Climate Dynamics
https://doi.org/10.1007/s00382-021-05737-5

Using regional scaling for temperature forecasts with the Stochastic Seasonal to Interannual Prediction System (StocSIPS)

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

Received: 15 June 2020 / Accepted: 15 March 2021

Abstract

Over time scales between 10 days and 10–20 years—the macroweather regime—atmospheric fields, including the temperature, respect statistical scale symmetries, such as power-law correlations, that imply the existence of a huge memory in the system that can be exploited for long-term forecasts. The Stochastic Seasonal to Interannual Prediction System (StocSIPS) is a stochastic model that exploits these symmetries to perform long-term forecasts. It models the temperature as the high-frequency limit of the (fractional) energy balance equation, which governs radiative equilibrium processes when the relevant equilibrium relaxation processes are power law, rather than exponential. They are obtained when the order of the relaxation equation is fractional rather than integer and they are solved as past value problems rather than initial value problems. StocSIPS was first developed for monthly and seasonal forecast of globally averaged temperature. In this paper, we extend it to the prediction of the spatially resolved temperature field by treating each grid point as an independent time series. Compared to traditional global circulation models (GCMs), StocSIPS has the advantage of forcing predictions to converge to the real-world climate. It extracts the internal variability (weather noise) directly from past data and does not suffer from model drift. Here we apply StocSIPS to obtain monthly and seasonal predictions of the surface temperature and show some preliminary comparison with multi-model ensemble (MME) GCM results. For 1 month lead time, our simple stochastic model shows similar—but somewhat higher—values of the skill scores than the much more complex deterministic models.

1 Introduction

The Navier–Stokes equations are the core of conventional numerical models for atmospheric prediction. These equations are derived from general conservation laws: energy, momentum, mass. Nevertheless, they have an implicit scale invariance symmetry, which is sometimes ignored in regard to other conservation laws (Lovejoy and Schertzer 2013; Palmer 2019). In this work, we exploit this symmetry as the basis for stochastic modelling and prediction of global temperature anomalies.

From hourly to centennial time scales, atmospheric fields are characterized by three scaling regimes: at high frequencies the weather, with fluctuations increasing with the time scale; there is a transition at $\tau_w \sim 10$ days to the macroweather, with fluctuations decreasing with scale; and at low frequencies the climate, again with increasing fluctuations. In recent times, the anthropogenic warming induces the transition to the climate regime at $\tau_c \sim 15$–20 years, but preindustrial records show $\tau_c > 100$ years (the Holocene transition scale is still not well known) (Lovejoy 2014). The transition time, τ_w, is the lifetime of planetary structures (Lovejoy and Schertzer 1986, 2010) and is therefore close to the deterministic predictability limit of conventional numerical weather prediction models. This predictability threshold for the models following a deterministic approach is imposed by the high complexity of the system and the sensitive dependence on initial conditions.

To extend the predictions to weekly, monthly and seasonal averages, stochasticity is incorporated at different levels in deterministic prediction systems. The ensemble approach, in which many different "random" realizations are obtained by integrating the model equations from slightly different initial conditions, is fundamentally stochastic. Sampling the attractor of the dynamic system is assumed to be equivalent to sampling the probability distribution of the possible outputs. Besides this implicit randomness product of chaos, explicit stochastic parameterization schemes are increasingly being incorporated in prediction systems.

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

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

Published online: 05 April 2021

Hybrid deterministic-stochastic approaches seem to be the future of macroweather forecasting (Williams 2012; Christensen et al. 2017; Davini et al. 2017; Rackow and Juricke 2020). The importance and the current state of stochastic climate modelling has been extensively discussed in the reviews: Franzke et al. (2015) and Palmer (2019).

In addition to these stochastic improvements to the deterministic core of conventional Global Circulation Models (GCMs), purely stochastic models have evolved as a complementary approach since the pioneering works of Hasselmann (1976). For these Hasselmann-type models, the high frequency "weather" is treated as a driving noise of the low frequency components described by integer-order linear ordinary differential equations. The most well-known are the linear inverse models (LIM) (Penland and Matrosova 1994; Penland and Sardeshmukh 1995; Winkler et al. 2001; Newman et al. 2003; Sardeshmukh and Sura 2009). These have been presented as a benchmark for decadal surface temperature forecasts. On the other hand, one of the main limitations of the LIM, is that it implicitly assumes short range exponential temporal decorrelations, while it has been shown that the true decorrelations are closer to long-range power laws (Koscielny-Bunde et al. 1998; Franzke 2012; Rypdal et al. 2013; Yuan et al. 2015). Consequently, LIM models underestimate the memory of the system, imposing a useful limit to the forecast horizon of roughly 1 year (Newman 2013).

In Lovejoy et al. (2015), the ScaLIng Macroweather Model (SLIMM) was introduced as an alternative stochastic model that respects the scaling symmetry. SLIMM generalizes LIM to use fractional differential equations that involve strong, long-range memories; it is these long-range memories that are exploited in SLIMM forecasts. The solution to the fractional differential equation in SLIMM is a fractional Gaussian noise process that is used to model the natural temperature variability.

In a recent series of papers (Lovejoy 2019, 2021a, b; Lovejoy et al. 2021), the classical Energy Balance Equation (EBE) is generalized to fractional orders: the Fractional EBE (FEBE). The phenomenological derivation of the FEBE complements derivations based on the classical continuum mechanics heat equation and of the more general Fractional Heat Equation (FHE) (Lovejoy et al. 2021), which is a fractional diffusion equation that has been studied in the statistical physics literature. When the FEBE is driven by a Gaussian white noise, the result is fractional Relaxation noise (fRn) that generalizes the classical Ornstein–Uhlenbeck process and its high-frequency limit is a fractional Gaussian noise process (fGn) that generalizes Brownian motion (Lovejoy 2019). In that sense, the fractional differential equation and the corresponding fGn solution exploited in SLIMM are the high-frequency approximations of the FEBE and its fRn solution, respectively.

In Del Rio Amador and Lovejoy (2019) (hereafter DRAL) the Stochastic Seasonal to Interannual Prediction System (StocSIPS) was introduced and applied to the prediction of globally averaged monthly temperature in the macroweather regime. StocSIPS 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. StocSIPS also improves the mathematical and numerical techniques used in the original SLIMM.

In DRAL, we presented the basic theory behind StocSIPS and applied it to the prediction of globally averaged series showing verification skill scores in both deterministic and probabilistic modes. We also compared hindcasts with Canada's operational long-range forecast system, the Canadian Seasonal to Interannual Prediction System (CanSIPS), and we showed that StocSIPS is just as accurate for 1-month forecasts, but significantly more accurate for longer lead times.

In this paper (specifically in Sects. 2.2 and 2.3), we verify that the scaling symmetry, which is the basis of StocSIPS, also holds at the regional level for monthly surface temperature, although some modifications must be introduced in the pre-processing of the tropical ocean temperature anomalies. In Sect. 2.4, we describe these particularities together with some theoretical details, although we purposely placed the most technical aspects in Appendix 1, so the main body of the article remains more results-based without too many overwhelming technicalities. Although all the equations and details relevant to this paper are given in the main text or in Appendix 1, the interested reader could refer to the more detailed theoretical description given in DRAL. The applicability of the model for all the regional series was confirmed through statistically testing in the second part of Sect. 2.4 and by contrasting the theoretically expected skill scores (if the model were perfect) with actual hindcast verification results for the natural temperature variability in Sect. 3.1. Finally, in Sect. 3.2 we apply StocSIPS to obtain monthly and seasonal predictions of the surface temperature and we show some preliminary comparisons with multi-model ensemble (MME) GCM results.

For 1 month lead time, our simple stochastic model shows similar values of the skill scores than the much more complex conventional models, with the advantage that it is much less expensive computationally and it can be easily adapted to direct hyperlocal prediction without need for downscaling. 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 obtained validate StocSIPS as a good alternative and a complementary approach to conventional numerical models. This complementarity is the basis for combining the two in a hybrid model that would bring the best of both worlds.

2 StocSIPS

2.1 Data preprocessing

In this study, the reference observational datasets are monthly average surface temperature (T2m) from the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis 1 (Kalnay et al. 1996; NCEP/NCAR 2020). The data were accessed on January 3, 2020 and it covers the period January 1948 to December 2019 (864 months in total). All data were interpolated to a 2.5° latitude × 12.5° longitude grid across the globe for a total of $73 \times 144 = 10{,}512$ grid points. Our objective is to model and predict this dataset using the Stochastic Seasonal to Interannual Prediction System (StocSIPS).

StocSIPS was presented in DRAL and applied to the prediction of globally averaged temperature in the macroweather regime. The main idea behind it is to consider the temperature series at position **x** as a combination of three independent signals:

$$T(\mathbf{x},t) = T_{\mathrm{ac}}(\mathbf{x},t) + T_{\mathrm{anth}}(\mathbf{x},t) + T_{\mathrm{nat}}(\mathbf{x},t). \tag{1}$$

The first component, $T_{\mathrm{ac}}(\mathbf{x},t)$, is the periodic annual cycle and is obtained from the mean temperature for each month in some reference period (here taken as the full length of the temperature datasets: 1948–2019). We assume that, for the time scales involved in the modelling and prediction problems, the annual cycle is unchanged. Also, for such a long verification period, the differences with the anomalies obtained using leave-one-out cross-validation methods are negligible. In Fig. 1a, we show an example of the raw temperature data, T (in red), and the periodic signal, T_{ac} (in blue), for the time series corresponding to the coordinates 50.0°N, 2.5°E (near Paris, France). In the graph, only the period 1981–2010 is shown for visual clarity.

The second component, $T_{\mathrm{anth}}(\mathbf{x},t)$, is a deterministic low-frequency response to anthropogenic forcings. It can be modelled as a response to equivalent-CO_2 (CO_2eq) radiative forcing as the one used in CMIP5 simulations (Meinshausen et al. 2011):

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

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 $\lambda_{2\times \mathrm{CO_2eq}}(\mathbf{x})$ is the transient climate sensitivity at position **x** (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_2eq includes not only greenhouse gases, but also aerosols, with their corresponding cooling effect. The sensitivity $\lambda_{2\times \mathrm{CO_2eq}}(\mathbf{x})$ is estimated from the linear regression of $T(\mathbf{x},t)$ vs. $\log_2\left[\rho_{\mathrm{CO_2eq}}(t)/\rho_{\mathrm{CO_2eq,pre}}\right]$. This relationship ignores memory effects, but these are not too strong during periods where the forcing continues to increase. The zero-mean residual natural variability component, $T_{\mathrm{nat}}(\mathbf{x},t)$, includes "internal" variability and the response of the system to other natural forcings (e.g.: volcanic and solar). Both components, T_{anth} and T_{nat}, are shown in Fig. 1b (blue and red, respectively) for the same point as in Fig. 1a with coordinates 50.0°N, 2.5°E. At this location, it could be argued that the anthropogenic trend is insignificant compared to the amplitude of the natural component, but at some other locations it is more relevant. Besides, the cumulative effect of T_{anth} for all the grid points is highly relevant for the globally averaged temperature (see Fig. 5 in DRAL).

Instead of using CO_2eq, alternatively, we could have used the CO_2 concentration in Eq. (2) as a surrogate for all anthropogenic effects, avoiding various uncertain radiative assumptions needed to estimate CO_2eq (especially aerosols). Nevertheless, from the point of view of detrending, the residuals, T_{nat}, would remain almost unchanged because of the nearly linear relation between the actual CO_2 concentration and the estimated equivalent concentration (correlation coefficient > 0.993). There are more rigorous methods of detrending the original signal to obtain independent components with "stationary" residuals while preserving the length of the time series [e.g.: empirical mode decomposition (EMD) (Zeiler et al. 2010), ensemble empirical mode decomposition (EEMD) (Wu and Huang 2009), LOESS (Cleveland and Devlin 1988; Clarke and Richardson 2021)]. Nevertheless, the method used here gives a direct physical meaning to the residual, T_{nat}, and to the low-frequency trend, T_{anth}. It is also accurate enough for obtaining the detrended temperature anomalies, whose characterization, modelling and prediction are the focus of the following sections. A more accurate method that takes into account the physics of the system adding memory effects to the heat balance equation, was presented in Procyk et al. (2020).

2.2 Spectra

The effects of the detrending in the frequency domain can be observed by comparing the spectra of the raw temperature series and the residual component, T_{nat}. In Fig. 1c we show

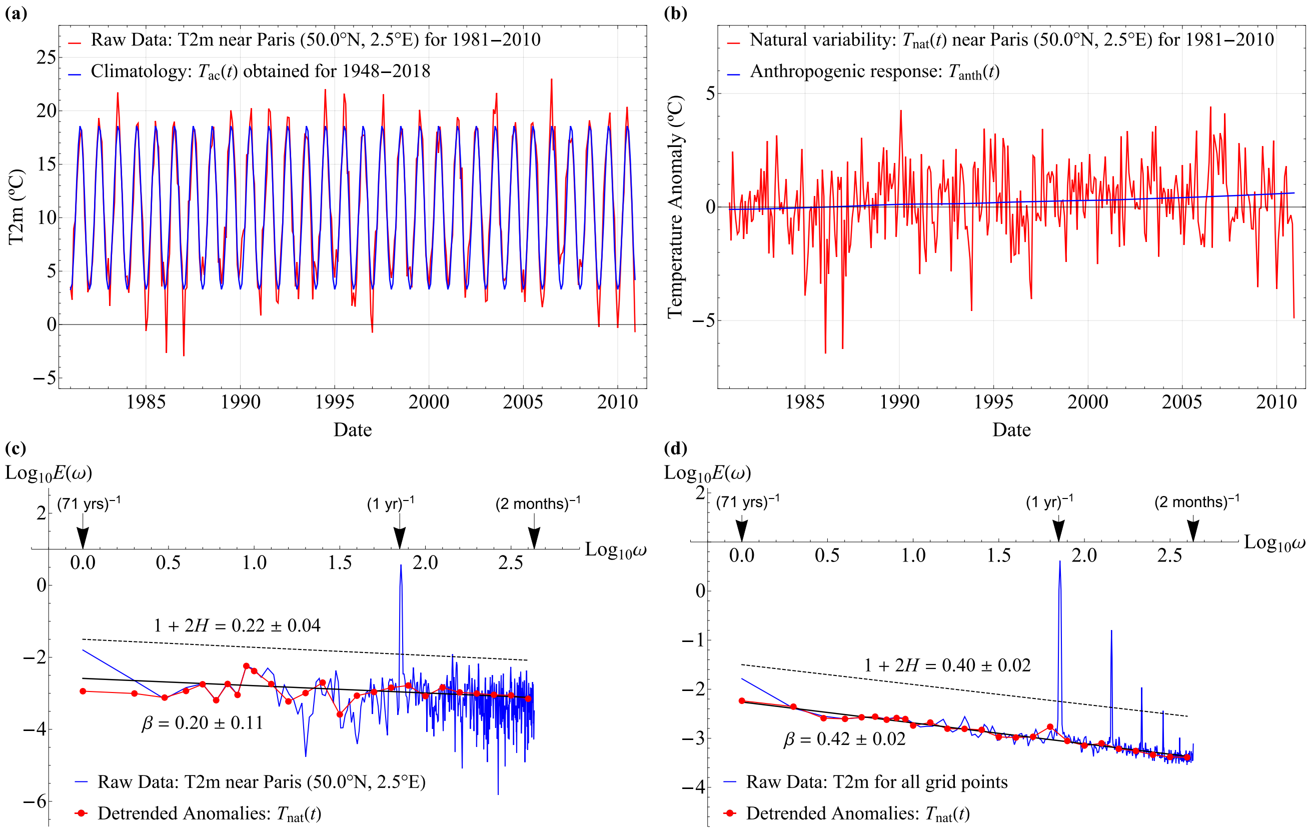

Fig. 1 Example of signal pre-processing and spectra for the grid point with coordinates 50.0°N, 2.5°E (near Paris, France). **a** Raw temperature data, T (in red), and the periodic signal, T_{ac} (in blue). Only the period 1981–2010 is shown for visual clarity. **b** The zero-mean residual natural variability component, T_{nat} and the anthropogenic trend, T_{anth} (red and blue, respectively). **c** Spectra of the raw temperature series and the residual component, T_{nat} (blue and red, respectively). The exponent, β was obtained from the linear regression of the spectrum averaged over equally spaced logarithmic bins. The reference dashed line with slope $1 + 2H$ was also included. **d** Similar to **c**, but now considering the average spectra for all the 10,512 grid points

these two spectra in a log–log scale in blue and red, respectively, for the grid point with coordinates 50.0°N, 2.5°E. The spectrum of the detrended series was smoothed by taking averages with logarithmically spaced bins. Notice that the peak corresponding to the annual cycle was removed along with the signal T_{ac}, as well as the low-frequency response corresponding to T_{anth}. The frequency, ω, is given in units of cycles per 72 years (72 years is the length of the series).

After removing the peaks corresponding to the annual cycle (and harmonics) and the low-frequency response, the only relevant feature of the spectrum of the detrended anomalies, $E(\omega)$, is its scale invariance (power–law behaviour):

$$E(\omega) \propto \omega^{-\beta}. \tag{3}$$

The exponent, $\beta = 0.20 \pm 0.11$, can be obtained from the linear regression of the spectrum averaged over equally spaced logarithmic bins (shown in red). The line corresponding to the best fit is shown in black in the figure. We also included a reference dashed line with slope $1 + 2H$, where H is the fluctuation exponent (see next section).

The scaling is even more noticeable in the less noisy spectrum shown in Fig. 1d, obtained by averaging the spectra of all the 10,512 grid points. Now the peaks corresponding to the periodic signal and the low-frequency contribution associated with anthropogenic effects are more clearly visible. The value of the exponent obtained in this case is $\beta = 0.42 \pm 0.02$. The implications of this scale-invariance will be treated in more detail in the following sections.

2.3 Scaling

In DRAL, it was shown that, for the case of globally averaged monthly atmospheric surface temperature, the statistics of $T_{nat}(t)$ are characterized by one main symmetry: the power-law (scaling) behaviour of the average of the fluctuations, ΔT, as a function of the time scale, Δt:

$$\langle|\Delta T(\Delta t)|\rangle \propto \Delta t^{H}, \tag{4}$$

where H is the fluctuation exponent and the brackets $\langle\bullet\rangle$ denote ensemble averaging. For $-1 < H < 0$, Haar fluctuations, not differences, should be used (Lovejoy and Schertzer 2012a). Many examples of the low intermittency ("spikiness") of the temperature fluctuations are given in Lovejoy and Schertzer (2013). Equivalently to Eq. (4), in the frequency domain the spectrum satisfies the previously mentioned equation: $E(\omega) \propto \omega^{-\beta}$, with $\beta = 1 + 2H$ for monofractal processes. These statistical symmetries are not exclusive to the globally averaged temperature. There are many empirical results that show a "colored noise" scaling behaviour in local temperature spectra as well as in many other atmospheric variables (Brockwell and Davis 1991; Blender et al. 2006; Box et al. 2008; Lovejoy and Schertzer 2013; Varotsos et al. 2013; Christensen et al. 2015).

For globally averaged temperature at scales between 1 month and several decades, there is a single scaling regime with $H < 0$. If we analyze temperature time series from daily (or shorter) time scales, we find that, in general, there is a transition between two scaling regimes: from the weather, characterized by fluctuations increasing with the time scale ($H > 0$), to the macroweather regime where fluctuations tend to cancel out as the time scale increases ($H < 0$).

This transition in the statistical properties of the atmosphere at scales of the order of $\tau_w \approx 10$ days, has been theorized by Lovejoy and Schertzer (1986) as the lifetime of planetary sized structures and estimated from first principles from knowledge of the solar output and the efficiency of conversion from solar to mechanical energy (Lovejoy and Schertzer 2010). A similar transition at $\tau_w \approx 1$ year was observed for the average surface temperature over the ocean (Lovejoy and Schertzer 2013).

