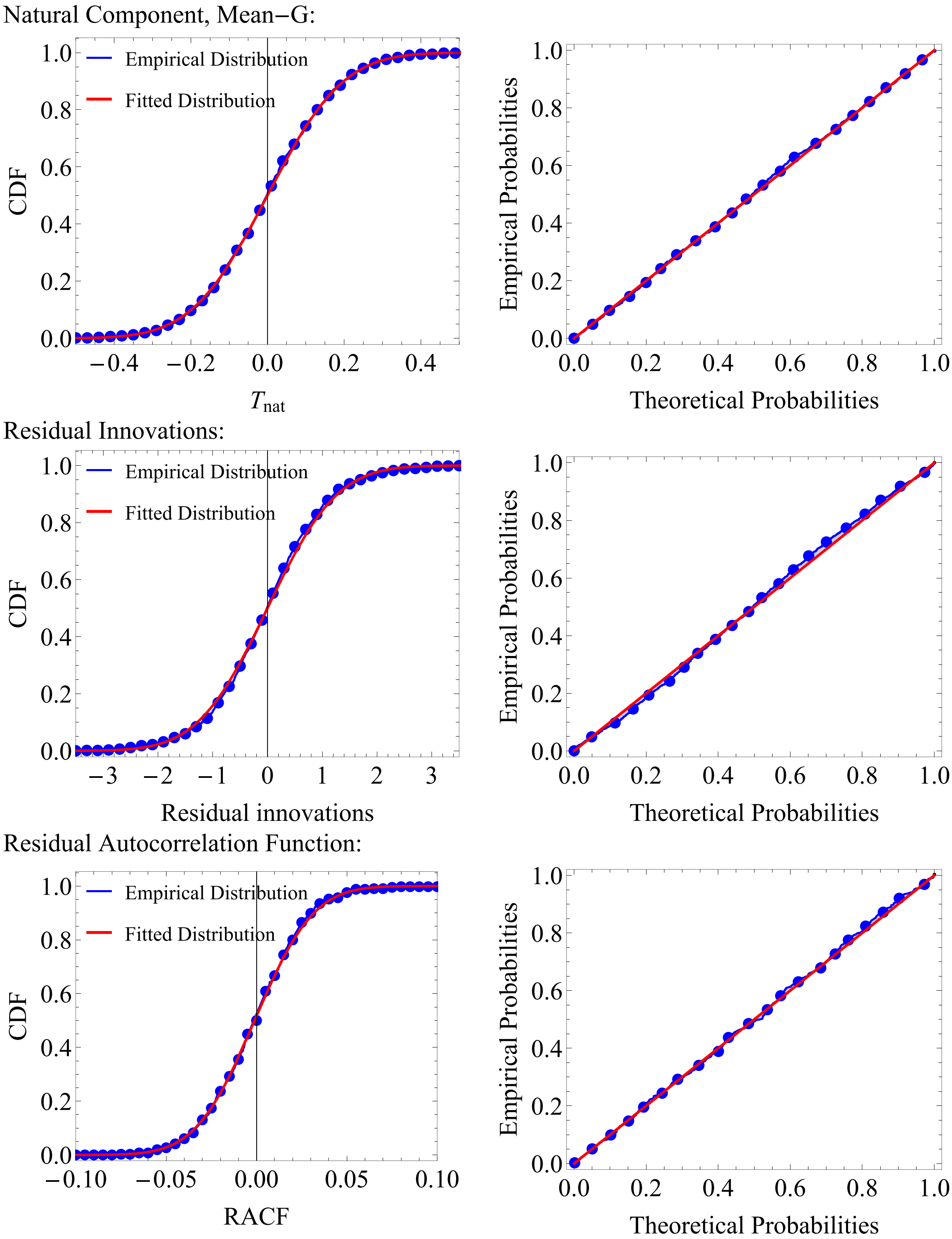

Fig. 24 From top to bottom, graphs for the natural variability component of Mean-G dataset, for the series of residual innovations and for the RACF. In the left, a comparison of the empirical CDF's (blue line with circles) to that of the respective fitted Gaussian distributions (red) and in the right the more detailed probability graphs where the empirical probabilities obtained from the graphs in the left are plotted against the theoretical probabilities (blue line with circles). The reference line shown in red corresponds to a perfect fit