The fluctuation exponents that characterize the weather and the macroweather regimes for air surface temperature (H_w and H_{mw}, respectively), as well as the transition scale τ_w, are functions of position with a strong dependence on the latitude. In Fig. 2, we show a map of the exponents obtained from the Haar fluctuation analysis (Lovejoy and Schertzer 2012a) in the high-frequency scaling regime between 2 months and 2 years. In general, there is a consistent difference between the macroweather exponents of surface temperature over the oceans and over land with $-1/2 < H_{mw}^{land} < H_{mw}^{ocean} < 0$ (the ocean is more persistent and the fluctuations cancel out more slowly). Also, for any position over land and for most of the ocean, we find that $\tau_w < 1$ month, so for surface temperature at monthly resolution, only the macroweather regime is observed. Only for the tropical ocean we do find a well-defined transition with τ_w as much as 2 years. Consequently, for this region, at time scales $\Delta t < \tau_w$ the statistics of the fluctuations are those of the weather regime with positive exponents (red in Fig. 2).

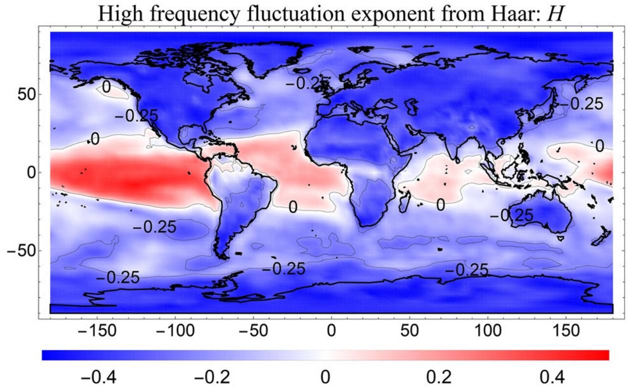

Fig. 2 Map of the fluctuation exponents obtained from the Haar fluctuation analysis (Lovejoy and Schertzer 2012a), in the high-frequency scaling regime between 2 months and 2 years

This longer transition in the SST corresponds to an analogous "ocean weather"–"ocean macroweather" transition (Lovejoy and Schertzer 2012b). It corresponds to lifetimes of large-scale ocean gyres (and other structures) that live much longer than atmospheric structures.

As an example, we show the Haar fluctuation analysis of the time series presented in Fig. 3a. We choose a point over land (time series in blue in Fig. 3a) with coordinates 50.0°N, 2.5°E (same grid point used before in Sect. 2.1) and a point in the tropical ocean (red in Fig. 3a) with coordinates 7.5°S, 30°W. In Fig. 3b, we show the average fluctuation as a function of the time scale before and after removing the anthropogenic trend for the point over land [red line with circles for the anomalies before removing the anthropogenic component ($T_{\text{anom}} = T_{\text{nat}} + T_{\text{anth}}$) and blue line with empty squares for the detrended anomalies (T_{nat})]. The reference line with slope $H_{mw} = -0.39 \pm 0.02$ was obtained from regression of the residuals' fluctuations between 2 months and 18.5 years. The units for Δt and ΔT are months and °C, respectively.

Notice that the anthropogenic warming breaks the scaling of the undetrended anomalies' fluctuations at a time scale of 15–20 years (the fluctuations start to increase with the scale at ~200 months). The fluctuation exponent for this low-frequency (climate) regime is $H_c = 1.0 \pm 0.1$—i.e., the fluctuations increase linearly with time following the almost linear growth of CO_2 concentration in recent epochs. The residual natural variability, on the other hand, shows reasonably good scaling for the whole period analyzed (66 years). In analysis of temperature records from preindustrial multiproxies and GCMs preindustrial control runs (Lovejoy 2014), evidence was presented showing that the range of scaling with decreasing fluctuations (pre-industrial macroweather) may extend to more than 100 years.

As we mentioned before, for this point over land, only one regime with fluctuations decreasing with the time scale (the macroweather regime) is present for the natural

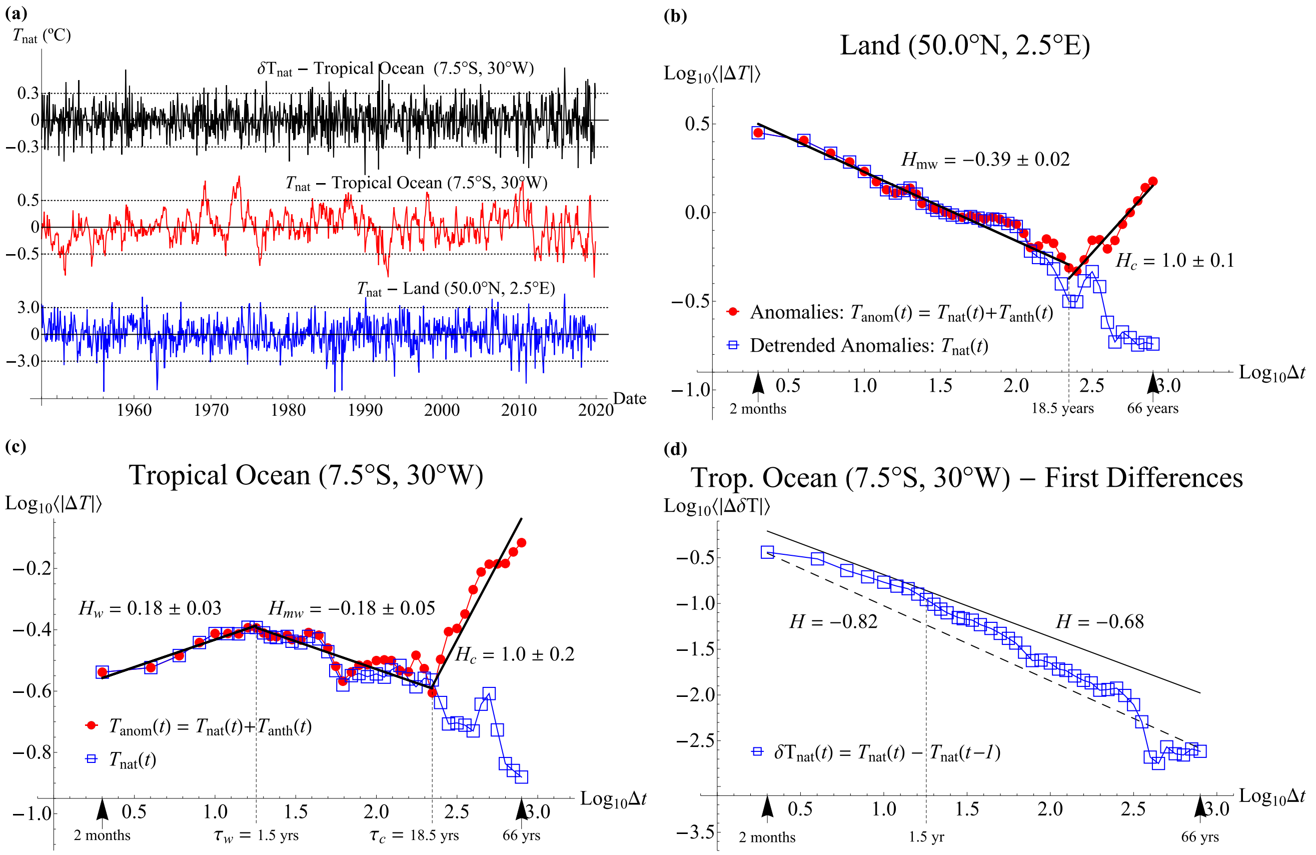

Fig. 3 Examples of Haar fluctuation analysis for two points, one over land and one over ocean. **a** In blue, time series for a point over land with coordinates 50.0°N, 2.5°E (same grid point used before in Sect. 2.1); in red, for a point over ocean located at 7.5°S, 30°W and in black, the series of the temperature differences, $\delta T_{nat}(t) = T_{nat}(t) - T_{nat}(t-1)$, for the same point over ocean (increments of the time series in red). **b** Average fluctuation as a function of the time scale before and after removing the anthropogenic trend for the point over land (red line with circles for the anomalies before removing the anthropogenic component and blue line with empty squares for the detrended anomalies). The reference lines with slopes $H_{mw} = -0.39 \pm 0.02$ and $H_{mw} = 1.0 \pm 0.1$ were obtained from regression of the anomalies' fluctuations in the respective macroweather and climate regimes. **c** Same as in **b** but now for the point over ocean. The three regimes (weather, macroweather and climate) are observed for this point. The corresponding transition scales and the respective exponents obtained from linear regression are also included in the graph. **d** Haar fluctuation analysis of the series of increments $\delta T_{nat}(t)$ for the point over ocean. The dashed line included as reference has slope $H = H_w - 1 = -0.82$, where H_w is the one shown in **c** and the solid line has a slope $H = -0.68$, which is the exponent obtained from the maximum likelihood method assuming that δT_{nat} is a fractional Gaussian noise (fGn) process (see next section)

variability. So, we can conclude that, at this location, the weather–macroweather transition occurs at $\tau_w < 1$ month (maximum resolution of the analyzed data). This was confirmed using 6-h resolution data. In contrast, as we show in Fig. 3c, if we analyze the grid point in the tropical region over the ocean (time series in red in Fig. 3a), there is a clear transition at $\tau_w \sim 1.5$ years from the weather regime (with fluctuations increasing with the scale) to the macroweather regime (with decreasing average fluctuations). A further transition occurs in the undetrended anomalies at $\tau_c \sim 18.5$ years to the climate regime, where fluctuations start to increase again with the time scale. As before, this transition in recent epochs is induced by anthropogenic effects. The actual transition in the natural variability, as obtained from preindustrial temperature records, apparently occurs at time scales longer than 100 years, which is consistent with the blue curves for T_{nat} in Fig. 3b,c after we remove the anthropogenic trend. The fluctuation exponent for the three regimes, weather–macroweather–climate, has values $H_w = 0.18 \pm 0.03$, $H_{mw} = -0.18 \pm 0.05$ and $H_c = 1.0 \pm 0.2$, respectively (shown in the graph) consistent with a smooth low-frequency behaviour.

A visual comparison between the blue and red curves in Fig. 3a shows a clear difference in the temperature behaviour at these two grid points. While over land, consecutive values of temperature tend to cancel out, over the ocean the temperature is more persistent and only after several time steps the anomalies change sign. This is confirmed in the Haar fluctuation analysis shown in Fig. 3b,c. This difference in the statistical behaviour imply that, while a fractional

Gaussian noise (fGn) model is a good fit for the extratropics, we cannot use it to describe the tropical region. Nevertheless, if we take the first differences in the time series for the grid point over the tropical ocean, the new series $\delta T_{\mathrm{nat}}(t) = T_{\mathrm{nat}}(t) - T_{\mathrm{nat}}(t-1)$ (shown in black in Fig. 3a) has a statistical behaviour which is clearly more similar to the series over land with consecutive fluctuations cancelling out. As we can see in the graph shown in Fig. 3d, the new series $\delta T_{\mathrm{nat}}(t)$ has a scaling regime for small Δt with negative fluctuation exponent similar to that of Fig. 3b. By taking first differences in the tropics, we are able to use fGn process everywhere to predict the time series, then we can go back to the original series for those places by taking cumulative sums.

There is still a change in the slope at $\tau_w \sim 1.5$ years, corresponding to the one in the original series shown in Fig. 3c. The dashed line included as reference has a slope $H = H_w - 1 = -0.82$. The series δT_{nat}, being the increments of the series T_{nat}, should have an exponent of the dominant high frequencies reduced by one. We also included in solid black, a reference line with slope $H = -0.68$, which is the exponent obtained from the maximum likelihood method assuming that δT_{nat} is an fGn process (see Sect. 2.4.2).

These examples—shown here for two different positions—are representative of the behaviour of the natural temperature variability all over the Earth. In fact, by taking the first differences of the time series in those places over the tropical ocean with weather regime at monthly resolution, we can reduce our analysis to only one case of self-similar time series with negative exponent in the range $-1 < H < 0$. This simplification emphasizes the role of the scaling symmetry, which is sometimes ignored in regard to other conservation laws, in spite of being also present in the Navier–Stokes equations (Lovejoy and Schertzer 2013; Palmer 2019), which are the core of conventional numerical models for atmospheric prediction and hence respected by them. In this work, we exploit this symmetry as the basis for stochastic modelling and prediction of global temperature anomalies.

2.4 Stochastic modelling using fGn and fRn

2.4.1 Properties of fGn, fRn

Together with the scaling symmetry presented in the previous section, we also assume the Gaussianity of the natural temperature variability. This Gaussian hypothesis was verified in DRAL for globally averaged monthly temperature in the macroweather regime. Although the Gaussian assumption is commonly made, it is worth underlining that it is somewhat surprising that it is a reasonable model for macroweather time series. Recall that Gaussian statistics imply that macroweather in time has little or no intermittency (the series are mono-, not multifractal, the transitions are not “spiky”). This contrasts with macroweather in space, which is highly intermittent, as well as the existence of highly intermittent, nonGaussian, multifractal spatial and temporal statistics in the weather and climate regimes (Lovejoy 2018).

The scaling of the temperature fluctuations and spectrum implies that there are power-law correlations in the system and hence a large memory effect that can be exploited. In Lovejoy (2019) and Lovejoy et al. (2021), it was argued that the origin of this memory are the Earth’s hierarchical, scaling energy storage mechanisms whereby anomalies in energy fluxes either external (e.g. anthropogenic) or internal can be stored for long periods. It was argued that to a good approximation, the temperature satisfies the Fractional Energy Balance Equation (FEBE) that has a high-frequency scaling storage term and a low-frequency energy balance term. When the FEBE is internally forced by a Gaussian white noise, the temperature response is the statistically stationary fractional Relaxation noise (fRn) process (Lovejoy 2019).

However, at time scales shorter than the relaxation time (of the order of a few years), the (scaling) storage term is dominant and, for exponents $-1/2 < H < 0$, the temperature response is a fractional Gaussian noise (fGn) process. This was the approximation made in DRAL and is empirically valid for all land areas and most of the oceans. The exceptions are some parts of the tropical ocean where $0 < H < 1$ (Figs. 2 and 4a), we return to these below.

The original idea of modelling the natural variability using an fGn process was presented in Lovejoy et al. (2015) as the ScaLIng Macroweather Model (SLIMM). In DRAL, StocSIPS was introduced as a general system that includes SLIMM as the core prediction model. StocSIPS also improves the mathematical and numerical techniques of SLIMM. It was applied to the prediction of globally averaged temperature series since 1880. The comparison of StocSIPS hindcasts with Canada’s operational long-range forecast system, the Canadian Seasonal to Interannual Prediction System (CanSIPS), showed that StocSIPS is just as accurate for 1-month forecasts, but significantly more accurate for longer lead times.

In this paper we extend the globally averaged version of StocSIPS for the prediction of a single temperature time series to the prediction of the full space–time temperature field. The basic theory for fGn processes, used here to model those places where $-1/2 < H < 0$ (most of the planet), is summarized in Appendix 1. An fGn process is fully characterized by two parameters (assuming zero mean): the fluctuation exponent, H, and the standard deviation, σ_T.

We mentioned that for most places in the tropical ocean, $0 < H < 1$. While these may still be modelled by fRn processes, the high-frequency approximation to fRn is no longer an fGn process, but rather a fractional Brownian motion

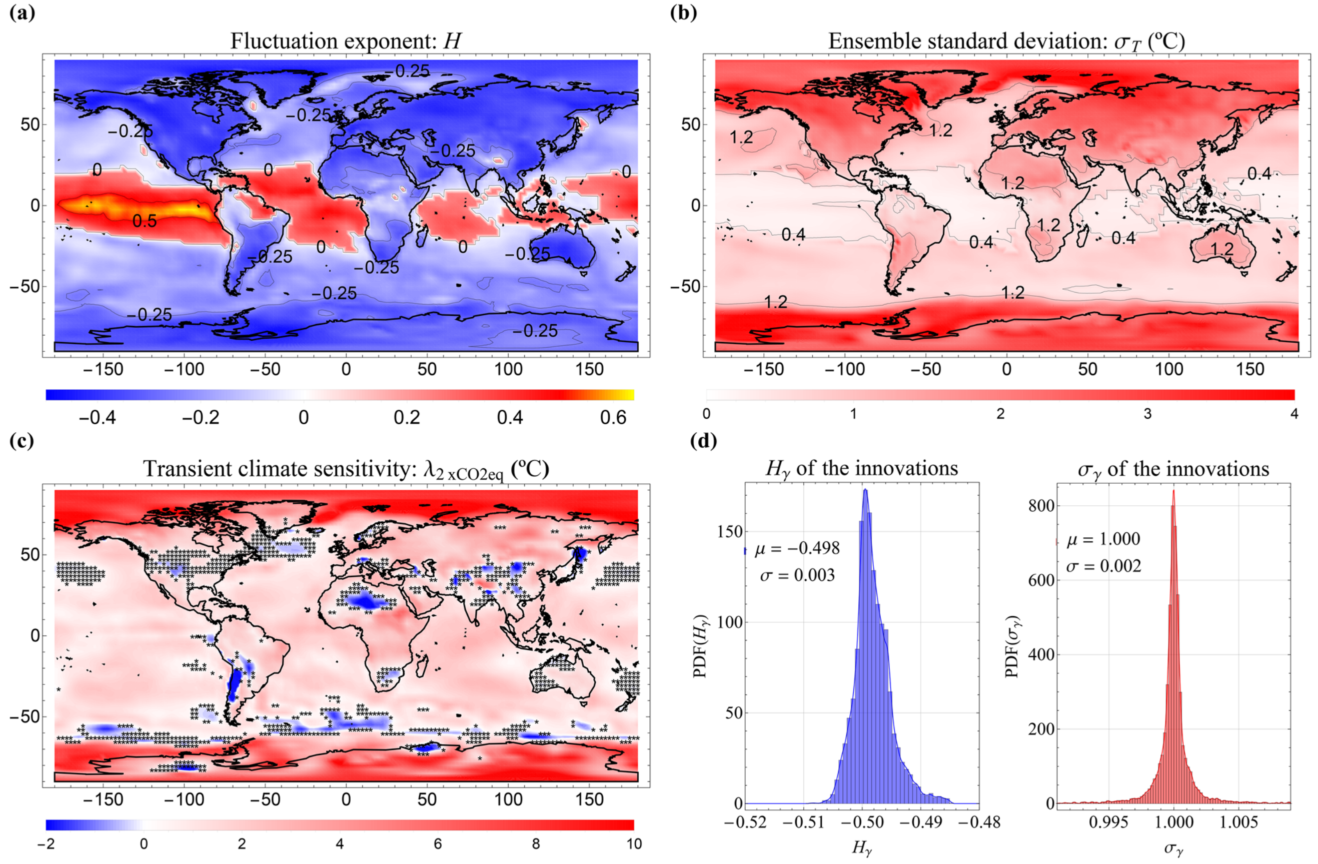

Fig. 4 Estimates of the three parameters $(H, \sigma_T, \lambda_{2\times \mathrm{CO_2eq}})$ obtained for each grid point and statistics of the innovations, $\gamma(t)$. **a** Maximum likelihood estimates of the temperature fluctuation exponent (compare with the estimates shown in Fig. 2). There is a discontinuity from negative to positive values of H as we approach the tropical ocean, corresponding to the change in model from fGn to fBm. **b** The standard deviation, σ_T, of the infinite ensemble fGn process. **c** Map of the transient climate sensitivity, defined in Eq. (2). The places marked with "*" indicate pixels where the null hypothesis, $\lambda_{2\times \mathrm{CO_2eq}} = 0$, cannot be rejected with more than 90% confidence. **d** Histograms of the fluctuation exponent and the standard deviation of the innovations (H_γ and σ_γ, respectively) for the 10,512 grid points. From the histograms, we can conclude that the innovations are very close to white noise for the whole planet ($H_\gamma = -0.498 \pm 0.003$ and $\sigma_\gamma = 1.000 \pm 0.002$)

(fBm) process, and we must use the correlation function for fRn given in Appendix 1.3, Eq. (23). For those regions with positive H, the first differences of the temperature, $\delta T_{\mathrm{nat}}(t) = T_{\mathrm{nat}}(t) - T_{\mathrm{nat}}(t-1)$, has H values reduced by 1, so for δT_{nat} we also have $-1 < H < 0$. That is, either the natural temperature variability itself or its first differences can be modelled by an fGn process. In those places where $H > 0$ for the high frequencies, it would be equivalent to modelling them with an fBm or fRn process. Of course, a true fBm would only have one scaling regime with positive fluctuation exponent, instead of the bi-scaling regime shown for the detrended anomalies in Fig. 3c. To model those series as an fBm process is an approximation that would work well for the high frequencies, but that would fail in reproducing the low frequency behaviour.

2.4.2 Parameter estimates and model adequacy

With the distinction in the tropical region where we take the first differences to adjust everything to an fGn model, we conclude that to model the actual temperature field for the globe (including the anthropogenic trend), for each grid point of the NCEP/NCAR Reanalysis 1 data we may estimate the three parameters $(H, \sigma_T, \lambda_{2\times \mathrm{CO_2eq}})$. For the first two, we use the maximum likelihood method described in Appendix 1 of DRAL and for the sensitivity we use the regression described in Sect. 2.1. To verify the model adequacy, we use Eq. (22) to obtain the residual innovations, $\gamma(t)$, then, using the maximum likelihood method, we obtain its variance, σ_γ, and its fluctuation exponent, H_γ; they should be equal to 1 and $-1/2$, respectively (white noise processes are particular cases of fGn with $H = -1/2$). The results are summarized in Fig. 4.

A map of the maximum likelihood estimates of the temperature fluctuation exponent is shown in Fig. 4a. These values are more accurate and give a better fit of our model than the high-frequency Haar estimates shown in Fig. 2. Notice that for most of the globe and all of the land, the values are in the range $-1/2 < H < 0$, which is characteristic of long-range memory fGn processes with nonsummable correlation functions, i.e. the sum over Δt of the series with elements given by Eq. (17) diverges for this range of H. There is a discontinuity from negative to positive values of H as we approach the tropical ocean, corresponding to the change in model from fGn to fBm (or, equivalently, from the description as an fGn of the natural temperature variability, T_{nat}, to the description of the temperature differences, δT_{nat}). In most of the tropical ocean (red regions in the map), the natural temperature variability has fluctuation exponents in the range $0 < H < 1/2$, whose fBm approximation has "antipersistent" increments (consecutive increments are negatively correlated). Only in the eastern equatorial Pacific (yellow region in the map), do we obtain fluctuation exponents in the range $1/2 < H < 1$, whose fBm approximation has persistent (positively correlated) increments. It is significant that it is precisely this more predictable region that is associated with the ENSO phenomenon (Trenberth 1997), the strongest interannual signal of climate variability on Earth.

In Fig. 4b we show the values of the parameter σ_T. Although this is the standard deviation of the infinite ensemble fGn process, for a given finite realization it does not coincide with the usual estimate (SD_T) based on the temporal average:

$$SD_T^2 = \frac{1}{N}\sum_{t=1}^{N}\left[T_{\mathrm{nat}}(t) - \overline{T}_N\right]^2, \tag{5}$$

where $\overline{T}_N = \sum_{t=1}^{N} T_{\mathrm{nat}}(t)/N$ (the over-bar notation is used to denote averaging in time). The biased estimate SD_T ignores correlations, that are however considered in the maximum likelihood estimate of σ_T. The relation between the two values for fGn processes depends on the length of the time series and the fluctuation exponent, H, and is given by:

$$SD_T^2 = \sigma_T^2\left(1 - N^{2H}\right) \tag{6}$$

(see Sect. 3.3 and Appendix 1 of DRAL). Notice that there is also a discontinuity in the map of σ_T for the same reasons explained previously. In general, the amplitude of the fluctuations is larger over land than over the ocean; the surface temperature over the ocean is less variable as this has a higher thermal inertia than land.

A map of the transient climate sensitivity, defined in Eq. (2), is shown in Fig. 4c. The places marked with "*" indicate grid boxes where the null hypothesis,$\lambda_{2\times\mathrm{CO_2eq}} = 0$, cannot be rejected with more than 90% confidence. Notice that these values depend on the reference dataset. In our case we used the NCEP/NCAR Reanalysis 1, which only has data since 1948. More precise estimates of the climate sensitivity were obtained by Hébert and Lovejoy (2018) using five observational datasets since 1880. In this paper, we are not aiming at an accurate study of the climate sensitivity. We should consider the values of $\lambda_{2\times\mathrm{CO_2eq}}$ reported here as a parameter used for detrending the temperature time series related to the anthropogenic effects.

Finally, in Fig. 4d, we show histograms of the fluctuation exponent and the standard deviation of the innovations (H_γ and σ_γ, respectively) for the 10,512 grid points. From the histograms, we conclude that the innovations are very close to white noise for all the places in the planet ($H_\gamma = -0.498 \pm 0.003$ and $\sigma_\gamma = 1.000 \pm 0.002$). So, with a high degree of accuracy, all the innovation series can be considered NID(0,1) (Normally and Independently Distributed with mean 0 and variance 1), and we can conclude that the fGn model is a good fit to the natural temperature variability (or its increments in the red and yellow places of the map in Fig. 4a).

3 Results

3.1 Natural variability forecast

3.1.1 Model validation through hindcast

In the previous section, we validated the fGn model as a good fit to the natural temperature variability (or to its increments) by checking the whiteness of the residual innovations. The goal of this section is to further validate the model by using the theory presented in Appendix 1.4 to hindcast only the natural variability—not the anthropogenic signal or the annual cycle—and seeing how well it performs. We test the assumptions made in the model by comparing the theoretically expected skill scores (expected values if the model were perfect) with the actual scores obtained from hindcasts. All the verification metrics used in this paper are detailed in Appendix 2.

Series of hindcasts at monthly resolution, were produced for forecast horizons from 1 to 12 months, in the period of verification (POV) from December 1950 to November 2019 (the verification starts in December in order to have the same number of conventional seasons: DJF, MAM, JJA, SON). In this 69-year verification period, each month was independently predicted using the information available m months before. For each horizon, k, we used a memory $m = 20$ months. For example, to predict the average temperature for December 1950 with $k = 1$ month, we used the previous 21 months, including November 1950, and the same was done for every verification date up to November 2019 and for all horizons up to $k = 12$ months. The dependence with the horizon of many scores [e.g. the root mean square error (RMSE)], is obtained from the difference between hindcasts series at a fixed k and the corresponding series of observations.

It is important to point out that the predictor $\widehat{T}_{\text{nat}}(t+k)$ (see Eq. 24) only depends on the previous $m+1$ months, from $T_{\text{nat}}(t-m)$ to $T_{\text{nat}}(t)$, weighted by coefficients that only depend on the fluctuation exponent H (see Fig. 4a). The estimates of H are quite robust and only small variations were obtained for different training periods, as long as the length of the training periods is larger than one third of the full length of the time series. Also, only small changes on the skill were found for small variations in H. In that sense, given the robustness of the estimates of the fluctuation exponent, we can use almost all the observational period for verification leaving only a few months before the first initialization date to use as memory. In all cases, the observational and forecast anomalies used for verification were calculated in the leave-one-out cross-validation mode.

3.1.1.1 Root Mean Square Error (RMSE) The infinite ensemble expectation of the RMSE is given in Appendix 1.4 (Eq. 27). This analytical expression is only a function of the model parameters and does not include any observational data. It is the theoretical RMSE value for a perfect model. To confirm the validity of the theoretical framework for the prediction of the natural variability component, we compare these expected values for each grid point with the actual verification RMSE obtained from hindcasts in the POV from December 1950 to November 2019. The all-month verification score for horizon k is obtained using Eqs. (32) and (33) with $N = 828$ months and $T_{\text{nat}}(t+k)$ and $\widehat{T}_{\text{nat}}(t+k)$ being the zero mean detrended observational and predicted anomalies, respectively.

The comparison between the theoretical and the actual (obtained from hindcasts) normalized root mean square error (NRMSE) is shown in Fig. 5 for horizon $k = 1$ month. The NRMSE is the RMSE normalized by the observed standard deviation (Eq. (5) for the natural variability). The NRMSE may vary from zero to infinity, with lower NRMSE values indicating more skillful forecasts. NRMSE values greater than 1 indicate that forecasts are less skillful than the climatological average value of the series. As we pointed out, for a fixed k, the theoretical RMSE only depends on the parameters σ_T and H. In general, there is very good agreement between theory and verification results. The maximum difference between the two maps in Fig. 5 is lower than 0.07. The forecast skill is higher over ocean than over land and takes the highest values over the tropical ocean, which corresponds to the spatial distribution of H values shown in Figs. 2 and 4a.

The maps in Fig. 5 were obtained for $k = 1$ month, but similar maps can be obtained for all forecast horizons from 1 to 12 months. The results of the comparison can be summarized in the scatter plots shown in Fig. 6. The graphs include the 10,512 grid points, showing the verification RMSE obtained from hindcasts vs. the expected theoretical $\text{RMSE}_{\text{nat}}^{\text{theory}}$ predicted by Eq. (27) for each horizon. As expected, the agreement between the theoretically expected scores and the hindcasts results decreases as the horizon increases, but it remains quite accurate in all cases with a correlation coefficient larger than 0.998. For the regions where $H > 0$, the fBm fit is less accurate; however, recall that in those places the actual statistics of the fluctuations are bi-scaling, while the fBm model assumes a perfectly scaling process. The accuracy of the theory decreases as the horizon approaches the transition time, τ_w.

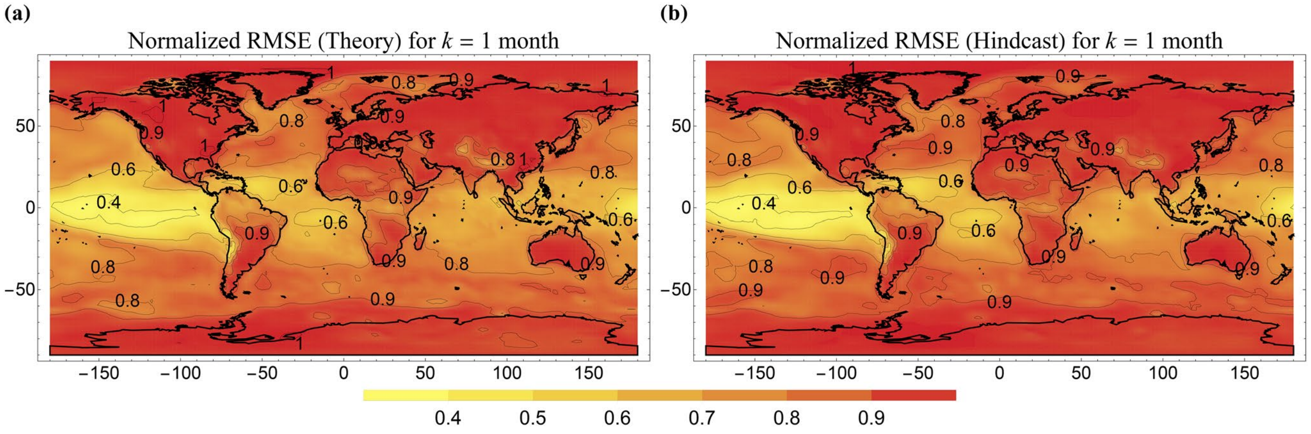
(a)
Normalized RMSE (Theory) for k = 1 month
(b)
Normalized RMSE (Hindcast) for k = 1 month
50
0
−50
−150
−100
−50
0
50
100
150
0.4
0.5
0.6
0.7
0.8
0.9

Fig. 5 Theoretical and hindcasts NRMSE for $k = 1$ month. The corresponding RMSEs were obtained using Eqs. (27) and (33), respectively, and the normalization standard deviation from Eq. (5) for the natural variability

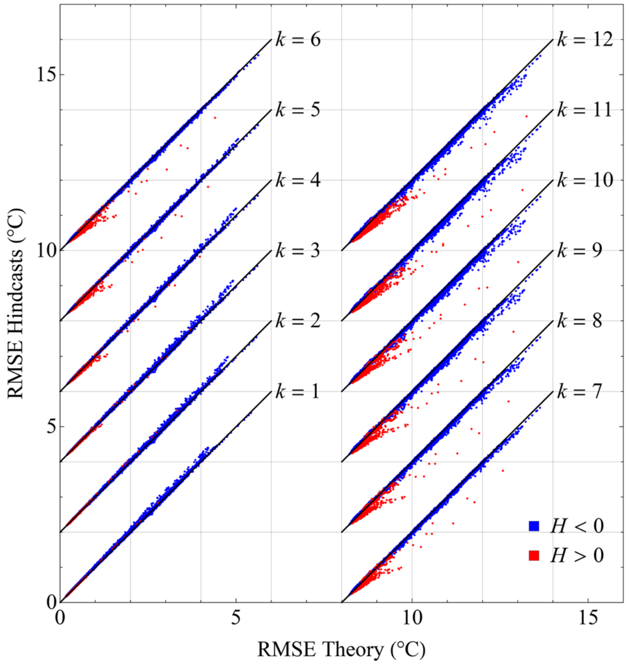

Fig. 6 Scatter plots for each horizon including the 10,512 grid points, showing the verification RMSE obtained from hindcasts vs. the expected theoretical $\mathrm{RMSE}_{\mathrm{nat}}^{\mathrm{theory}}$ predicted by Eq. (27). The graphs were displaced vertically by 2 °C (plus a horizontal displacement of 8 °C for $k \geq 7$ months) for visual clarity. The black line at 45° is a reference indicating perfect agreement between theory and verification. The blue points represent locations where $H < 0$ and the natural variability is modeled as an fGn process and the red points are for places where $H > 0$ and we use the fBm model

3.1.1.2 Mean Square Skill Score (MSSS) Related to the RMSE score, the MSSS is a commonly used metric (see Eq. 35). The guidelines of the World Meteorological Organization (WMO) Standard Verification System for Long-Range Forecasts (LRFs) (WMO 2010a), suggests the MSSS as a metric for deterministic forecasts (based on the ensemble mean). For leave-one-out cross-validated data in the POV (WMO 2010a), the mean square error (MSE) of the reference climatology forecasts (including the deterministic anthropogenic trend forecast) is:

$$\mathrm{MSE}_{\mathrm{C}} = \left(\frac{N}{N-1}\right)^{2} SD_T^2 \tag{7}$$

(see Eq. 34), where ${SD_T}^2$ is the variance of the detrended anomaly series (natural variability component). The MSSS for horizon k for the natural variability forecast is:

$$\mathrm{MSSS}_{\mathrm{nat}}(k) = 1 - \frac{\mathrm{MSE}_{\mathrm{nat}}(k)}{\left(\frac{N}{N-1}\right)^{2} SD_T^2}, \tag{8}$$

where $\mathrm{MSE}_{\mathrm{nat}}$ is obtained using Eq. (32) with $N = 828$ months and $T_{\mathrm{nat}}(t+k)$ and $\hat{T}_{\mathrm{nat}}(t+k)$ being the zero mean detrended observational and predicted anomalies, respectively.

One consequence of the memory effects in the natural variability is the increase of ${SD_T}^2$ with the length of the verification period given by Eq. (6). This implies that some metrics, such as the MSSS or the NRMSE, will actually have the same dependence with the duration of the verification period. The longer the verification period, the higher the value of MSSS (lower for NRMSE), even for a fixed prediction system (with fixed RMSE). Comparisons between skill scores of different models should always be made for the same POV (or at least the same length of the POV). As the number of months used for verification increases, ${SD_T}^2 \to {\sigma_T}^2$ and the MSSS approaches the asymptotic value (determined by H). This effect is small for most values of H, but is significant if too short verification periods are used or if H is close to zero (e.g.: the bias ${SD_T}^2/{\sigma_T}^2 \approx 0.6$ for $N = 100$ months and $H = -0.1$). See Fig. 9 in DRAL for an example in monthly globally averaged temperature.

3.1.1.3 Temporal correlation coefficient (TCC) The TCC is another commonly used verification score for deterministic forecasts (see Eq. 36). For the natural variability forecast, the TCC for horizon k is:

$$\mathrm{TCC}_{\mathrm{nat}}(k) = \frac{\overline{T_{\mathrm{nat}}(t+k)\hat{T}_{\mathrm{nat}}(t+k)}}{SD_T\sqrt{\overline{\hat{T}_{\mathrm{nat}}(t)^2}}}, \tag{9}$$

where the overbars indicate temporal average for a constant k.

For the natural variability forecast, the autoregressive coefficients in our predictor were obtained as analytical functions of only the fluctuation exponent, H (see Eqs. 24 and 25). As we showed in Appendix 2.3, if our model is adequate for describing the natural temperature variability, then the following relationship between the verification $\mathrm{TCC}_{\mathrm{nat}}$ and $\mathrm{MSSS}_{\mathrm{nat}}$ should be satisfied for $k = 1$ month:

$$\mathrm{TCC}_{\mathrm{nat}}(1) \approx \sqrt{\mathrm{MSSS}_{\mathrm{nat}}(1)}. \tag{10}$$

It does not hold for all horizons in the tropical region due to the use of the fBm rather than fGn model.

In Fig. 7 we show maps of the $\mathrm{TCC}_{\mathrm{nat}}$ and the absolute difference $\left|\mathrm{TCC}_{\mathrm{nat}} - \sqrt{\mathrm{MSSS}_{\mathrm{nat}}}\right|$ obtained from hindcasts for $k = 1$ month. The color scale in (b) was rescaled 100 times with respect to (a) so the differences could be perceptible. They are negligible compared to the values in (a). The maximum differences in Fig. 7b is almost always lower than 0.01 (mean value of 0.001), which strongly corroborates the adequacy of the fGn model to describe the natural variability.