Table 6 Normality tests and standard deviations of the distributions obtained for the series of anomalies (SD_T), the series of residual innovations (SD_{innov}) and the RACF (SD_{RACF}) for each global dataset

Dataset	Temperature anomalies			Residual innovations			RACF		
	SD_T	K–S	p value	SD_{innov}	K–S	p value	SD_{RACF}	K–S	p value
NASA	0.155	0.020	0.497	1.001	0.024	0.277	0.026	0.026	0.939
NOAA	0.144	0.029	0.114	1.000	0.044	0.003	0.025	0.033	0.747
HAD4	0.159	0.016	0.775	1.000	0.041	0.006	0.025	0.021	0.992
CowW	0.163	0.013	0.951	1.000	0.016	0.752	0.025	0.022	0.982
Berk	0.174	0.013	0.922	1.000	0.02	0.485	0.026	0.022	0.986
Mean-G	0.153	0.016	0.755	1.001	0.026	0.193	0.026	0.023	0.979
NASA-L	0.371	0.041	0.008	0.999	0.039	0.011	0.029	0.04	0.511
NOAA-L	0.325	0.040	0.009	1.000	0.051	0.000	0.029	0.063	0.072
HAD4-L	0.285	0.036	0.028	1.000	0.047	0.001	0.028	0.032	0.774
CowW-L	0.333	0.032	0.065	1.000	0.036	0.027	0.03	0.047	0.317
Berk-L	0.342	0.034	0.043	1.000	0.033	0.056	0.032	0.041	0.486
Mean-L	0.321	0.035	0.038	1.000	0.039	0.013	0.03	0.032	0.767

The mean values for each distribution were very small compared to the standard deviations and they were omitted. The K–S test statistics with the corresponding p values are also shown

Table 7 Skill scores $\mathrm{RMSE}_{\mathrm{raw}}$ and $\mathrm{ACC}_{\mathrm{nat}}$ for forecast horizons $k = 1, 3, 6$ and 12 months for the twelve datasets, including the mean series for the global and the land surface

Dataset	Normalization factor (°C)			$\mathrm{RMSE}_{\mathrm{raw}}$ (°C)				$\mathrm{ACC}_{\mathrm{nat}}$			
	$\hat{\sigma}_T$	SD_T	SD_{raw}	$k = 1$	$k = 3$	$k = 6$	$k = 12$	$k = 1$	$k = 3$	$k = 6$	$k = 12$
NASA	0.183	0.149	0.315	0.108	0.128	0.139	0.148	0.688	0.515	0.373	0.218
NOAA	0.183	0.140	0.301	0.093	0.113	0.127	0.137	0.744	0.587	0.434	0.264
HAD4	0.194	0.152	0.276	0.100	0.120	0.133	0.145	0.752	0.612	0.487	0.340
CowW	0.183	0.158	0.285	0.107	0.126	0.137	0.147	0.738	0.601	0.497	0.377
Berk	0.197	0.163	0.301	0.109	0.131	0.142	0.151	0.741	0.597	0.497	0.391
Mean-G	0.195	0.147	0.293	0.098	0.119	0.131	0.142	0.743	0.588	0.459	0.314
NASA-L	0.373	0.338	0.509	0.305	0.327	0.332	0.333	0.435	0.257	0.204	0.174
NOAA-L	0.331	0.327	0.521	0.296	0.318	0.324	0.325	0.429	0.238	0.167	0.140
HAD4-L	0.297	0.268	0.449	0.223	0.248	0.256	0.261	0.554	0.375	0.296	0.239
CowW-L	0.337	0.327	0.503	0.286	0.311	0.317	0.320	0.482	0.313	0.249	0.205
Berk-L	0.348	0.331	0.506	0.293	0.318	0.325	0.326	0.462	0.277	0.206	0.168
Mean-L	0.327	0.312	0.492	0.274	0.299	0.305	0.307	0.476	0.293	0.224	0.184