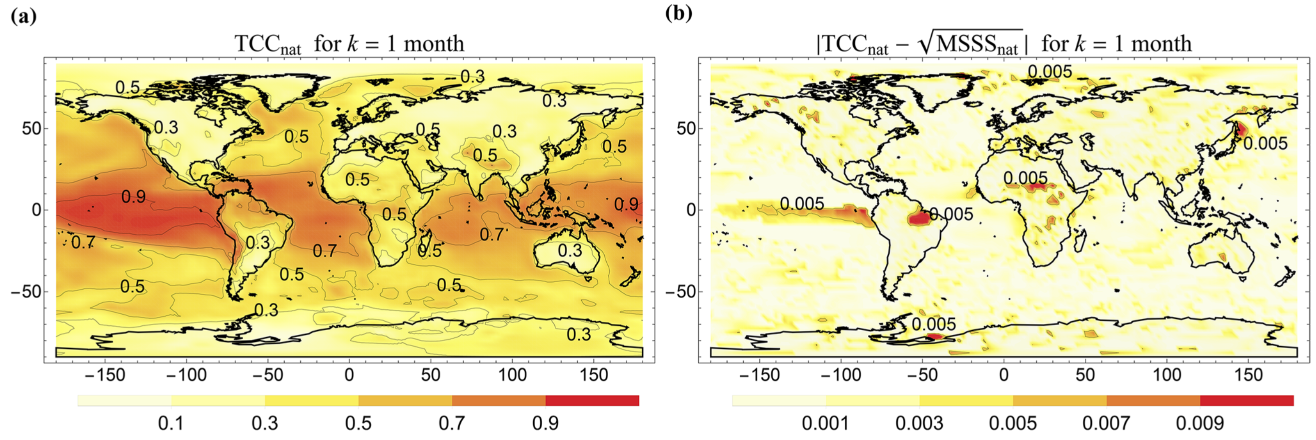

Fig. 7 Maps of $\mathrm{TCC}_{\mathrm{nat}}$ and the absolute difference $\left|\mathrm{TCC}_{\mathrm{nat}}-\sqrt{\mathrm{MSSS}_{\mathrm{nat}}}\right|$ obtained from hindcasts for $k=1$ month. The colour scale in **b** was rescaled 100 times with respect to **a** so the differences could be perceptible

3.1.2 Probabilistic scores and reliability

All the skill scores discussed above are recommended by the WMO for assessing deterministic prediction of long-range forecasts (WMO 2010b). These forecasts are deterministic in the sense that only the ensemble mean is considered, disregarding the ensemble variance, or more accurately, the prediction of the probability distribution. In this study we only focus on deterministic predictions (deterministic in the previously mentioned sense, recall that we use a stochastic model) because, given a Gaussian approximation of a probability distribution function, the skill of probabilistic forecasts is mainly dependent upon the skill of ensemble mean predictions and much less upon predictions of ensemble variances (Kryjov et al. 2006). In fact, in DRAL it was shown that, assuming a Gaussian distribution for the errors, the Continuous Ranked Probability Score (CRPS) (Hersbach 2000; Gneiting et al. 2005), which is a commonly used metric for probabilistic forecasts, is related to the RMSE by:

$$\mathrm{CRPS}(k)=\frac{\mathrm{RMSE}(k)}{\sqrt{\pi}}\left[\sqrt{2(1+\mathrm{ESS})}-\sqrt{\mathrm{ESS}}\right], \tag{11}$$

where:

$$\mathrm{ESS}=\frac{\overline{\sigma^2_{\mathrm{ensemble}}}}{\mathrm{MSE}} \tag{12}$$

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

For the case of StocSIPS, which by definition is a Gaussian model with ensemble spread $\sigma_{\mathrm{ensemble}}=\mathrm{RMSE}^{\mathrm{theory}}_{\mathrm{nat}}$ (given by Eq. (27)), the agreement between $\mathrm{RMSE}^{\mathrm{theory}}_{\mathrm{nat}}$ and $\mathrm{RMSE}_{\mathrm{nat}}$ (summarized in Fig. 6 for all horizons) implies that ESS ≈ 1 almost everywhere.

The graphs shown in Fig. 6 are analogous to spread-error scatterplots (Leutbecher and Palmer 2008). In our case, each point represents the ensemble spread and the temporal average RMSE for each pixel, instead of the spatially averaged values shown in Fig. 4 of Leutbecher and Palmer (2008). We could group up and average the values in equally populated bins to produce more similar spread-error plots, but as they all fall near to the reference diagonal, the conclusions would remain the same. Other measures used to assess the reliability [like the error-spread score (Christensen et al. 2015)] depend on the third or higher order moments of the forecast probability distribution. Since the StocSIPS forecast is Gaussian by definition, the ESS used here (Eq. 12) gives enough information assuming the near Gaussianity of the observational probability distribution.

In Fig. 8 we show maps of the ESS of StocSIPS for horizon from 1 to 4 months. Notice that, from Eq. (27), $\sigma_{\mathrm{ensemble}}=\mathrm{RMSE}^{\mathrm{theory}}_{\mathrm{nat}}$ is a function of the forecast horizon and the location, following the spatial distribution of the model parameters σ_T and H, but for all pixels the ESSs are very close to 1, except for the tropical ocean where it tends to be "overdispersive" (ESS > 1). The average values for the globe with one standard deviation are shown in brackets in each map label. They increase monotonically from 0.96 ± 0.05 for $k=1$ month, 0.98 ± 0.03 for $k=2$ months, 0.98 ± 0.04 for $k=3$ months, 1.00 ± 0.06 for $k=4$ months, up to 1.09 ± 0.21 for $k=12$ months (only the first four values are included in the maps). From Eq. (11), it can be shown that for a system with perfect reliability where ESS = 1, the CRPS takes its minimum value $\mathrm{CRPS}_{\mathrm{min}}=\mathrm{RMSE}/\sqrt{\pi}$. For

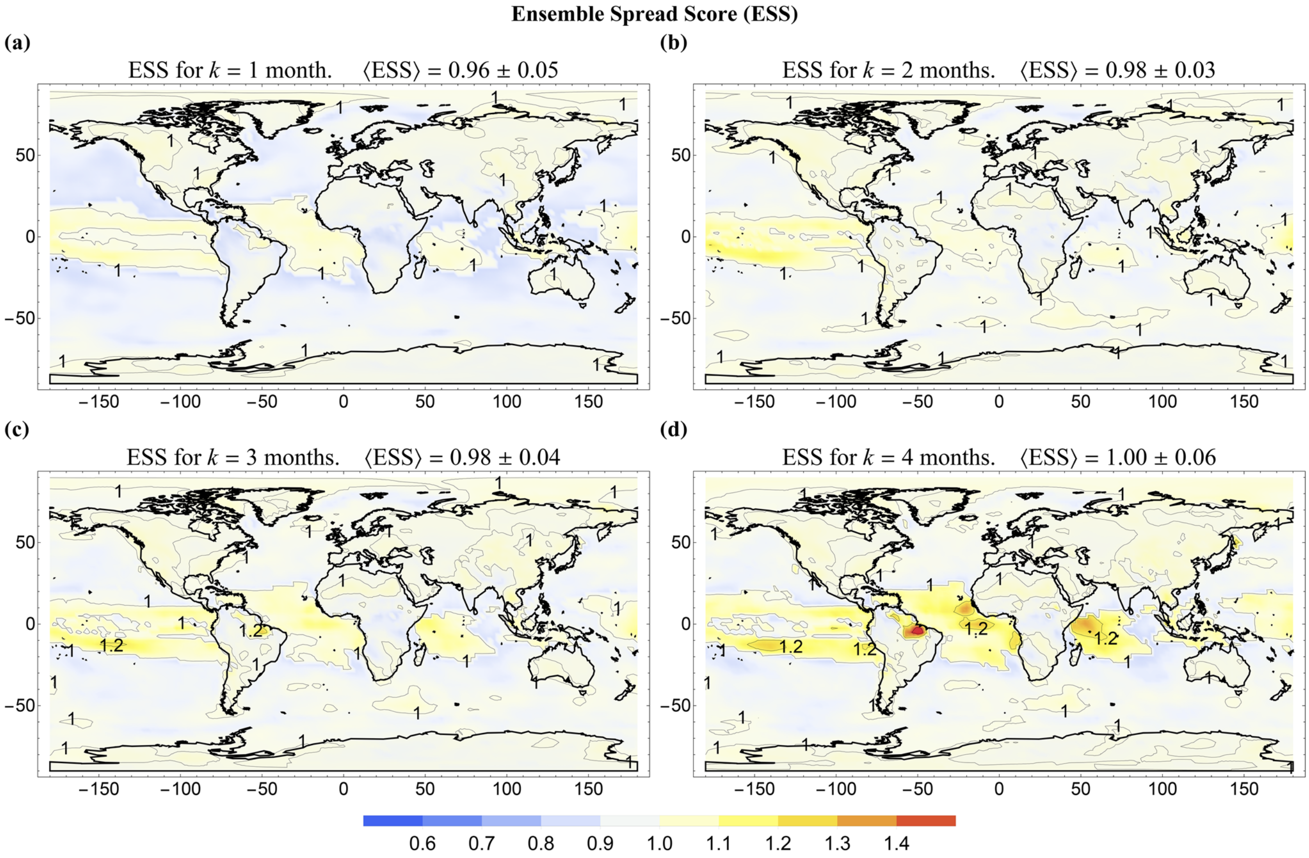
Ensemble Spread Score (ESS)
(a)
ESS for k = 1 month. ⟨ESS⟩ = 0.96 ± 0.05
(b)
ESS for k = 2 months. ⟨ESS⟩ = 0.98 ± 0.03
(c)
ESS for k = 3 months. ⟨ESS⟩ = 0.98 ± 0.04
(d)
ESS for k = 4 months. ⟨ESS⟩ = 1.00 ± 0.06
50
0
−50
−150
−100
−50
0
50
100
150
1
1.2
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4

Fig. 8 Maps of ESS of StocSIPS for horizons k from 1 to 4 months (**a**–**d**, respectively). The values of the ESS are very close to 1, with the exception of the tropical ocean where it tends to be "overdispersive" (ESS >1). The average values for the globe with one standard deviation are shown in brackets in the map labels

any other case when we have an "overconfident" (ESS <1) or an "overdispersive" (ESS > 1) system, CRPS > RMSE/$\sqrt{\pi}$. In conclusion, StocSIPS is a nearly perfectly reliable system (except for the tropical ocean) without need of recalibration of the forecast probability distribution.

3.2 Hindcast verification

3.2.1 Monthly and 3-month average predictions

The results presented in Sect. 3.1 confirm the validity of the stochastic model on forecasting the natural temperature variability. In this section, we show the verification scores for the forecast of the raw (undetrended) anomalies including the forecast of the CO$_2$eq deterministic trend. All the scores were computed following the definitions shown in Appendix 2.

Given the smooth variation of the CO$_2$eq concentration at monthly scales, we can use simple extrapolation in Eq. (2) to obtain the predictor $\widehat{T}_{\text{anth}}(t+k)$ from the knowledge of the CO$_2$eq concentration path up to time t. As the function $T_{\text{anth}}(t)$ is almost linear in a k-vicinity of any t, the error of projecting the anthropogenic component is negligible compared to the error of the natural variability. In fact, as we assume the same global CO$_2$eq forcing affecting all locations, the error of predicting the anthropogenic trend for a given k, is proportional to the sensitivity map shown in Fig. 4c. It was found that this error is lower than 2% of the RMSE of the natural variability for all locations and for all horizons. In any case, the projection of the trend was still included in the following verification results.

3.2.1.1 Normalized root mean square error (NRMSE) Figure 9 shows maps of the NRMSE for horizons k = 1, 2 and 3 months (panels a–c, respectively) and for the seasonal forecast (including all seasons, average for k=1–3 months) in panel (d). The values in brackets in the figure labels are the NRMSE globally area averaged over the grid points (see Eq. 39). In general, the skill of the forecasts is larger over ocean than over land, with the lower values of NRMSE attained over the tropical ocean. This corresponds to the distribution of H shown in Figs. 2 and 4a.

Since small NRMSE implies large skill, according to the global-averaged NRMSE, the seasonal skill is larger than that of any of the first three individual monthly forecasts.

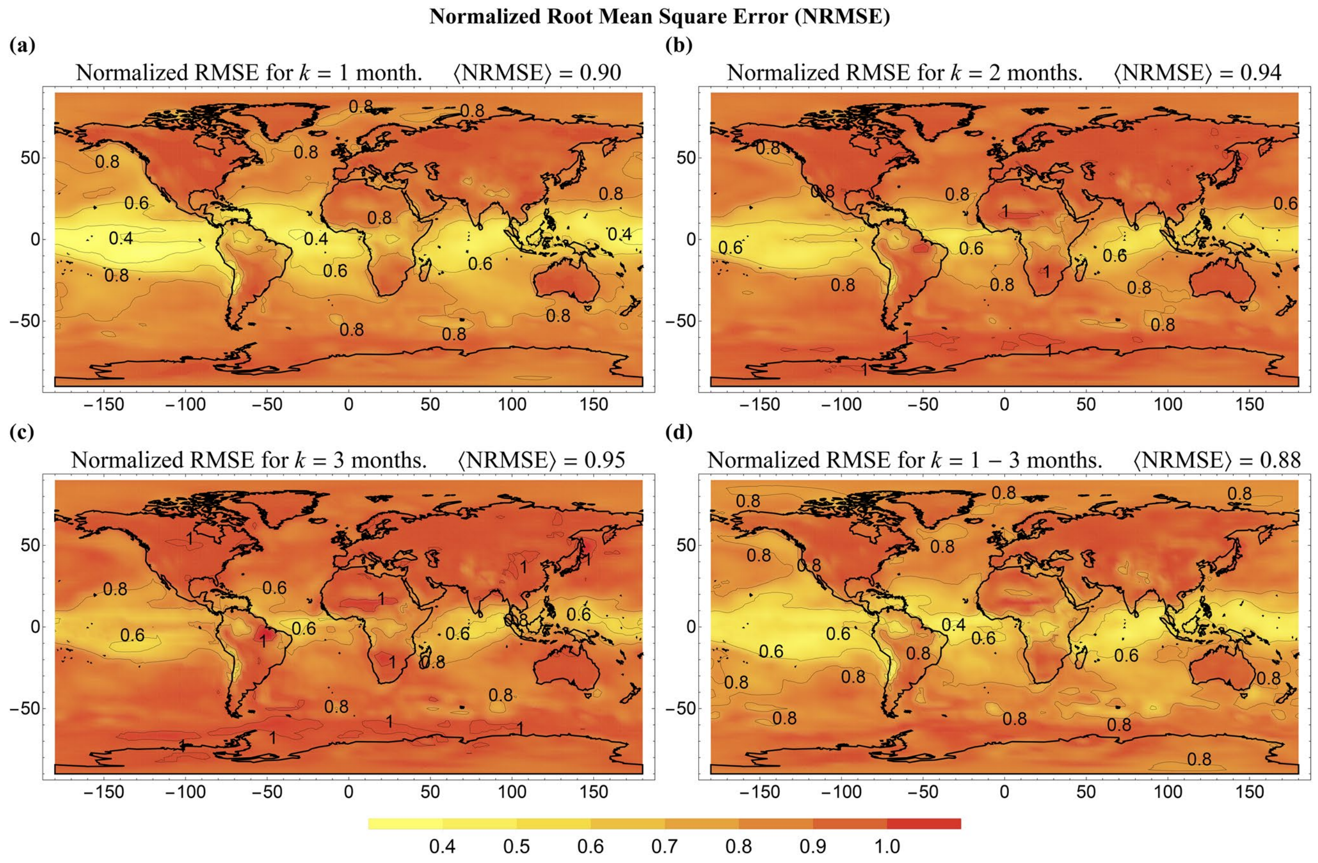
Normalized Root Mean Square Error (NRMSE)
(a)
Normalized RMSE for k = 1 month. ⟨NRMSE⟩ = 0.90
(b)
Normalized RMSE for k = 2 months. ⟨NRMSE⟩ = 0.94
(c)
Normalized RMSE for k = 3 months. ⟨NRMSE⟩ = 0.95
(d)
Normalized RMSE for k = 1 – 3 months. ⟨NRMSE⟩ = 0.88
50
0
−50
−150
−100
−50
0
50
100
150
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Fig. 9 Normalized root mean square error NRMSE for: (**a**) k = 1 month, (**b**) k = 2 months, (**c**) k = 3 months and (**d**) for the all-seasons mean (average for k = 1–3 months). The values in brackets in the figure labels represent the areal mean of global NRMSE

This is possible because although the horizon is further in the future, the seasonal forecast is for a longer (3 months) average. For scaling processes, the two effects exactly compensate. For the prediction of the natural variability component using fGn, the skill on predicting the next month using monthly averaged data is the same as the skill on predicting the next season using 3-month averaged data. This is reflected in Eq. (27), where k is in units of τ, which is the resolution (smallest sampling temporal scale) of the data. The similarity between the average values in the captions of panels (a) and (d) of Figs. 9, 10 and 11 confirms this consequence of the scaling.

The values in Fig. 9a for the forecast of the raw anomalies are lower than those shown in Fig. 5b for the natural variability because, while the RMSE of both are almost the same (we can neglect the error on projecting the anthropogenic trend), the normalization factor (standard deviation of the respective anomalies) is larger for the undetrended anomalies.

3.2.1.2 Mean Square Skill Score (MSSS) To compute the MSSS for the raw anomalies, the MSE of the reference climatology forecasts (forecast produced using only the annual cycle signal without removing the anthropogenic variation) is in this case:

$$\mathrm{MSE_C} = \left(\frac{N}{N-1}\right)^2 SD^2_{\mathrm{anom}}, \qquad (13)$$

where $SD_{\mathrm{anom}}{}^2$ is the variance of the anomalies series without removing the anthropogenic component:

$$SD^2_{\mathrm{anom}} = \overline{T^2_{\mathrm{anom}}} = \overline{\left(T_{\mathrm{anth}} + T_{\mathrm{nat}}\right)^2} = \overline{T^2_{\mathrm{anth}}} + SD^2_T \qquad (14)$$

(assuming that the natural and anthropogenic variabilities are independent).

Because $SD_{\mathrm{anom}}{}^2 > SD_T{}^2$ and the MSE of the forecast of the raw and the detrended anomalies are almost equal, then from Eq. (8) we obtain that the MSSS for the undetrended series forecast is larger than for the natural variability.

Maps of MSSS, corresponding to those shown in Fig. 9, are shown in Fig. 10. The difference in skill between ocean and land is more evident in these maps. In many places over land, the MSSS is close to zero, meaning that most of the skill comes from the projection of the anthropogenic trend. The global averages shown in brackets in the map labels are

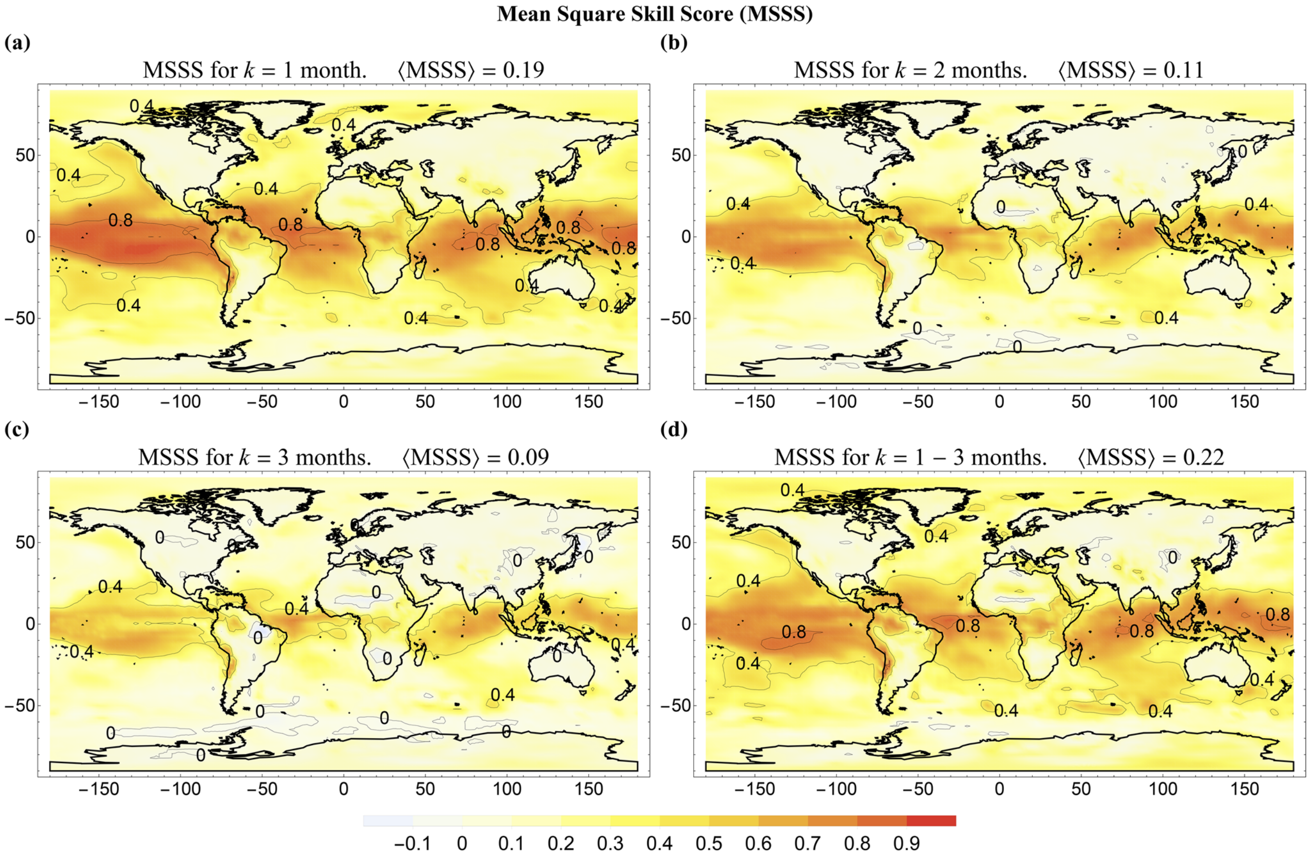
Mean Square Skill Score (MSSS)
(a)
MSSS for k = 1 month. ⟨MSSS⟩ = 0.19
(b)
MSSS for k = 2 months. ⟨MSSS⟩ = 0.11
(c)
MSSS for k = 3 months. ⟨MSSS⟩ = 0.09
(d)
MSSS for k = 1 − 3 months. ⟨MSSS⟩ = 0.22
50
0
−50
−150 −100 −50 0 50 100 150
0.4
0.8
0
−0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Fig. 10 Mean square skill score (MSSS) for: (**a**) k = 1 month, (**b**) k = 2 months, (**c**) k = 3 months and (**d**) for the all-seasons mean (average for k = 1–3 months). The values in brackets in the figure labels represent the areal mean of global MSSS

computed following the guidelines of the WMO (2010a). Note that the maps and the average values shown in Figs. 9 and 10 are related as MSSS ≈ 1 − NRMSE^2 if the reference forecast for the MSSS is the climatological annual cycle.