The values $\mathrm{MSSS}_{\mathrm{nat}}$ and $\mathrm{MSSS}_{\mathrm{raw}}$ can be obtained from Eqs. (34) taking $\mathrm{MSE} = \mathrm{RMSE}^2$ and the respective $\mathrm{MSE}_{\mathrm{ref}} = SD_T^2$ or $\mathrm{MSE}_{\mathrm{ref}} = SD_{\mathrm{raw}}^2$. The values of $\hat{\sigma}_T$, were included for reference

The values of SD_T are the same shown previously in Table 5. As expected from the theory, $SD_{\mathrm{innov}} = 1$ for all dataset and the values obtained for SD_{RACF} are close to the theoretical value $1/\sqrt{N} = 0.025$ ($N = 1656$). With the exceptions of the residual innovations of NOAA and HAD4 for the global datasets, the p-values are above 0.05, so there is not enough evidence to reject normality at that level. Moreover, the p-values obtained are, in general, larger than those obtained for series of the same length based on pseudorandom number generators [for a numerical experiment using 10000 samples, the p-values were uniformly distributed in the range (0–1)]. For the land surface datasets, the p-values for the temperature anomalies and the innovations are low and a different distribution for the white noise innovations could be proposed.

As we mentioned before, the normality of the innovations is less important to confirm the adequacy of the model than its whiteness, which is confirmed by the Gaussianity of the RACF in all cases (see the large p-values in the last column of Table 6). A main deviation from normal behavior is the existence of extremes in the original data. This "fat-tailed" property of the probability distributions was evidenced in Lovejoy (2014) in a paper of statistical hypothesis testing of anthropogenic warming. In the present work, it does not have major implications or compromise the applicability of the model to the global data.

Appendix 3: Forecast and validation for all datasets

Some results of the hindcast validation are summarized in Table 7 for the twelve datasets, including the mean series for the global and the land surface. Only the error, $\mathrm{RMSE}_{\mathrm{nat}}$, and the $\mathrm{ACC}_{\mathrm{nat}}$, for the natural variability component were presented for horizons $k = 1$, 3, 6 and 12 months. The values $\mathrm{MSSS}_{\mathrm{nat}}$ and $\mathrm{MSSS}_{\mathrm{raw}}$ can be obtained from Eq. (34) taking $\mathrm{MSE} = \mathrm{RMSE}^2$ and the respective $\mathrm{MSE}_{\mathrm{ref}} = SD_T^2$ or $\mathrm{MSE}_{\mathrm{ref}} = SD_{\mathrm{raw}}^2$. Also, we can use the values of $\mathrm{ACC}_{\mathrm{nat}}$ to obtain very good approximations of $\mathrm{MSSS}_{\mathrm{nat}}$ for these horizons thanks to the relationship $\mathrm{MSSS}_{\mathrm{nat}} \approx \mathrm{ACC}_{\mathrm{nat}}^2$ (Eq. (38)). Only the spurious values of $\mathrm{ACC}_{\mathrm{raw}}$ cannot be obtained from this table, but it is worth mentioning that, even for $k = 12$ months, they are higher than 0.75 for all datasets. Notice the large difference between the values of SD_T and SD_{raw}, for the detrended and the raw anomalies respectively, due to the presence of the anthropogenic trend. The values of $\hat{\sigma}_T$, were included for reference.

References

Baillie RT, Chung S-K (2002) Modeling and forecasting from trend-stationary long memory models with applications to climatology. Int J Forecast 18:215–226. https://doi.org/10.1016/S0169-2070(01)00154-6

Berkeley Earth (2018) Land + Ocean (1850–Recent). http://berkeleyearth.org/data/. Accessed 21 May 2018

Berner J, Achatz U, Batté L et al (2017) Stochastic parameterization: toward a new view of weather and climate models. Bull Am Meteorol Soc 98:565–588. https://doi.org/10.1175/BAMS-D-15-00268.1

Biagini F, Hu Y, Øksendal B, Zhang T (2008) Stochastic calculus for fractional brownian motion and applications. Springer, London

Blender R, Fraedrich K, Hunt B (2006) Millennial climate variability: GCM-simulation and Greenland ice cores. Geophys Res Lett 33:L04710. https://doi.org/10.1029/2005GL024919

CanSIPS (2016) CanSIPS data in GRIB2 format. https://weather.gc.ca/grib/grib2_cansips_e.html. Accessed 16 Feb 2016

Cowtan K, Way RG (2014) Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Q J R Meteorol Soc 140:1935–1944. https://doi.org/10.1002/qj.2297

Cowtan and Way (2018) Coverage bias in the HadCRUT4 temperature record. http://www-users.york.ac.uk/~kdc3/papers/coverage2013/series.html. Accessed 21 May 2018

Crochemore L, Ramos M-H, Pappenberger F (2016) Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrol Earth Syst Sci 20:3601–3618. https://doi.org/10.5194/hess-20-3601-2016

Franzke C (2012) Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. J Clim 25:4172–4183. https://doi.org/10.1175/JCLI-D-11-00293.1

Franzke CLE, O'Kane TJ, Berner J et al (2014) Stochastic climate theory and modeling. Wiley Interdiscip Rev Clim Chang 6:63–78. https://doi.org/10.1002/wcc.318

Garcia LE, Dawdy DR, Mejia JM (1972) Long memory monthly streamflow simulation by a broken line model. Water Resour Res 8:1100–1105. https://doi.org/10.1029/WR008i004p01100

GISTEMP Team (2018) GISS Surface Temperature Analysis (GISTEMP). NASA Goddard Institute for Space Studies. https://data.giss.nasa.gov/gistemp/. Accessed 21 May 2018

Gleason B, Williams C, Menne M, Lawrimore J (2015) GHCN-M Technical Report No. GHCNM-15-01 Modifications to GHCN-Monthly (version 3.3.0) and USHCN (version 2.5.5) processing systems. Asheville, NC

Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133:1098–1118. https://doi.org/10.1175/MWR2904.1

Gripenberg G, Norros I (1996) On the prediction of fractional brownian motion. J Appl Probab 33:400–410. https://doi.org/10.2307/3215063

Hansen J, Ruedy R, Sato M, Lo K (2010) Global surface temperature change. Rev Geophys 48:RG4004. https://doi.org/10.1029/2010rg000345

Hasselmann K (1976) Stochastic climate models Part I. Theory. Tellus 28:473–485. https://doi.org/10.1111/j.2153-3490.1976.tb00696.x

Hébert R, Lovejoy S, Tremblay B (2019) An observation-based scaling model for climate sensitivity estimates and global projections to 2100. Clim Dyn **(under review)**

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15:559–570. https://doi.org/10.1175/1520-0434(2000)015%3c0559:DOTCRP%3e2.0.CO;2

Hipel KW, McLeod AI (1994) Time series modelling of water resources and environmental systems, 1st edn. Elsevier, Amsterdam

Huybers P, Curry W (2006) Links between annual, Milankovitch and continuum temperature variability. Nature 441:329

Jeong H-DJ, Pawlikowski K, McNickle DC (2003) Generation of self-similar processes for simulation studies of telecommunication networks. Math Comput Model 38:1249–1257. https://doi.org/10.1016/S0895-7177(03)90127-0

Keller JD, Hense A (2011) A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms. Meteorol Zeitschrift 20:107–117. https://doi.org/10.1127/0941-2948/2011/0217

Kharin VV, Zwiers FW (2003) Improved seasonal probability forecasts. J Clim 16:1684–1701. https://doi.org/10.1175/1520-0442(2003)016%3c1684:ISPF%3e2.0.CO;2

Kharin VV, Teng Q, Zwiers FW et al (2009) Skill assessment of seasonal hindcasts from the Canadian historical forecast project. Atmos Ocean 47:204–223. https://doi.org/10.3137/AO1101.2009

Kharin VV, Merryfield WJ, Boer GJ, Lee W-S (2017) A postprocessing method for seasonal forecasts using temporally and spatially smoothed statistics. Mon Weather Rev 145:3545–3561. https://doi.org/10.1175/MWR-D-16-0337.1

Koscielny-Bunde E, Bunde A, Havlin S et al (1998) Indication of a universal persistence law governing atmospheric variability. Phys Rev Lett 81:729–732. https://doi.org/10.1103/PhysRevLett.81.729

Lettenmaier DP, Burges SJ (1977) Operational assessment of hydrologic models of long-term persistence. Water Resour Res 13:113–124. https://doi.org/10.1029/WR013i001p00113

Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141. https://doi.org/10.1175/1520-0469(1963)020%3c0130:DNF%3e2.0.CO;2

Lorenz EN (1972) Predictability; does the flap of a butterfly's wings in Brazil set off a tornado in Texas? Presented before the American Association for the Advancement of Science, December 29, 1972. http://eaps4.mit.edu/research/Lorenz/Butterfly_1972.pdf

Lovejoy S (2014) Scaling fluctuation analysis and statistical hypothesis testing of anthropogenic warming. Clim Dyn 42:2339–2351. https://doi.org/10.1007/s00382-014-2128-2

Lovejoy S (2015) Using scaling for macroweather forecasting including the pause. Geophys Res Lett 42(17):7148–7155. https://doi.org/10.1002/2015GL065665

Lovejoy S (2017) How accurately do we know the temperature of the surface of the earth? Clim Dyn 49:4089–4106. https://doi.org/10.1007/s00382-017-3561-9

Lovejoy S, de Lima MIP (2015) The joint space-time statistics of macroweather precipitation, space-time statistical factorization and macroweather models. Chaos An Interdiscip J Nonlinear Sci 25:075410. https://doi.org/10.1063/1.4927223

Lovejoy S, Schertzer D (2012) Low-frequency weather and the emergence of the climate. Extrem Events Nat Hazards Complex Perspect 196:231–254

Lovejoy S, Schertzer D (2013) The weather and climate: emergent laws and multifractal cascades. Cambridge University Press, Cambridge

Lovejoy S, Schertzer D, Varon D (2013) Do GCMs predict the climate… or macroweather? Earth Syst Dyn 4:439–454. https://doi.org/10.5194/esd-4-439-2013

Lovejoy S, del Rio Amador L, Hébert R (2015) The ScaLIng Macroweather Model (SLIMM): using scaling to forecast global-scale macroweather from months to decades. Earth Syst Dyn 6:637–658. https://doi.org/10.5194/esd-6-637-2015

Lovejoy S, Del Rio Amador L, Hébert R (2018) Harnessing butterflies: theory and practice of the stochastic seasonal to interannual prediction system (StocSIPS) BT. In: Tsonis AA (ed) Advances in nonlinear geosciences. Springer, Cham, pp 305–355

Mandelbrot BB (1971) A fast fractional gaussian noise generator. Water Resour Res 7:543–553. https://doi.org/10.1029/WR007i003p00543

Mandelbrot BB (1972) Broken line process derived as an approximation to fractional noise. Water Resour Res 8:1354–1356. https://doi.org/10.1029/WR008i005p01354

Mandelbrot BB, Van Ness JW (1968) Fractional Brownian motions, fractional noises and applications. SIAM Rev 10:422–437. https://doi.org/10.1137/1010093

Mandelbrot BB, Wallis JR (1969) Computer Experiments with fractional gaussian noises: part 3, mathematical appendix. Water Resour Res 5:260–267. https://doi.org/10.1029/WR005i001p00260

Matalas NC, Wallis JR (1971) Statistical properties of multivariate fractional noise processes. Water Resour Res 7:1460–1468. https://doi.org/10.1029/WR007i006p01460

Meinshausen M, Smith SJ, Calvin K et al (2011) The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. Clim Change 109:213–241. https://doi.org/10.1007/s10584-011-0156-z

Mejia JM, Rodriguez-Iturbe I, Dawdy DR (1972) Streamflow simulation: 2. The broken line process as a potential model for hydrologic simulation. Water Resour Res 8:931–941. https://doi.org/10.1029/WR008i004p00931

Merryfield WJ, Bertrand D, Fontecilla J-S et al (2011) The Canadian Seasonal to Interannual Prediction System (CanSIPS)—an overview of its design and operational implementation—technical Note

Merryfield WJ, Lee W-S, Boer GJ et al (2013) The Canadian Seasonal to Interannual Prediction System. Part I: models and initialization. Mon Weather Rev 141:2910–2945. https://doi.org/10.1175/MWR-D-12-00216.1

Met Office Hadley Centre (2018) Met Office Hadley Centre observations datasets. http://www.metoffice.gov.uk/hadobs/hadcrut4/data/current/download.html. Accessed 21 May 2018