3.2.1.3 Temporal correlation coefficient (TCC)

Similarly to the MSSS, if the TCC is obtained for the undetrended anomalies (with only the annual cycle, but not with the anthropogenic trend removed), then, because of the extra correlation associated to the trend, higher values are often obtained compared to the TCC for the natural variability. For most of the long-term forecasts reported in the literature only the annual cycle is removed: the increasing trend related to anthropogenic warming is kept to obtain the anomalies used for verification, resulting in artificially boosted skill scores.

In Fig. 11, we show maps of the TCC for the prediction of the raw anomalies. The number in brackets in the caption of each plot indicates the area-averaged over the globe of the grid-point correlation coefficients. The area average was computed taking the Fisher Z-transform of the correlations following Eq. (42) (Fisher 1915; WMO 2010b). The StocSIPS predictions over the ocean are highly correlated with the observations and the highest correlations are in the tropical regions. Over land, although the skill is poorer (using the correlation coefficient), it is still significantly high for the forecast of the first 3 months. The TCC of the prediction is positive almost everywhere and, compared to the NRMSE or the MSSS, it shows significantly larger skill. This "extra" skill shown in the correlations for the raw anomalies comes from the presence of the anthropogenic signal.

3.2.2 Global averages

To summarize, in Fig. 12 we show graphs of the area-averaged NRMSE, MSSS and TCC for the monthly and the 3-month average forecasts as a function of the forecast horizon. In all the graphs, the red lines with circles correspond to the average considering the grid points for the whole planet, the blue lines with open squares are for places over the ocean and the green lines with triangles are for grid points over land. The corresponding dashed lines of the same colours represent the respective scores obtained if only the anthropogenic trend is forecast. In all cases, the reference forecast is the climatological annual cycle. Attending to the average values, we can conclude that the skill over ocean is always

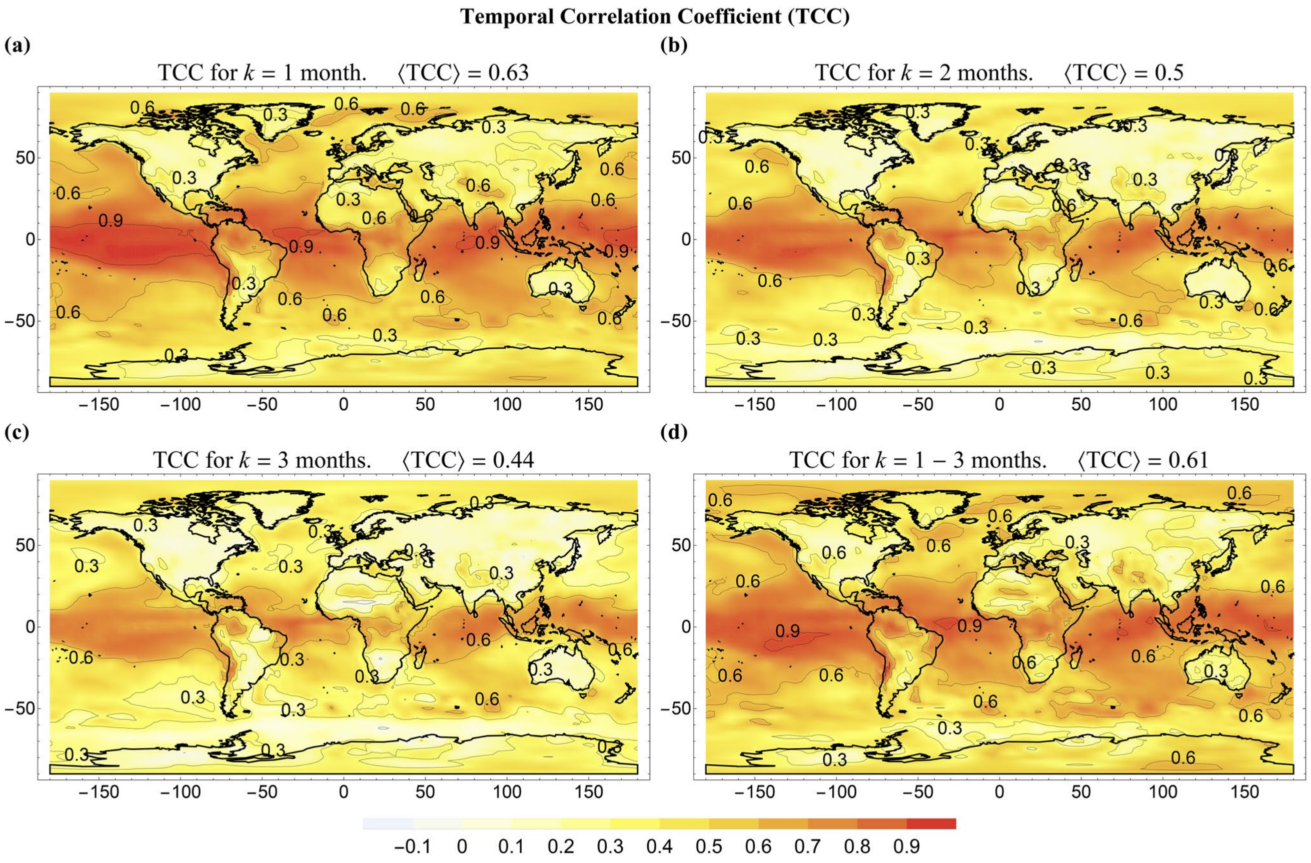
Temporal Correlation Coefficient (TCC)
(a)
TCC for k = 1 month. ⟨TCC⟩ = 0.63
(b)
TCC for k = 2 months. ⟨TCC⟩ = 0.5
(c)
TCC for k = 3 months. ⟨TCC⟩ = 0.44
(d)
TCC for k = 1 − 3 months. ⟨TCC⟩ = 0.61
50
0
−50
−150 −100 −50 0 50 100 150
0.3
0.6
0.9
−0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

Fig. 11 Anomaly correlation coefficient (TCC) for: (**a**) k = 1 month, (**b**) k = 2 months, (**c**) k = 3 months and (**d**) for the all-seasons mean (average for k=1–3 months). The values in brackets in the figure labels represent the areal mean of global TCC

greater than over land, with the global skill in between the two.

As we mentioned previously, for a perfectly scaling process, the 3-month average forecasts for k=1–3 months would have the same skill as the monthly forecast for k = 1 month. In the same way, the seasonal for k=4–6 months would correspond to the monthly for k = 2 months, for k=7–9 months to k = 3 months and for k=10–12 months to k = 4 months. A comparison between panels (a) and (d) and (b) and (e) in Fig. 12, show that this is reasonably well satisfied for the NRMSE and MSSS, respectively. Of course, the actual comparison should be made for the forecast of the natural temperature variability, which is the true scaling process. The results shown in Fig. 12 are for the raw anomalies which include the anthropogenic trend, that breaks the scale invariance of the fluctuations.

The difference between the curves and the dashed horizontal lines (showing the skill if only the anthropogenic trend is forecast) corresponds to the skill on forecasting the natural variability using the fGn model. While the forecast of the natural variability is reasonably skillful for $k \leq 6$ months, for horizons larger than 6 months, most of the overall skill comes from projecting the anthropogenic trend. Finally, note that the global and ocean averages vary monotonically with k, but the land averages show some oscillation that indicates a seasonality effect in the forecasts. This seasonality is analyzed in the next section.

3.2.3 Multiplicative seasonality

The results shown in panel (d) of Figs. 9, 10 and 11, were obtained for the seasonal forecast without distinguishing specific seasons. In fact, StocSIPS assumes that each month has the same anomaly statistics. It is actually 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 behaviour. This seasonal interannual variability is shown in Fig. 13, where the standard deviation of the 3-month averaged anomalies is shown for each conventional season: (a) December to February (DJF), (b) March to May (MAM), (c) June to August (JJA) and (d) September to November (SON). The difference in the variability between the spring and the fall seasons (panels (b) and (d)) is low. In comparison, the interannual variability over the land area in

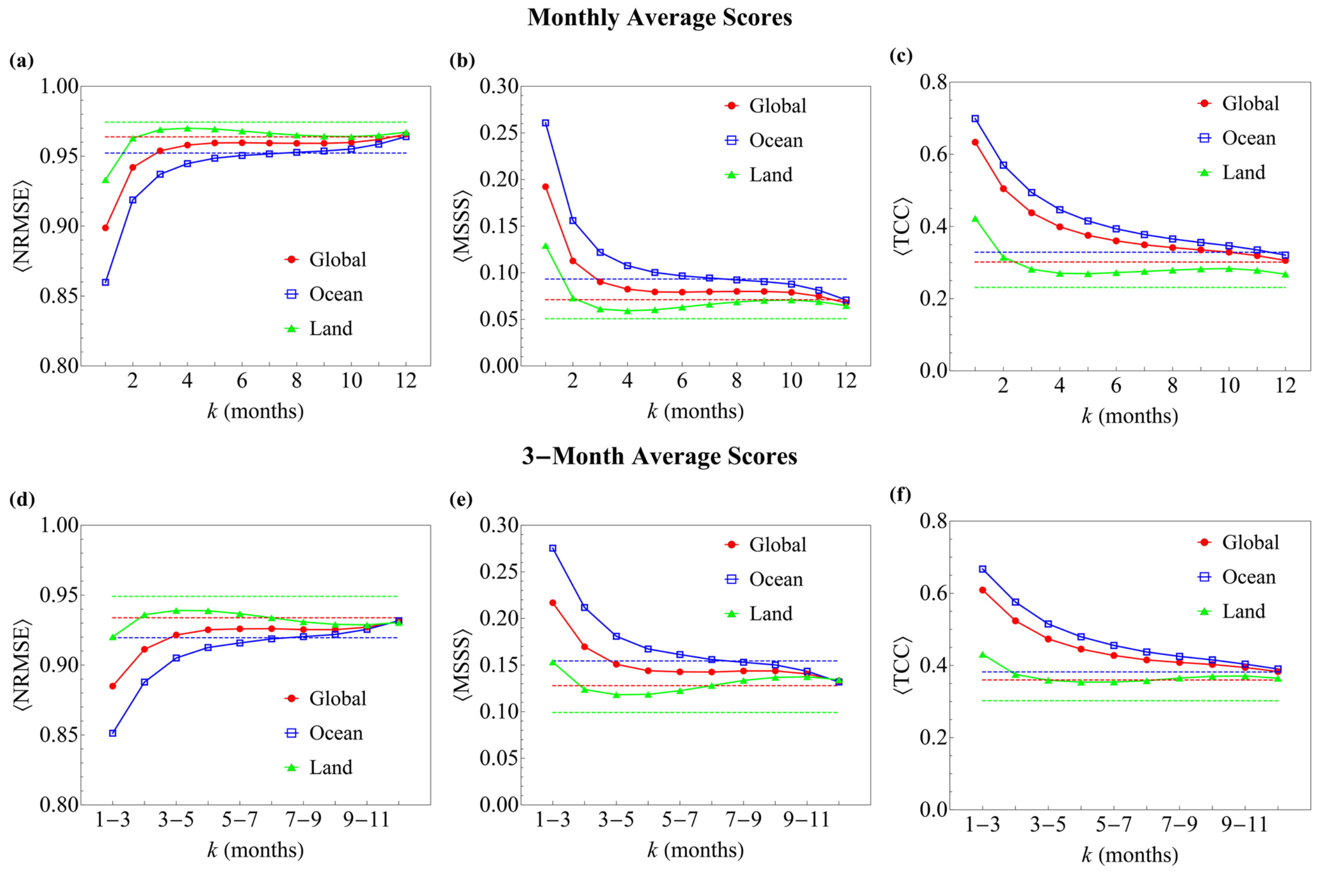

Fig. 12 Graphs of the area-averaged NRMSE, MSSS and ACC for the monthly (**a**–**c**) and the 3-month average (**d**–**f**) forecasts as a function of the forecast horizon. In all the graphs, the red lines with circles correspond to the average considering the grid points for the whole planet, the blue lines with open squares are for places over the ocean and the green lines with triangles are for grid points over land. The corresponding dashed lines of the same colours represent the respective scores obtained if only the anthropogenic trend is forecast

the northern hemisphere is larger during the boreal winter (DJF) and lower during the summer (JJA).

The largest seasonality is observed in the polar regions, where the winter temperatures are much more intermittent compared to the summer values. Conversely, during the summer the standard deviation of the anomalies in the poles is much lower compared to the other seasons. The values in brackets in the figure labels represent the areal mean of global standard deviation, $\langle \mathrm{SD} \rangle$, and the areal mean excluding the poles (between 60°S and 60°N), $\langle \mathrm{SD} \rangle_{-60}^{+60}$. Notice that the polar regions contribute substantially to the interannual variability and also, that the boreal winter season is in general significantly more variable than the other seasons (the $\langle \mathrm{SD} \rangle$ goes from 1.16 °C for DJF to roughly 1.03 °C for the others). A possible explanation for this seasonality is that, when removing the annual cycle and the trend associated with the anthropogenic warming, we assumed that both were statistically independent. This is not completely true for the polar region. While for the rest of the planet the anthropogenic temperature response increases uniformly for every month following the increasing CO$_2$ concentrations, in the poles during the summer, the temperature is tied to the freezing point of water. This is a shortcoming of the model that could be considered to improve future versions of StocSIPS.

3.2.4 Preliminary comparison with GCMs' seasonal predictions

In the previous sections, we validated StocSIPS as a good model for describing the monthly surface temperature field and we assessed its skill by computing monthly and 3-month average scores, without distinguishing specific seasons. To account for the effects of the multiplicative seasonality on the predictions, we can stratify the observations and the forecasts series to show dependencies with the targeted season and the forecast horizon. Usually, this is the kind of forecast published by several major operational centers for seasonal prediction. In this section we show the skill scores obtained for stratified data and we make a preliminary comparison with other models' skill to assess the relative advantages and shortcomings of StocSIPS.

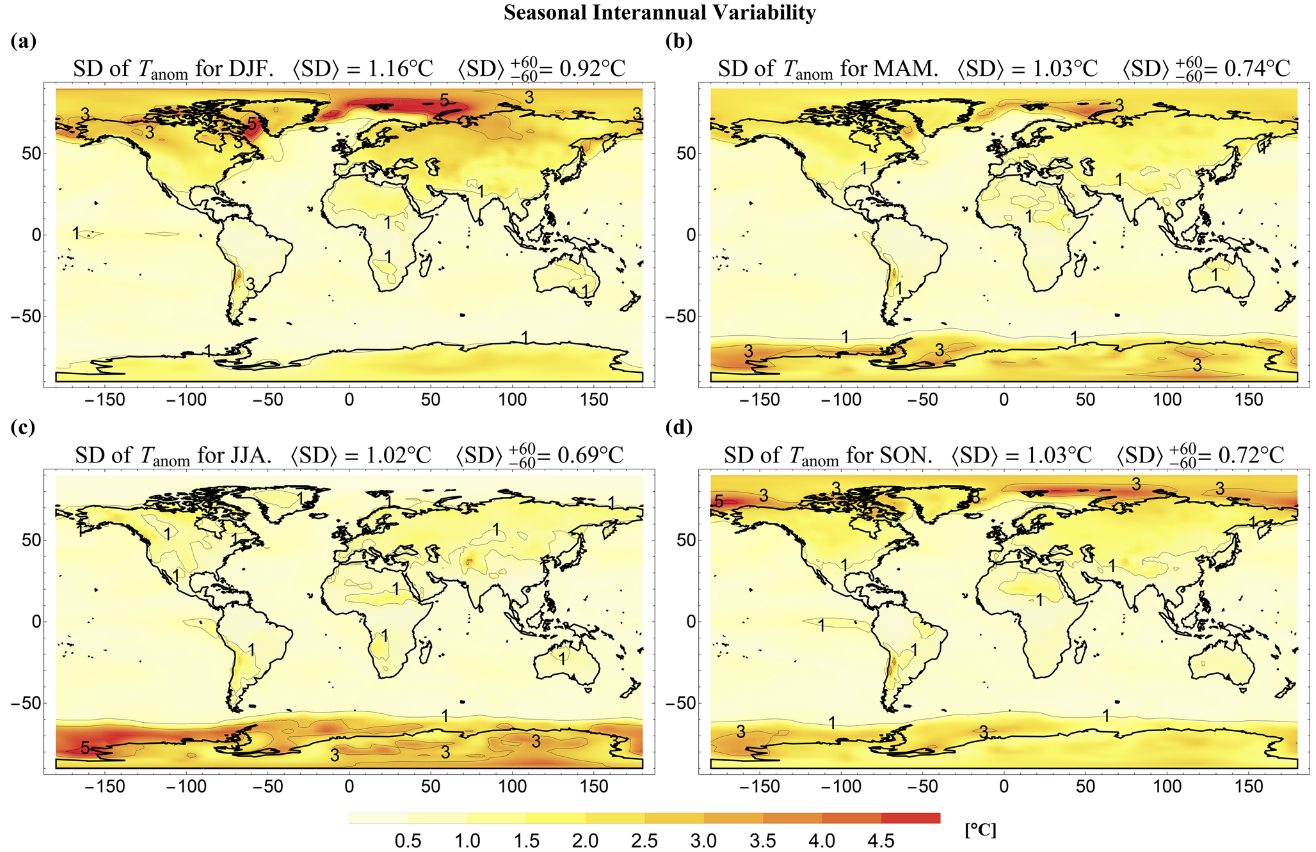
Seasonal Interannual Variability
(a)
SD of T_{anom} for DJF. ⟨SD⟩ = 1.16°C ⟨SD⟩$^{+60}_{-60}$= 0.92°C
(b)
SD of T_{anom} for MAM. ⟨SD⟩ = 1.03°C ⟨SD⟩$^{+60}_{-60}$= 0.74°C
(c)
SD of T_{anom} for JJA. ⟨SD⟩ = 1.02°C ⟨SD⟩$^{+60}_{-60}$= 0.69°C
(d)
SD of T_{anom} for SON. ⟨SD⟩ = 1.03°C ⟨SD⟩$^{+60}_{-60}$= 0.72°C
50
0
−50
−150 −100 −50 0 50 100 150
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5
[°C]

Fig. 13 Interannual standard deviation (SD) of the temperature anomalies for the conventional seasons: (**a**) DJF, (**b**) MAM, (**c**) JJA and (**d**) SON. The values in brackets in the figure labels represent the areal mean of global standard deviation and the areal mean excluding the poles (between 60°S and 60°N)

Our purpose in this paper is not to make an exhaustive and detailed comparison with other long-term prediction models' results. This detailed comparison and also the combination of StocSIPS with conventional numerical models to produce merged forecasts is the subject of a future paper currently in preparation. Those results are too extensive to include them in the present paper, so we limited ourselves to compare with already published skill scores from other models. For this purpose, we selected the multi-model ensemble (MME) predictions recently published by Kim et al. (2020). An important aspect is that Kim et al. offer a detailed description of the scores and the methods used, which we try to closely follow here to guarantee reproducibility. The definitions of these metric are given in Appendix 2, following the guidelines of the WMO Standardized verification system for long-range forecasts (SVS-LRF) (WMO 2010a, b).

In Kim et al. (2020), the authors assess different MME combination methods for seasonal prediction using hindcast datasets of six models from five Global Producing Centers (GPCs) for long-range forecasts (LRFs) designated by the WMO (Graham et al. 2011). The six models included in their analysis cover 27 years of common hindcast period from 1983 to 2009. The selected GPCs were: Melbourne, Montreal (two models), Moscow, Seoul and Tokyo. References and details of the individual models can be found in Kim et al. (2020). The authors study seven experimental deterministic MME methods to merge the six seasonal forecast systems: simple composite method (SCM), simple linear regression (SLR), multiple linear regression (MLR), best selection anomaly (BSA), multilayer perceptron (MLP), radial basis function (RBF) and genetic algorithm (GA). Their reported scores for 2-m temperature were obtained for 1-month lead retrospective forecasts in a grid with a resolution of 2.5° in both longitude and latitude. To produce the figures in this section, we used and adapted some of the figures from Kim et al. (2020) (including supporting information).

3.2.4.1 Mean Square Skill Score (MSSS) For a better comparison with Kim et al. results, all the seasonal scores for StocSIPS were obtained for observational and forecast seasonal anomalies calculated as departures from the climatology in the leave-one-out cross-validation scheme for the period 1983–2009. In Fig. 14, we show maps of the MSSS

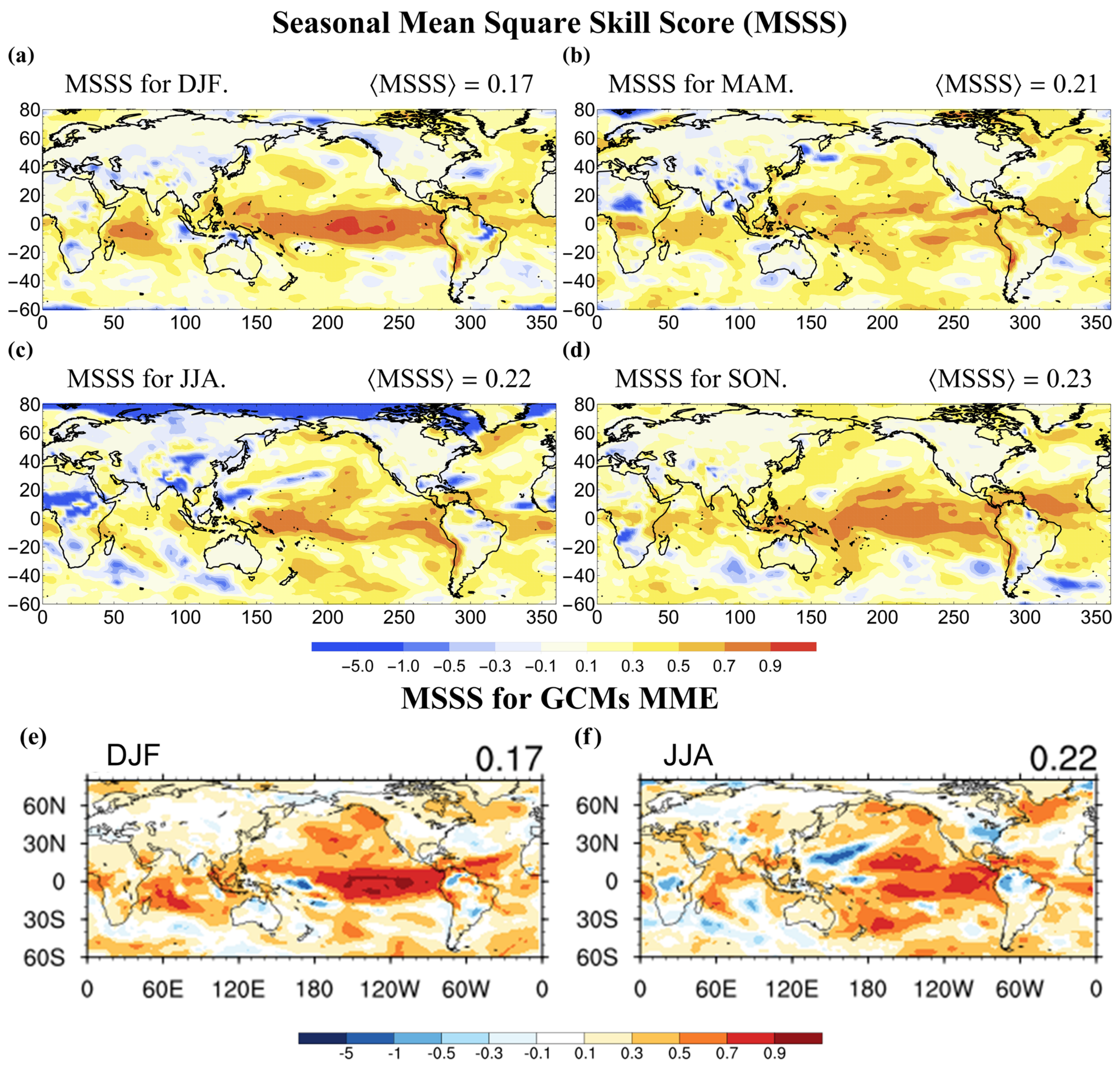

Fig. 14 MSSS for: (**a**) DJF, (**b**) MAM, (**c**) JJA and (**d**) SON. In all cases, the forecasts used data up to the beginning of each respective season (average for $k = 1–3$ months). The values in brackets in the figure labels represents the globally averaged MSSS (see Eq. (40)). The maps shown in **e** and **f** for the GCMs MME predictions of DJF and JJA, respectively, were reproduced from Figs. S1 and S2 of Kim et al. (2020) (supporting information) for their best MME combination method (GA). Kim et al. (2020) is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium

of StocSIPS for: (a) DJF, (b) MAM, (c) JJA and (d) SON. In all cases, the forecasts used data up to the beginning of each respective season (average for $k = 1–3$ months), i.e. including November for DJF, up to February for MAM and so on. The values in brackets in the figure labels represent the globally averaged score, ⟨MSSS⟩, (see Eq. 40). In panels (e) and (f) we reproduce maps of MSSS for DJF and JJA, respectively, from Figs. S1 and S2 of Kim et al. (2020) (supporting information) for their best MME combination method (GA).

In agreement with the previous results shown in Fig. 10 for the independent months and the all-season MSSS, the predictions for the individual seasons in general show better skill over the ocean than over land. The MSSS values are particularly high in the tropical region with the highest values obtained in the equatorial Pacific for DJF. Similar results were obtained for the GCM forecasts shown in Fig. 14e,f. The globally averaged scores (shown in the top right corner of each plot), are identical for StocSIPS and the MME results: 0.17 and 0.22 for DJF and JJA, respectively. The negative values of MSSS for StocSIPS near the north pole for JJA are associated to the multiplicative seasonality effect described in Sect. 3.2.3.

3.2.4.2 Temporal correlation coefficient (TCC) Similar to Fig. 14 for the MSSS, in Fig. 15 we show maps of the TCC of StocSIPS for: (a) DJF, (b) MAM, (c) JJA (d) SON and the best GCMs MME combination (GA) from Kim et al. (2020) in panels (e) and (f) for DJF and JJA, respectively. The values in brackets in the figure labels represent the globally averaged score, ⟨TCC⟩, computed using Eq. (42). As before, the highest correlation values are achieved in tropical

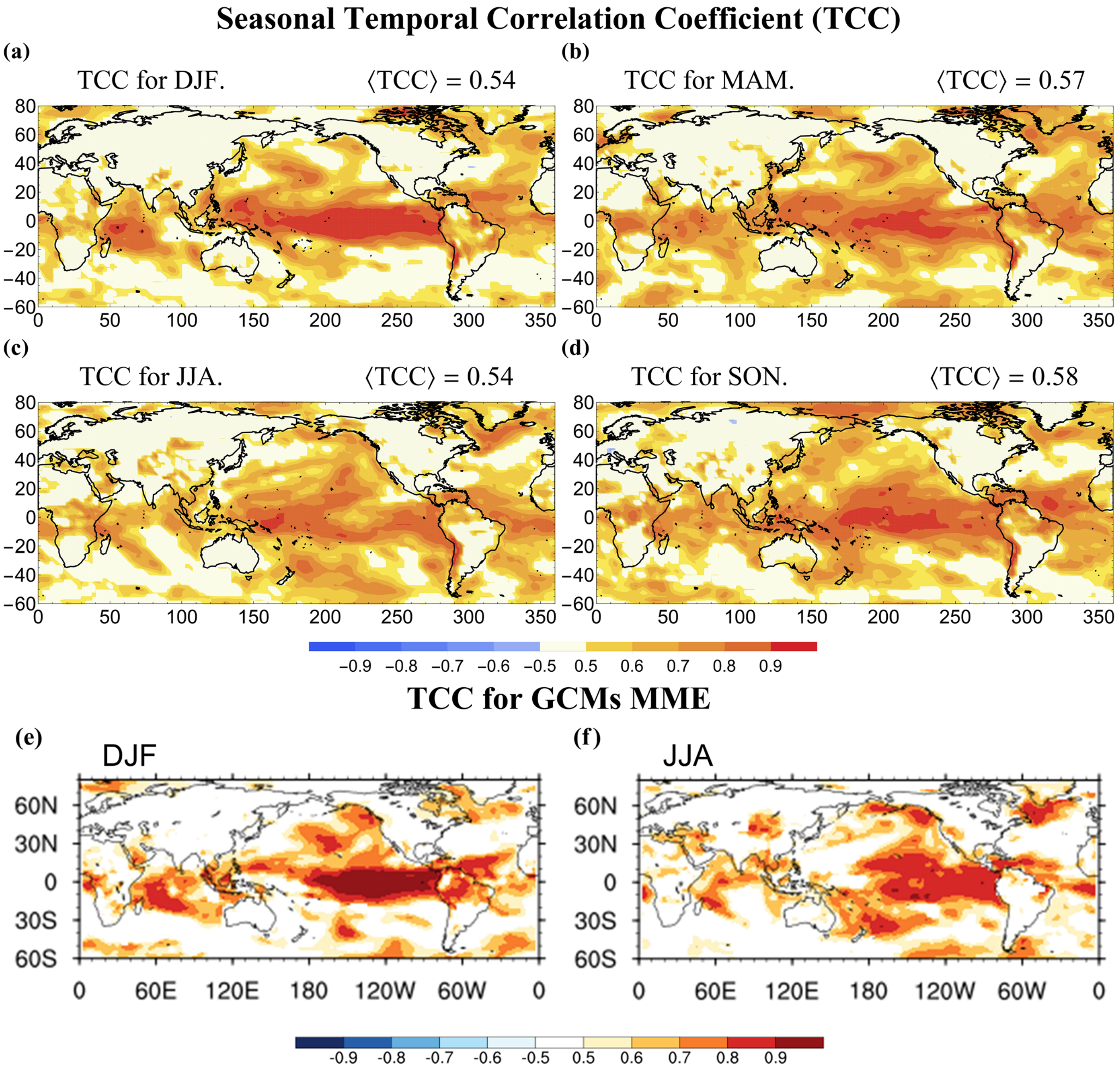

Fig. 15 TCC for: (**a**) DJF, (**b**) MAM, (**c**) JJA and (**d**) SON. In all cases, the forecasts used data up to the beginning of each respective season (average for $k=1$–3 months). The shaded areas indicate the regions over the 5% significance level using two-tailed student's t-test. The values in brackets in the figure labels represent the globally averaged score, ⟨TCC⟩, computed using Eq. (42). The maps shown in **e** and **f** for the GCMs MME prediction of DJF and JJA, respectively, were reproduced from Figs. S5 and S6 of Kim et al. (2020) (supporting information) for their best MME combination method (GA)

regions. Considering the average scores, there is no significant reduction in the TCC for DJF compared to JJA. There are also no considerably low values near the north pole for JJA. Compared to the MSSS, the multiplicative seasonality effects are less reflected in the TCC, since the latter is a measure of the skill in predicting the phase (sign), so less dependent on the variability of the anomaly magnitudes.

3.2.4.3 Anomaly pattern correlation coefficient (ACC) The temporal evolution of the forecast skill can be assessed using the anomaly pattern correlation coefficient (ACC), which is the spatial correlation for any given date between the observational and forecast anomalies (see Eq. 37). This shows how well the model reproduces the temperature anomaly distribution around the globe for any given season. Figure 16 shows the evolution of the ACC for StocSIPS (black line with solid circles) and for each of the seven MME combination methods studied by Kim et al. (colored lines with markers) in the 27-year verification period 1983–2009. Graphs for each season are shown independently: (a) MAM, (b) JJA, (c) SON and (d) DJF. This figure was adapted from Fig. 3 in Kim et al. (2020) to include the StocSIPS scores. In the original figure, the authors also show the absolute value of the El Niño 3.4 index (black line without markers) to study the dependence of the ensemble predictions with the El Niño phase. The main conclusion is that the GCM MMEs perform better during ENSO events than during non-ENSO events for all seasons. A similar behaviour was not found for the case of StocSIPS, where the performance based on the ACC varies independently of the ENSO phase. The average scores for the POV (see Eq. 43) are shown in the right panels for each of the respective seasons. Comparing these values, we can see that StocSIPS has better overall skill than most of the GCM MME combinations for all seasons. Only for JJA, the StocSIPS score is lower than the best three MME (SCM, SLR and GA). For the rest of the seasons, its average score is almost equal (or slightly larger) than the best MME (using GA or SCM).

3.2.4.4 Globally averaged TCC and RMSE Comparisons for globally averaged TCC and RMSE (see Eqs. (39) and (42)) are shown in Fig. 17a,b, respectively, for each season. The bars are for the MME combination methods in Kim et al. (2020), together with the mean of single model skills (MSMS). The scores for StocSIPS were included as horizontal lines with the same color code as the bars for each respective season. The dashed black line indicates that the estimated TCC is statistically significant at the 5% level using the one-tailed Student's t test. The GA methods shows the best performance, although it is very close to the SCM with equal weights for each model. Most MME predictions show higher skill than the corresponding MSMS for all four seasons, although sometimes (like the TCC for MLP), the MME combination does not improve over the single model predictions. In all cases, the TCC of the StocSIPS forecasts is larger than the best GCM MME. Similarly, the StocSIPS RMSE is lower than most of the MME combinations for all seasons. Only the SCM, GA and SLR show lower errors than StocSIPS for JJA predictions. The globally averaged TCC does not show a large seasonal variation, but there is still a reduction in skill for JJA and DJF associated to the high variability in the poles discussed in Sect. 3.2.3. This multiplicative seasonality effect is clear in the average RMSE, which follows the average SD values in the caption of Fig. 13.

For an overall comparison, in Fig. 18 we show a plot of the 4-season-averaged RMSE vs. TCC for the six individual models used in Kim et al. (2020) (red crosses) and the seven MME combinations (letters). The StocSIPS scores were included as a blue asterisk. For the GCMs, the GA method has the best performance—very close to the SCM—with the highest TCC (0.51) and the lowest RMSE (0.64). The StocSIPS forecasts have similar RMSE (0.64), but better average TCC (0.55).

4 Summary and discussion

In this paper we applied the Stochastic Seasonal to Interannual Prediction System (StocSIPS) to the monthly and seasonal prediction of the surface temperature with a 2.5° × 2.5° spatial resolution. The theory and the basis of the numerical methods used in StocSIPS were previously presented and applied to the forecast of globally averaged temperature in Del Rio Amador and Lovejoy (2019). StocSIPS is based on two statistical properties of the macroweather regime: the near Gaussianity of temperature fluctuations and the temporal scaling symmetry of the natural variability. The model is a high-frequency approximation to the Fractional Energy Balance Equation (FEBE), a fractional generalization of the usual EBE.

StocSIPS models the temperature series at each grid point independently 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 only one parameter: the climate sensitivity, itself obtained from linear regression with 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 three parameters for each grid

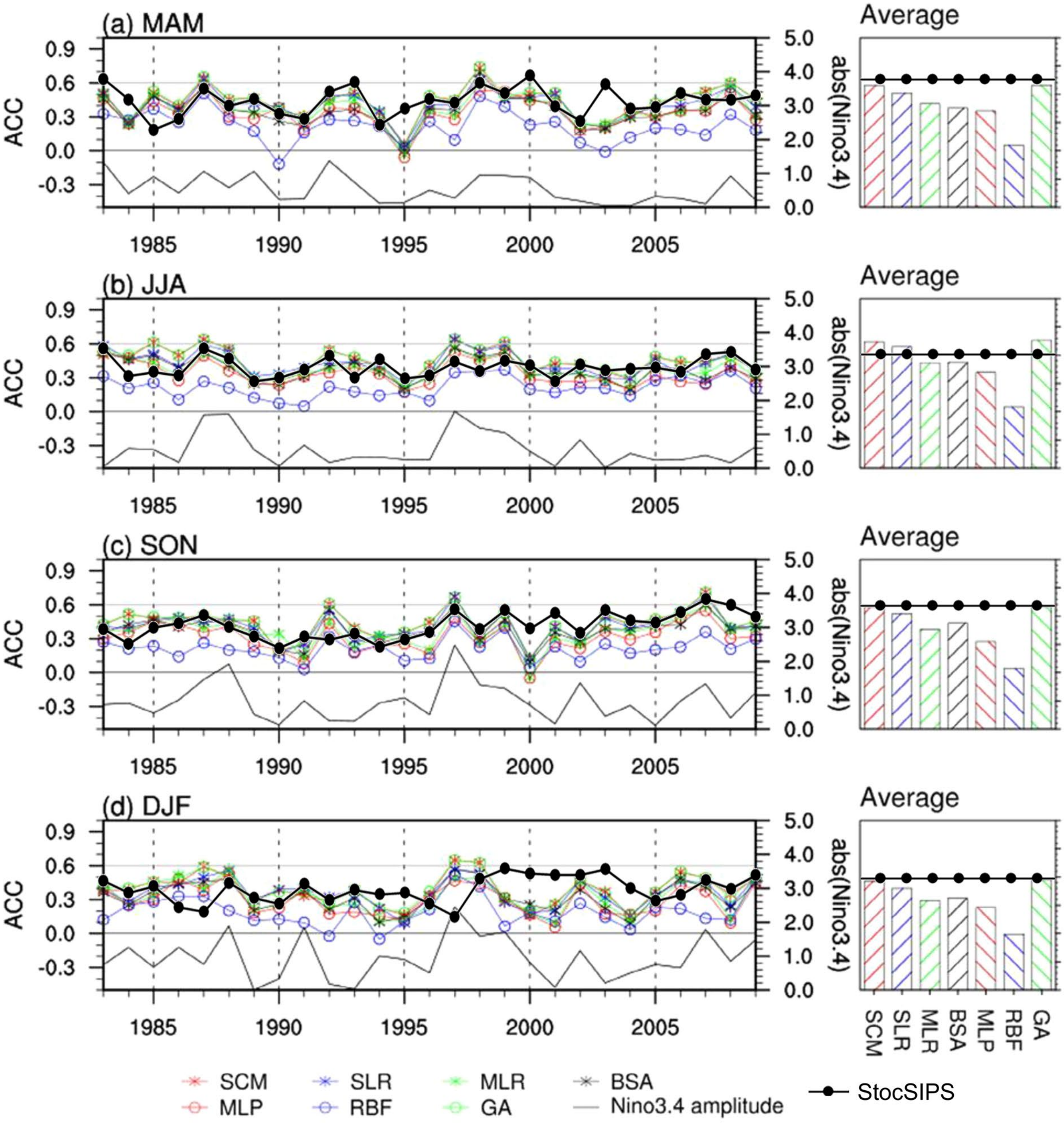

Fig. 16 ACC for StocSIPS (black line with solid circles) and for each of the seven MME combination methods studied by Kim et al. (colored lines with markers) in the 27-year verification period 1983–2009 for: (**a**) MAM, (**b**) JJA, (**c**) SON and (**d**) DJF. The average scores for the POV (see Eq. (43)) are shown in the right panels for each of the respective seasons. The absolute value of the El Niño 3.4 index (black line without markers) is also shown. This figure was adapted from Fig. 3 in Kim et al. (2020) to include the StocSIPS scores

point for modeling and predicting the surface temperature. Those parameters are quite robust and can be estimated with good accuracy from past data. The same procedure could be extended to any other field assuming it satisfies the Gaussianity and the scaling behaviour of the fluctuations.