Morice CP, Kennedy JJ, Rayner NA, Jones PD (2012) Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set. J Geophys Res Atmos. https://doi.org/10.1029/2011jd017187

Newman M, Sardeshmukh PD, Winkler CR, Whitaker JS (2003) A study of subseasonal predictability. Mon Weather Rev 131:1715–1732. https://doi.org/10.1175//2558.1

NOAA-NCEI (2018) Global Surface Temperature Anomalies. NOAA National Center for Environmental Information. https://www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php. Accessed 21 May 2018

Norros I (1995) On the use of fractional Brownian motion in the theory of connectionless networks. IEEE J Sel Areas Commun 13:953–962. https://doi.org/10.1109/49.400651

O'Connell PE (1974) Stochastic modelling of long-term persistence in streamflow sequences. Doctoral Thesis. Imperial College, London

Palma W (2007) Long-memory time series. Wiley, Hoboken

Palmer T, Buizza R, Hagedorn R et al (2006) Ensemble prediction: a pedagogical perspective. ECMWF Newsl 2:10–17. https://doi.org/10.21957/ab129056ew

Papoulis A, Pillai SU (2002) Probability, random variables and stochastic processes, 4th edn. McGraw-Hill, New York

Pasternack A, Bhend J, Liniger MA et al (2018) Parametric decadal climate forecast recalibration (DeFoReSt 1.0). Geosci Model Dev 11:351–368. https://doi.org/10.5194/gmd-11-351-2018

Paxson V (1997) Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic. ACM SIGCOMM Comput Commun Rev 27:5–18. https://doi.org/10.1145/269790.269792

Penland C, Matrosova L (1994) A balance condition for stochastic numerical models with application to the El niño–Southern oscillation. J Clim 7:1352–1372. https://doi.org/10.1175/1520-0442(1994)007%3c1352:ABCFSN%3e2.0.CO;2

Penland C, Sardeshmukh PD (1995) The optimal growth of tropical sea surface temperature anomalies. J Clim 8:1999–2024. https://doi.org/10.1175/1520-0442(1995)008%3c1999:TOGOTS%3e2.0.CO;2

Rodriguez-Iturbe I, Mejia JM, Dawdy DR (1972) Streamflow Simulation: 1. A new look at Markovian Models, fractional Gaussian noise, and Crossing Theory. Water Resour Res 8:921–930. https://doi.org/10.1029/WR008i004p00921

Rohde R, Muller RA, Jacobsen R et al (2013) A new estimate of the average earth surface land temperature spanning 1753 to 2011. Geoinform Geostat Overv. https://doi.org/10.4172/2327-4581.1000101

Rypdal K, Østvand L, Rypdal M (2013) Long-range memory in Earth's surface temperature on time scales from months to centuries. J Geophys Res Atmos 118:7046–7062. https://doi.org/10.1002/jgrd.50399

Sardeshmukh PD, Sura P (2009) Reconciling non-Gaussian climate statistics with linear dynamics. J Clim 22:1193–1207. https://doi.org/10.1175/2008JCLI2358.1

Smith TM, Reynolds RW, Peterson TC, Lawrimore J (2008) Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). J Clim 21:2283–2296. https://doi.org/10.1175/2007JCLI2100.1

Stanski HR, Wilson LJ, Burrows WR (1989) Survey of common verification methods in meteorology. WWW Tech. Rep. 8, WMO/TD 358. Ontario

Van Schaeybroeck B, Vannitsem S (2018) Postprocessing of long-range forecasts. Statistical postprocessing of ensemble forecasts. Elsevier, Oxford, pp 267–290

Winkler CR, Newman M, Sardeshmukh PD (2001) A linear model of wintertime low-frequency variability. Part I: formulation and forecast skill. J Clim 14:4474–4494. https://doi.org/10.1175/1520-0442(2001)014%3c4474:ALMOWL%3e2.0.CO;2

Wold H (1938) A study in the analysis of stationary time series. Almqvist und Wiksell, Uppsala

Yuan N, Fu Z, Liu S (2015) Extracting climate memory using fractional integrated statistical model: a new perspective on climate prediction. Sci Rep 4:6577. https://doi.org/10.1038/srep06577