Although we mentioned that the fGn with fluctuation exponent in the range $-1/2 < H < 0$ is a good model for the natural variability, a distinction must be made for most of the tropical ocean region, for which a positive fluctuation exponent was found. Instead of using the fGn model there, we must use the general fRn model or its high frequency fBm approximation with $0 < H < 1$. It is significant that within this tropical ocean region, only in the more predictable region that is associated with the ENSO phenomenon

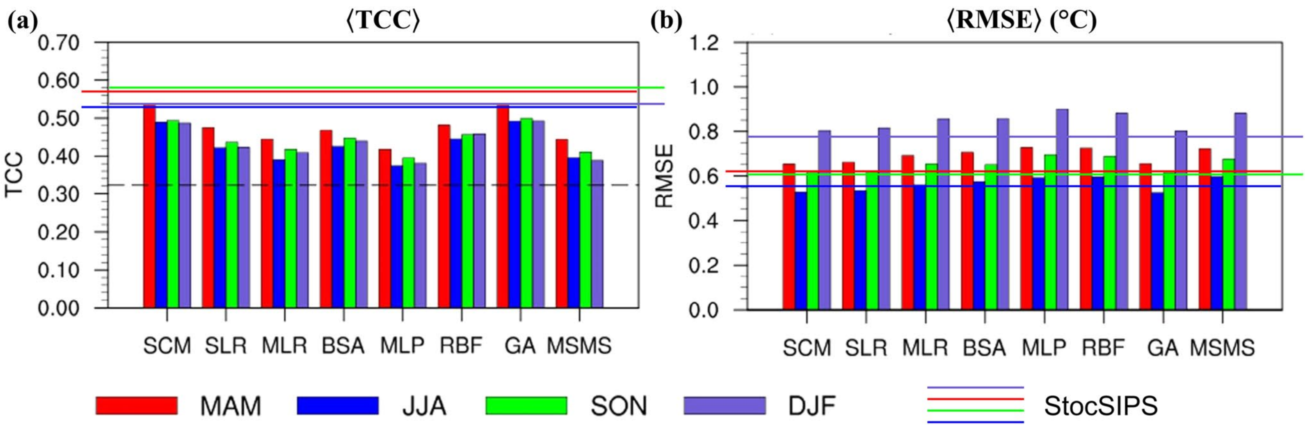

Fig. 17 Globally averaged TCC (**a**) and RMSE (**b**) (Eqs. (42) and (39), respectively) for MAM (red), JJA (blue), SON (green) and DJF (purple) for the period 1983–2009. The bars are for the MME combination methods in Kim et al. (2020), together with the mean of single model skills (MSMS). The scores for StocSIPS were included as horizontal lines with the same color code for each respective season. The dashed black line indicates that the estimated TCC is statistically significant at the 5% level using the one-tailed Student's t test. This figure was adapted from Figs. 5 and 6 in Kim et al. (2020) to include the StocSIPS scores

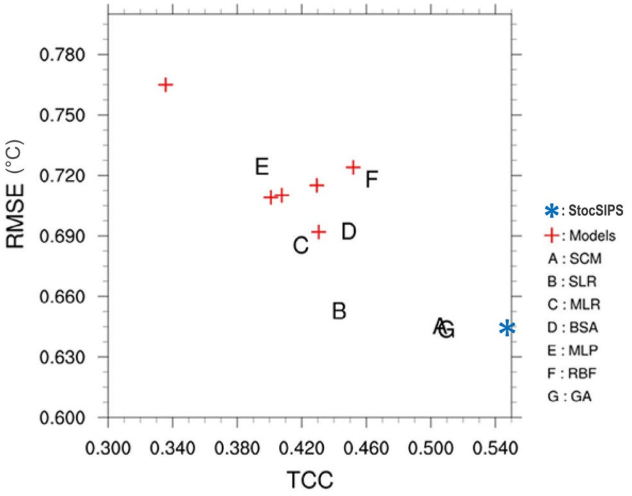

Fig. 18 4-season-averaged RMSE vs. TCC for the six individual models used in Kim et al. (2020) (red crosses), the seven MME combinations (letters) and StocSIPS (blue asterisk). This figure was adapted from Fig. 7 in Kim et al. (2020) to include the StocSIPS scores

we obtain fluctuation exponents in the range $1/2 < H < 1$, whose fBm approximation has persistent (positively correlated) increments.

It is surprising that by using only three parameters for each location (the fluctuation exponent, H, the standard deviation, σ_T, and the transient climate sensitivity, $\lambda_{2\times\mathrm{CO}_2\mathrm{eq}}$), we can build a model that accurately describes the temperature field. The adequacy of the model was verified by testing the whiteness of the residual innovations and validating the theoretically expected scores (if the model were perfect) vs. the actual hindcasts results. This also implies that, for probabilistic forecast, StocSIPS is a nearly perfectly reliable system without need of recalibration of the forecast probability distribution.

The hindcast verification results show that the skill is generally greater over the ocean than over land, in particular over the more persistent tropical ocean region. One of the implications of the scaling that was verified is that the 3-month average forecast has the same skill as the 1 month ahead monthly prediction. This is possible because although the horizon is further in the future, the seasonal forecast is for a longer (3 months) average. For scaling processes, the two effects exactly compensate.

The seasonal predictions show a decreased skill in the polar regions during the summer. A possible explanation for this seasonality is that, when removing the annual cycle and the trend associated with the anthropogenic warming, we assumed that both were statistically independent. This is not true for the polar region. While for the rest of the planet the anthropogenic temperature response increases uniformly for every month following the increasing CO_2 concentrations, in the poles during the summer, the temperature is tied to the freezing point of water. This spurious seasonality introduced in the preprocessing of the data, can be corrected in future versions of StocSIPS to improve the global forecasts.

Besides this seasonality near the poles, the globally averaged skill score values are also lower during the boreal winter. This can be explained by the asymmetric distribution of land mass between the northern and the southern hemispheres and the fact that the atmospheric temperature near the surface is more stable over ocean than over land. Further improvements in the model may be possible using recalibration of the individual forecasts for every season.

Although the purpose of this paper is not to make a detailed and exhaustive comparison with other long-term prediction models, it is important to at least show a preliminary comparison with already published skill scores from other models to assess the advantages and shortcomings of StocSIPS. The evaluation against seven different MME combination methods using six models from the Lead Centers for Long-range forecasts published by Kim et al. (2020), showed that the skill scores obtained with StocSIPS are comparable (or better in the case of the temporal correlation coefficient) than the best MME combination (which has larger skill than any individual ensemble member). This is in agreement with the previous results in Del Rio Amador and Lovejoy (2019) that show that StocSIPS outperformed the Canadian MME (CanSIPS) for all but the first month of the forecast. This preliminary comparison for seasonal forecast validates StocSIPS as a good alternative and a complementary option to conventional numerical models.

StocSIPS and GCMs are based on entirely different approaches. While the GCMs only take the initial state of the system (with perturbations to produce multiple ensemble realizations), they do exploit all possible interactions with other atmospheric variables and other locations to produce their forecast through the integration of the dynamical equations. Conversely, StocSIPS neglects all the spatial (and other variables) relations to produce forecasts based on the past states at any single location by exploiting the large memory of the system. Another way to view this is that for forecasts, GCMs are initial value models that generate many “stochastic” realizations of the state of the atmosphere, whereas StocSIPS is effectively a “past value problem” that directly estimates the most probable future state (conditional expectation).

Although there is no evident mechanism that explains how the distant past affects the current state of the system, model reduction as explained by the Mori-Zwanzig formalism (Mori 1965; Zwanzig 1973, 2001; Gottwald et al. 2017) shows that if we only look at one part of the system (e.g. the temperature at a given location), memory effects arise. All the interactions coming from other degrees of freedom are embedded in the past values. Recent works (Lovejoy et al. 2015; Lovejoy 2019; Lovejoy et al. 2021) hypothesize that, for the case of temperature, scaling behaviour are a result of a hierarchy of energy storage mechanisms acting at different temporal and spatial scales. In addition, it was shown that a scaling Mori-Zwanzig effect naturally arises even from the classical heat equation (Lovejoy 2021a, b).

One evident question that arises from our treatment is why not to exploit the teleconnections in the temperature field to improve the forecast instead of predicting each series independently. To answer this, in a recent publication (Del Rio Amador and Lovejoy 2021), StocSIPS was extended to the multivariate case (m-StocSIPS), to include and realistically reproduce all the space–time cross-correlation structure. By using Granger causality, it was shown that, although large spatial correlations exist in the temperature field, the optimal predictor of the temperature at a given location is obtained from its own past if long enough time series are given. These cross-correlations “were already used” to build that past. This means that the predictions given here (in the univariate StocSIPS version) are optimal in this stochastic framework. Improvements on the MSSS values of only 1–2% are possible, which is in the noise level of our current predictions, so in that sense, the forecast of the individual series is nearly optimal. Nevertheless, the fact that the GCMs remain “deterministic” up to approximately 1–2 years over the oceans (mostly in the tropics) and in the poles, where having a dynamic sea ice model is apparently crucial for subseasonal to seasonal forecasts (Zampieri et al. 2018), suggests that StocSIPS can be combined with GCM outputs to produce a single hybrid forecasting system that improves on both.

Appendix 1: Basic theory for fGn processes

Continuous-in-time fGn

In DRAL, the stochastic natural variability component of the globally averaged temperature was represented as an fGn process. The main properties of fGn relevant for the present paper are summarized in the following.

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

$$T_\tau(t) = \frac{1}{\tau}\frac{c_H\sigma_T}{\Gamma(H+3/2)}\left[\int_{-\infty}^{t}\left(t-t'\right)^{H+1/2}\gamma\left(t'\right)dt' - \int_{-\infty}^{t-\tau}\left(t-\tau-t'\right)^{H+1/2}\gamma\left(t'\right)dt'\right], \tag{15}$$

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

$$c_H^2 = \frac{\pi}{2\cos(\pi H)\Gamma(-2-2H)}. \tag{16}$$

This is the canonical value for the constant c_H that was chosen to make the expression for the statistics particularly simple. In particular, the variance is $\langle T_\tau(t)^2\rangle = {\sigma_T}^2\tau^{2H}$ for all t, where $\langle\bullet\rangle$ denotes ensemble (infinite realizations) averaging. The parameter H, with $-1 < H < 0$, is the fluctuation exponent of the corresponding fractional Gaussian noise process, the Hurst exponent, $H' = H + 1$. Fluctuation exponents are used due to their wider generality; they are well defined even for strongly intermittent non-Gaussian multifractal processes and they can be any real value. For a discussion, see page 643 in Lovejoy et al. (2015).

Equation (15) can be interpreted as the smoothing of the fractional integral of a white noise process or as the power-law weighted average of past innovations, $\gamma(t)$. This power-law weighting accounts for the memory effects in the temperature series. The closer the fluctuation exponent is to zero, the larger is the influence of past values on the current temperature. This is evidenced by the behaviour of the autocorrelation function:

$$R_H(\Delta t) = \frac{\langle T_\tau(t) T_\tau(t+\Delta t)\rangle}{\langle T_\tau(t)^2\rangle} = \frac{1}{2}\left(\left|\frac{\Delta t}{\tau}+1\right|^{2H+2} + \left|\frac{\Delta t}{\tau}-1\right|^{2H+2} - 2\left|\frac{\Delta t}{\tau}\right|^{2H+2}\right), \tag{17}$$

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

$$R_H(\Delta t) \approx (H+1)(2H+1)\left(\frac{\Delta t}{\tau}\right)^{2H}, \tag{18}$$

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

Discrete-in-time fGn

A detailed explanation of the theory for modeling and predicting using the discrete version of fGn processes was presented in DRAL; the main results are summarized next. The analogue of Eq. (15) in the discrete case for a finite series, $\{T_t\}_{t=1,\ldots,N}$, with length N and zero mean is:

$$T_t = \sum_{j=1}^{t} m_{tj}\gamma_{t+1-j} = m_{t1}\gamma_t + \cdots + m_{tt}\gamma_1, \tag{19}$$

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

$$\mathbf{C}^N_{H,\sigma_T} = \mathbf{M}^N_{H,\sigma_T}\left(\mathbf{M}^N_{H,\sigma_T}\right)^T, \tag{20}$$

with $m_{ij} = 0$ for $j > i$ (we assume $\tau = 1$ is the smallest scale in our system). The superscript T denotes transpose operation.

In vector form, Eq. (19) can be written as:

$$\mathbf{T}_N = \mathbf{M}^N_{H,\sigma_T}\boldsymbol{\gamma}_N \tag{21}$$

Equations (19–21) can be used to create synthetic samples of fGn with a given length N, autocorrelation function given by Eq. (17) and set of parameters $\sigma_T > 0$ and $-1 < H < 0$ (the mean of the series is always assumed equal to zero). Conversely, given an actual temperature series with vector $\mathbf{T}_N = \left[T_1, \ldots, T_N\right]^T$, we can estimate the parameters σ_T and H using the maximum likelihood method (details are given in Appendix 1 of DRAL) and we can verify that it could be well approximated by an fGn model by inverting Eq. (21) and obtaining the residual vector of innovations:

$$\boldsymbol{\gamma}_N = \left(\mathbf{M}^N_{H,\sigma_T}\right)^{-1}\mathbf{T}_N. \tag{22}$$

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

fRn correlation function for 0 < *H* < 1

The fractional Relaxation noise (fRn) process was introduced in Lovejoy (2019) generalizing both fGn, fBm and Ornstein–Uhlenbeck processes. For short time scales (compared to some characteristic relaxation time, τ_r) and for exponents $-1/2 < H < 0$, the fRn is close to an fGn process. For fluctuation exponents in the range $0 < H < 1$ the high-frequency approximation to fRn is no longer an fGn process. In this case, to leading order, the correlation function is:

$$R_{\mathrm{fRn}}(\Delta t) = 1 - A_H\left(\frac{\Delta t}{\tau_r}\right)^{2H} + O\left(\frac{\Delta t}{\tau_r}\right)^{3H+1/2};\ \ \Delta t < \tau_r;\ 0 < H < 1 \tag{23}$$

where τ_r is the relaxation time and A_H is an H-dependent numerical factor [see (Lovejoy 2019)]. The same correlation function was obtained by Delignières (2015) as an approximation to short segments of discrete-in-time fractional Brownian motion (fBm) process that is the integral of an fGn process (but with H increased by 1). This shows that although fBm is nonstationary, short segments approximate (the stationary) fRn process. When $0 < H < 1$, fBm is a high-frequency approximation to an fRn process.

Prediction

In DRAL it was shown that, if $\left\{T_t\right\}_{t<0}$ is an fGn process, the optimal k-steps predictor for T_k $(k > 0)$, based on a finite number, m (memory), of past values, is given by:

$$\hat{T}_k = \sum_{j=-m}^{0} \phi_j(k)T_j = \phi_{-m}(k)T_{-m} + \cdots + \phi_0(k)T_0, \tag{24}$$

where the vector of predictor coefficients, $\boldsymbol{\phi}(k) = \left[\phi_{-m}(k), \ldots, \phi_0(k)\right]^T$, satisfies the Yule–Walker equations:

$$\mathbf{R}_H\boldsymbol{\phi}(k) = \mathbf{r}_H(k), \tag{25}$$

with the vector $\mathbf{r}_H(k) = \left[R_H(k-i)\right]^T_{i=-m,\ldots,0} = \left[R_H(m+k), \ldots, R_H(k)\right]^T$ and $\mathbf{R}_H = \left[R_H(i-j)\right]_{i,j=-t,\ldots,0}$ being the autocorrelation matrix (see Eq. 17). In those regions with consecutive values positively correlated (blue regions in Fig. 4a with $-1/2 < H < 0$ or the increments in the yellow region with$1/2 < H < 1$), the elements $R_H(\Delta t)$ are obtained from Eq. (17). In the places with consecutive increments negatively correlated, where $0 < H < 1/2$ (red in Fig. 4a), instead of forecasting the fGn increments, we forecast directly the fRn process and we get the elements $R_H(\Delta t)$ from Eq. (23). To use this autocorrelation for fRn, we estimate the constant A_H in Eq. (23) for each location by fitting the empirical autocorrelation function.

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

$$\mathrm{RMSE}(k,m) = \sqrt{\left\langle\left[T_k - \hat{T}_k(m)\right]^2\right\rangle}. \tag{26}$$

Following the results presented in DRAL and using that, for positive H the fRn is the integral of the corresponding fGn process, we obtain the following analytical expression for the RMSE of the predictor of the natural variability component:

$$\mathrm{RMSE}^{\mathrm{theory}}_{\mathrm{nat}}(k) = \mathrm{RMSE}\left(k,m,\sigma_T,H\right) = \begin{cases} \sigma_T\sqrt{1-\mathbf{r}_H(k)^T\left(\mathbf{R}_H\right)^{-1}\mathbf{r}_H(k)}; & \text{for } -1/2 < H < 0 \\ \sigma_T k^H\sqrt{1-\mathbf{r}_{H-1}(1)^T\left(\mathbf{R}_{H-1}\right)^{-1}\mathbf{r}_{H-1}(1)}; & \text{for } 0 < H < 1 \end{cases}. \tag{27}$$

For a given forecast horizon, k, the RMSE only depends on the parameters σ_T and H, and the memory used, m. In Fig. 3 of DRAL it was shown that only a few past datapoints are needed as memory to obtain an error approaching—with more than 95% agreement—the asymptotical value corresponding to $m = \infty$, for all possible values of H.

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

$$\mathrm{MSSS}(k) = 1 - \frac{\left\langle\left[T(k) - \hat{T}(k)\right]^2\right\rangle}{\left\langle T(k)^2\right\rangle} \tag{28}$$

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

From the definition of the RMSE, Eq. (26), we obtain the theoretical value for fGn:

$$\mathrm{MSSS}^{\mathrm{theory}}_{\mathrm{nat}}(k) = \mathrm{MSSS}(k,m,H)= 1 - \frac{\mathrm{RMSE}\left(k,m,\sigma_T,H\right)^2}{\sigma_T^2} \tag{29}$$

or, replacing Eq. (27) for $-1/2 < H < 0$:

$$\mathrm{MSSS}(k,m,H) = \mathbf{r}_H(k)^T\left(\mathbf{R}_H\right)^{-1}\mathbf{r}_H(k) = \boldsymbol{\phi}(k)\cdot\mathbf{r}_H(k). \tag{30}$$

In Fig. 19 we show graphs of the theoretical MSSS as a function of H for different values of k. A memory $m = 50$ was used for computing the MSSS. As expected, the skill decreases as the forecast horizon increases. For $H = -0.5$, the fGn process is a white noise process and $\mathrm{MSSS} = 0$. The skill increases with H and (with infinite past data) the process becomes perfectly predictable when $H \to 0$.

Appendix 2: Verification metrics

Definitions

The verification metrics used in this paper were defined following the recommendations in the Standardized verification system for long-range forecasts (SVS-LRF) for the practical details of producing and exchanging appropriate verification scores (WMO 2010a, b). Let $x_i(t)$ and $f_i(t)$, $(t = 1, \ldots, N)$ denote time series of observations and forecasts, respectively, for a grid point i over the period of verification (POV) with N time steps. Then, their averages for the POV, $\bar{x}_i$ and $\bar{f}_i$ and their sample variances ${s_{x_i}}^2$ and ${s_{f_i}}^2$ are given by:

$$\begin{aligned} \bar{x}_i &= \frac{1}{N}\sum_{t=1}^{N} x_i(t), \quad \bar{f}_i = \frac{1}{N}\sum_{t=1}^{N} f_i(t) \\ s^2_{xi} &= \frac{1}{N}\sum_{t=1}^{N}\left[x_i(t) - \bar{x}_i\right]^2, \quad s^2_{fi} = \frac{1}{N}\sum_{t=1}^{N}\left[f_i(t) - \bar{f}_i\right]^2. \end{aligned} \tag{31}$$

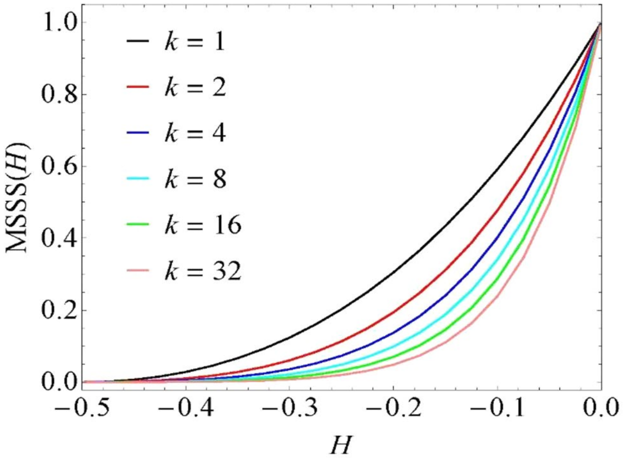

Fig. 19 Graphs of the theoretical MSSS (Eq. 46) as a function of H for different values of k. A memory $m = 50$ was used for computing the MSSS

The mean square error (MSE) of the forecast for grid point i is:

$$\mathrm{MSE}_i = \frac{1}{N}\sum_{t=1}^{N}\left[f_i(t) - x_i(t)\right]^2 \tag{32}$$

and the root mean square error (RMSE) is:

$$\mathrm{RMSE}_i = \sqrt{\mathrm{MSE}_i}. \tag{33}$$

For leave-one-out cross-validated data in the POV (WMO 2010a), the MSE of climatology forecasts is:

$$\mathrm{MSE}_{\mathrm{C}i} = \left(\frac{N}{N-1}\right)^2 s_{xi}^2. \tag{34}$$

The mean square skill score (MSSS) for grid point i, taking as reference the climatology forecast, is defined as:

$$\mathrm{MSSS}_i = 1 - \frac{\mathrm{MSE}_i}{\mathrm{MSE}_{\mathrm{C}i}}. \tag{35}$$

The temporal correlation coefficient (TCC) is:

$$\mathrm{TCC}_i = \frac{\frac{1}{N}\sum_{t=1}^{N}\left[x_i(t) - \bar{x}_i\right]\left[f_i(t) - \bar{f}_i\right]}{s_{xi}s_{fi}}. \tag{36}$$

Both the MSE_i and the TCC_i are computed using temporal averages for a given location i, conversely, the anomaly pattern correlation coefficient (ACC) (Jolliffe and Stephenson 2011) is defined using spatial averages for a given time t:

$$\mathrm{ACC}(t) = \frac{\sum_{i=1}^{n}\cos\theta_i\left[x_i'(t) - \langle x'(t)\rangle\right]\left[f_i'(t) - \langle f'(t)\rangle\right]}{\sqrt{\sum_{i=1}^{n}\cos\theta_i\left[x_i'(t) - \langle x'(t)\rangle\right]^2}\sqrt{\sum_{i=1}^{n}\cos\theta_i\left[f_i'(t) - \langle f'(t)\rangle\right]^2}}, \tag{37}$$

where n is the number of grid points, θ_i is the latitude at location i, $x_i^{'}(t)$ and $f_i^{'}(t)$ are observation and forecast anomalies for the POV, respectively, and the spatial averages $x'(t)$ and $f'(t)$ are given by:

$$\left\langle x'(t)\right\rangle = \frac{\sum_{i=1}^{n}\cos\theta_i x_i'(t)}{\sum_{i=1}^{n}\cos\theta_i}, \quad \left\langle f'(t)\right\rangle = \frac{\sum_{i=1}^{n}\cos\theta_i f_i'(t)}{\sum_{i=1}^{n}\cos\theta_i}. \tag{38}$$

Averaged scores

To take the average of nonlinear scores, they should be transformed so the corresponding variables are Gaussian. The spatial average RMSE (considering the area factor) is:

$$\langle\mathrm{RMSE}\rangle = \sqrt{\frac{\sum_{i=1}^{n}\mathrm{MSE}_i\cos\theta_i}{\sum_{i=1}^{n}\cos\theta_i}}. \tag{39}$$

Similarly, the average MSSS is:

$$\langle\mathrm{MSSS}\rangle = 1 - \frac{\sum_{i=1}^{n}\mathrm{MSE}_i\cos\theta_i}{\sum_{i=1}^{n}\mathrm{MSE}_{\mathrm{C}i}\cos\theta_i}. \tag{40}$$

For the correlation coefficients, the Fisher Z-transform must be taken first. This is defined as:

$$Z(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \tanh^{-1} r \tag{41}$$

The spatial average TCC is the defined as:

$$\langle\mathrm{TCC}\rangle = Z^{-1}\left[\sqrt{\frac{\sum_{i=1}^{n}Z\left(\mathrm{TCC}_i\right)\cos\theta_i}{\sum_{i=1}^{n}\cos\theta_i}}\right] \tag{42}$$

and the temporal average ACC is

$$\langle\mathrm{ACC}\rangle = Z^{-1}\left\{\frac{1}{N}\sum_{t=1}^{N}Z[\mathrm{ACC}(t)]\right\}. \tag{43}$$

Orthogonality principle and MSSS decomposition

The MSSS (Eq. 35), can be expanded for leave-one-out cross-validated forecasts (Murphy 1988). Using Eqs. (31), (32), (34) and (36) in (35), we obtain:

$$\mathrm{MSSS}_i = \left\{ 2\frac{s_{fi}}{s_{xi}}\mathrm{TCC}_i - \left(\frac{s_{fi}}{s_{xi}}\right)^2 - \left(\frac{\left[\overline{f}_i - \overline{x}_i\right]}{s_{xi}}\right)^2 + \frac{2N-1}{(N-1)^2} \right\} \Big/ \left\{ 1 + \frac{2N-1}{(N-1)^2} \right\}. \tag{44}$$

This equation gives a relation between the MSSS and the TCC. For forecasts with the same variance as that of observations and no overall bias, the MSSS is only positive (MSE lower than for climatology) if the TCC is larger than approximately 0.5.

A more simplified relation can be obtained in our case for the prediction of the detrended anomalies (natural variability). As we mentioned in Appendix 1.4, the predictor (Eq. 24) is built in such a way that the coefficients satisfy the Yule Walker equations, which are derived from the orthogonality principle (Wold 1938; Brockwell and Davis 1991; Hipel and McLeod 1994; Palma 2007; Box et al. 2008). This principle states that the error of the optimal predictor, $e_i(t) = x_i(t) - f_i(t)$ (in a mean square error sense) is orthogonal to any possible estimator:

$$\langle e_i(t) f_i(t) \rangle = 0. \tag{45}$$

From this ensemble average condition, we get the analytical expressions for the coefficients as a function of the fluctuation exponent, H, for the fGn process. If the model realistically describes the actual temperature anomalies, then the condition Eq. (45) can be approximated by the temporal average in the POV:

$$\frac{1}{N}\sum_{t=i}^{N}\left[e_i(t) f_i(t)\right] = 0. \tag{46}$$

or

$$\frac{1}{N}\sum_{t=i}^{N}\left[x_i(t) - f_i(t)\right] f_i(t) = 0. \tag{47}$$

from which:

$$\frac{1}{N}\sum_{t=i}^{N}\left[x_i(t) f_i(t)\right] = \frac{1}{N}\sum_{t=i}^{N} f_i(t)^2. \tag{48}$$

For $\overline{x}_i = \overline{f}_i = 0$, dividing by the product $s_{x_i} s_{f_i}$ and using Eqs. (31) and (36), we can rewrite Eq. (48) as:

$$\mathrm{TCC}_i = \frac{s_{fi}}{s_{xi}}. \tag{49}$$

Using this ratio in Eq. (44) we finally obtain:

$$\mathrm{MSSS}_i = \frac{\mathrm{TCC}_i^2 + \frac{2N-1}{(N-1)^2}}{1 + \frac{2N-1}{(N-1)^2}}. \tag{50}$$

A more detailed analysis gives the same expression with the weaker condition of overall unbiased estimates $\overline{x}_i - f_i = 0$ (not necessarily each of them must be zero).

In our case, for the forecast of the detrended anomalies (natural variability) at monthly resolution in the POV 1951–2019 ($N = 828$ months), the N-dependent term in Eq. (50) is negligible:

$$\frac{2N-1}{(N-1)^2} \approx 0.0024, \tag{51}$$

so, with good approximation we obtain:

$$\mathrm{MSSS}_i \approx \mathrm{TCC}_i^2. \tag{52}$$

The orthogonality principle, Eq. (47) (or equivalently, Eq. (49) or Eq. (52)), is the condition that maximizes the MSSS. In our case, where the autoregressive coefficients in our predictor are analytical functions of only one parameter (H), if Eq. (52) is verified then our predictor is optimal in a mean square error sense and our model is suitable for describing the natural temperature variability.

Funding This study was funded by Hydro-Québec (Bourse de doctorat Hydro-Québec en science-F213013R02) and Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-2016-04796).

References

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

Box GEP, Jenkins GM, Reinsel GC (2008) Time series analysis. Wiley, New York

Brockwell PJ, Davis RA (1991) Time series: theory and methods. Springer, New York

Christensen HM, Moroz IM, Palmer TN (2015) Evaluation of ensemble forecast uncertainty using a new proper score: application to medium-range and seasonal forecasts. Q J R Meteorol Soc 141:538–549. https://doi.org/10.1002/qj.2375

Christensen HM, Berner J, Coleman DRB, Palmer TN (2017) Stochastic parameterization and El Niño-Southern Oscillation. J Clim 30:17–38. https://doi.org/10.1175/JCLI-D-16-0122.1

Clarke DC, Richardson M (2021) The benefits of continuous local regression for quantifying global warming. Earth Sp Sci, 8:e2020EA001082. https://doi.org/10.1029/2020EA001082

Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 83:596. https://doi.org/10.2307/2289282

Davini P, von Hardenberg J, Corti S et al (2017) Climate SPHINX: evaluating the impact of resolution and stochastic physics parameterisations in the EC-Earth global climate model. Geosci Model Dev 10:1383–1402. https://doi.org/10.5194/gmd-10-1383-2017

Delignières D (2015) Correlation properties of (discrete) fractional Gaussian noise and fractional brownian motion. Math Probl Eng 2015:1–7. https://doi.org/10.1155/2015/485623

Del Rio Amador L, Lovejoy S (2019) Predicting the global temperature with the Stochastic Seasonal to Interannual Prediction System (StocSIPS). Clim Dyn 53:4373–4411. https://doi.org/10.1007/s00382-019-04791-4

Del Rio Amador L, Lovejoy S (2021) Long-range forecasting as a past value problem: untangling correlations and causality with scaling. Geophys Res Lett Rev. https://doi.org/10.1002/essoar.10505160.1

Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10:507. https://doi.org/10.2307/2331838

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 (2015) Stochastic climate theory and modeling. Wiley Interdiscip Rev Clim Change 6:63–78. https://doi.org/10.1002/wcc.318

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

Gottwald GA, Crommelin DT, Franzke CLE (2017) Stochastic climate theory. In: Franzke CLE, Okane TJ (eds) Nonlinear and stochastic climate dynamics. Cambridge University Press, Cambridge, pp 209–240

Graham R, Yun W, Kim J et al (2011) Long-range forecasting and the Global Framework for Climate Services. Clim Res 47:47–55. https://doi.org/10.3354/cr00963

Gripenberg G, Norros I (1996) On the prediction of fractional Brownian motion. J Appl Probab 33:400–410. https://doi.org/10.1017/S0021900200099812

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 (2018) Regional climate sensitivity- and historical-based projections to 2100. Geophys Res Lett 45:4248–4254. https://doi.org/10.1002/2017GL076649

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. In: Hipel KW, Mcleod AI (eds) Time series modelling of water resources and environmental systems. Elsevier, Amsterdam, pp 1–1013

Jolliffe IT, Stephenson DB (2011) Forecast verification: a practitioner's guide in atmospheric science, 2nd edn. Wiley, Hoboken

Kalnay E, Kanamitsu M, Kistler R et al (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77:437–471. https://doi.org/10.1175/1520-0477(1996)077%3c0437:TNYRP%3e2.0.CO;2

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

Kim G, Ahn J, Kryjov VN et al (2020) Assessment of MME methods for seasonal prediction using WMO LC-LRFMME hindcast dataset. Int J Climatol. https://doi.org/10.1002/joc.6858

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

Kryjov VN, Kang H-W, Nohara D et al (2006) Assessment of the climate forecasts produced by individual models and MME methods. APCC Technical Report 2006, APEC Climate Center. Busan, South Korea

Leutbecher M, Palmer TN (2008) Ensemble forecasting. J Comput Phys 227:3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014

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 (2018) Spectra, intermittency, and extremes of weather, macroweather and climate. Sci Rep 8:12697. https://doi.org/10.1038/s41598-018-30829-4

Lovejoy S (2019) Fractional relaxation noises, motions and the fractional energy balance equation. Nonlin Process Geophys Discuss. https://doi.org/10.5194/npg-2019-39

Lovejoy S (2021a) The half-order energy balance equation, Part 1: the homogeneous HEBE and long memories. Earth Syst Dynam. https://doi.org/10.5194/esd-2020-12

Lovejoy S (2021b) The half-order energy balance equation, Part 2: the inhomogeneous HEBE and 2D energy balance models. Earth Syst Dynam Discuss. https://doi.org/10.5194/esd-2020-13

Lovejoy S, Schertzer D (1986) Scale invariance in climatological temperatures and the spectral plateau. Ann Geophys 4B:401–410

Lovejoy S, Schertzer D (2010) Towards a new synthesis for atmospheric dynamics: space–time cascades. Atmos Res 96:1–52. https://doi.org/10.1016/j.atmosres.2010.01.004

Lovejoy S, Schertzer D (2012a) Haar wavelets, fluctuations and structure functions: convenient choices for geophysics. Nonlinear Process Geophys 19:513–527. https://doi.org/10.5194/npg-19-513-2012

Lovejoy S, Schertzer D (2012b) Low-Frequency Weather and the Emergence of the Climate. In: Sharma AS, Bunde A, Dimri VP and Baker DN (eds) Extreme Events and Natural Hazards: The Complexity Perspective. https://doi.org/10.1029/2011GM001087

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

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, Procyk R, Hébert R, Del Rio Amador L (2021) The fractional energy balance equation. Q J R Meteorol Soc. https://doi.org/10.1002/qj.4005

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

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

Mori H (1965) Transport, collective motion, and Brownian motion. Prog Theor Phys 33:423–455. https://doi.org/10.1143/PTP.33.423

Murphy AH (1988) Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon Weather Rev 116:2417–2424. https://doi.org/10.1175/1520-0493(1988)116%3c2417:SSBOTM%3e2.0.CO;2

Ncep/ncar (2020) Ncep/ncar reanalysis 1. https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html. Accessed 3 Jan 2020

Newman M (2013) An empirical benchmark for decadal forecasts of global surface temperature anomalies. J Clim 26:5260–5269. https://doi.org/10.1175/JCLI-D-12-00590.1

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

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

Palmer TN (2019) Stochastic weather and climate models. Nat Rev Phys 1:463–471. https://doi.org/10.1038/s42254-019-0062-2

Palmer T, Buizza R, Hagedorn R et al (2006) Ensemble prediction: a pedagogical perspective. ECMWF Newsl 106:10–17. https://doi.org/10.21957/ab129056ew

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

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

Procyk R, Lovejoy S, Hébert R (2020) The fractional energy balance equation for climate projections through 2100. Earth Syst Dynam Discuss. https://doi.org/10.5194/esd-2020-48

Rackow T, Juricke S (2020) Flow-dependent stochastic coupling for climate models with high ocean-to-atmosphere resolution ratio. Q J R Meteorol Soc 146:284–300. https://doi.org/10.1002/qj.3674

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

Trenberth KE (1997) The definition of El Niño. Bull Am Meteorol Soc 78:2771–2777. https://doi.org/10.1175/1520-0477(1997)078%3c2771:TDOENO%3e2.0.CO;2

Varotsos CA, Efstathiou MN, Cracknell AP (2013) On the scaling effect in global surface air temperature anomalies. Atmos Chem Phys 13:5243–5253. https://doi.org/10.5194/acp-13-5243-2013

Williams PD (2012) Climatic impacts of stochastic fluctuations in air–sea fluxes. Geophys Res Lett. https://doi.org/10.1029/2012GL051813

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

WMO (2010a) Standardised verification system (SVS) for long-range forecasts (LRF). New attachment II-8 to the manual on the GDPS. WMO-No. 485, vol 1. Geneva, Switzerland.

WMO (2010b) Manual on the Global Data-processing and Forecasting System Volume I. (WMO-No. 485). Geneva, Switzerland.

Wold H (1938) A study in analysis of stationary time series. J R Stat Soc, Almqvist und Wiksell, Uppsala

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal 01:1–41. https://doi.org/10.1142/S1793536909000047

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

Zampieri L, Goessling HF, Jung T (2018) Bright prospects for arctic sea ice prediction on subseasonal time scales. Geophys Res Lett 45:9731–9738. https://doi.org/10.1029/2018GL079394

Zeiler A, Faltermeier R, Keck IR et al (2010) Empirical mode decomposition—an introduction. In: The 2010 international joint conference on neural networks (IJCNN). IEEE, pp 1–8

Zwanzig R (1973) Nonlinear generalized Langevin equations. J Stat Phys 9:215–220. https://doi.org/10.1007/BF01008729

Zwanzig R (2001) Nonequilibrium statistical mechanics, 1st edn. Oxford University Press, Oxford