Long-term variability in global sea surface temperature (SST) is often quantified by the slope from a linear regression fit. Attention is then focused on assessing the statistical significance of the derived slope parameter, but the adequacy of the linear model itself, and the inherent assumption of a deterministic linear trend, is seldom tested. Here, a parametric statistical test is applied to test the hypothesis of a linear deterministic trend in global sea surface temperature. The results show that a linear slope is not adequate for describing the long-term variability of sea surface temperature over most of the earth’s surface. This does not mean that sea surface temperature is not increasing, rather that the increase should not be characterized by the slope from a linear fit. Therefore, describing the long-term variability of sea surface temperature by implicitly assuming a deterministic linear trend can give misleading results, particularly in terms of uncertainty, since the actual increase could be considerably larger than the one predicted by a deterministic linear model.

1. Introduction

The identification and estimation of trends is a frequent and fundamental task in most geosciences disciplines dealing with time series of observational or model data. However, this endeavor is hindered by the lack of a precise, objective definition of what a trend is exactly. Implicit in the intuitive notion of trend are concepts such as long-term, smoothness, and monotonicity, but it is not possible to define unambiguously how long is “long term,” or how smooth a pattern should be in order to qualify as a trend.

Even time series generated by purely random processes can exhibit visually appealing trends. For example, stationary zero-mean Gaussian red noise time series exhibit long runs without a zero crossing that can be misleadingly taken as evidence of nonstationary behavior (Wunsch 1999). Since red noise signals are ubiquitous in climate, as the result of the integration by the ocean of white noise atmospheric forcings with stationary characteristics (Hasselmann 1976), it is particularly important in a climate context to discriminate between nonstationary, deterministic trends and red noise, stationary variability.

To complicate matters further, time series characterized by long-range dependence (also called long memory or long-term persistence; i.e., with autocorrelations decaying hyperbolically) exhibit “trendlike” features (e.g., Beran 1994) that can be mistakenly interpreted as a trend (in the conventional sense), leading to erroneous forecasts and interpretations of the variability structure of a time series, particularly in terms of statistical uncertainty (Koutsoyiannis 2006; Koutsoyiannis and Montanari 2007). In this context, trend assessment has received considerable attention, in particular in the case of hydro-climatic time series (e.g., Kallache et al. 2005; Cohn and Lins 2005; Rybski et al. 2006; Barbosa et al. 2008; Lennartz and Bunde 2009; Fatichi et al. 2009; Rybski and Bunde 2009). These studies focused mainly on trend identification in the presence of long-range dependence. Still, though a long-range-dependent process has a precise, rigorous definition, a trend has not. And despite the growing amount of literature dealing with trend evaluation and long-range dependence issues, in practice most trends in hydro-climatic time series are reported as such without a careful consideration of alternative approaches. This is hardly surprising, considering that the lack of a definition for trend renders trend assessment a delicate exercise in itself.

A way out is to go from the ill-defined concept of trend to the well-defined notion of stationarity, and to constrain trend assessment to the evaluation of a linear
regression model (i.e., to perform a parametric test). Although this restriction is theoretically limiting, in practice the overwhelming majority of trends in hydroclimatic records are reported as the slope from a linear regression model. Thus, a pressing issue in geosciences practice is to assess when a linear regression model is an appropriate description for a time series. One could think that if a derived slope is statistically significant, particularly if inference is performed carefully, for example by taking into account the data’s autocorrelation (e.g., Fomby and Vogelsang 2002) or via simulation by means of bootstrap techniques, then the linear regression model would be appropriate. However, stochastic features, such as long-range dependence, can produce statistically significant linear trends. Therefore, the plausibility of the linear regression model needs to be tested itself, in addition to testing if the trend slope is statistically significant.

Although a rigorous definition of trend does not exist, stationarity is a well-defined statistical concept. A time series \( X_t \) is (weakly) stationary if its first and second moments are time invariant (i.e., the mean and variance are constant and the covariance depends on the time lag between observations but not on the time itself).

Red noise time series described by a first-order autoregressive process \( X_t = c + \phi X_{t-1} + \varepsilon_t \) with \( 0 < \phi < 1 \), \( c = \text{constant} \) and \( \varepsilon_t \sim \text{iid}(0, \sigma^2) \), verify the above definition of stationarity. The mean and variance are asymptotically time invariant and the process is mean reverting (i.e., it will cross the mean line a infinite number of times, the mean acting as an attractor around which the process fluctuates). However, for \( \phi = 1 \), the process \( X_t = c + \phi X_{t-1} + \varepsilon_t \) describes a random walk and is no longer stationary, since the variance of the process is \( t \sigma^2 \) (i.e., not constant in time; e.g., Hamilton 1994). This process is called difference stationary (or integrated of order 1) because it becomes stationary after being differentiated once. It is no longer a mean reverting process, rather it is a process that tends to move away from the mean.

A distinct kind of nonstationary behavior, named trend stationarity, is associated with deterministic trends. Trend-stationary time series are described by a process \( X_t = a + b_t + \varepsilon_t \), and detrending renders the time series stationary (while for a difference stationary series, detrending is not enough to render the time series stationary). The mean of a trend-stationary process is time varying, but the variance is constant in time.

A trend-stationary process is usually implicitly assumed when fitting a linear regression model to a time series of observations. This assumption can and should be tested and this is the subject of this work. Both trend-stationary and difference-stationary time series exhibit a tendency behavior, but trend-stationary series are characterized by a deterministic trend tendency with stable variance, while difference-stationary series are characterized by a stochastic tendency with increasing variance. The distinction between the two kinds of nonstationary behavior has practical implications (e.g., forecasting) and for understanding the underlying generating process (e.g., whether the time series can be characterized by a deterministic trend or whether the apparent trend is due to stochastic fluctuations). Woodward and Gray (1995) developed a bootstrap-based classification procedure aiming to distinguish between difference and trend-stationary series while Zheng and Basher (1999) considered several structural models for (already) differenced temperature time series with selection of the most appropriate model on the basis of an information criterion. In this work, well-defined parametric statistical tests are applied in order to evaluate the trend-stationary assumption in global sea surface temperature (SST).

2. Methods

Parametric statistical tests have been developed in econometrics for discriminating between difference-stationary and trend-stationary time series. The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test (Kwiatkowski et al. 1992) tests the null hypothesis of a trend-stationary process against a difference-stationary alternative. Rejection of the null hypothesis when applying the test to a time series indicates that its long-term variability should not be characterized by the slope of a linear regression model (even if it is statistically significant), since the assumption of a deterministic trend is not itself plausible.

The KPSS test is based on the model \( X_t = \beta t + r_t + \nu_t \), where \( r_t \) is a random walk,

\[ r_t = r_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2_r) \],

and \( \nu_t \) is a stationary noise process. It allows to test for stationarity (in the form of a constant level) and trend stationarity (in the form of a deterministic trend plus a stationary stochastic noise). In the constant level case, the null hypothesis is \( H_0 : \beta = 0 \land \sigma^2_r = 0 \) against the alternative \( H_1 : \sigma^2_r = 0 \neq 0 \). In the general trend-stationary case, the null hypothesis is \( H_0 : \beta = 0 \land \sigma^2_r = 0 \) against \( H_1 : \sigma^2_r = 0 \neq 0 \).

While the application of the KPSS test is sufficient for assessing trend stationarity, and specifically if a linear deterministic trend is a reasonable model for a given time series, further insight can be obtained by applying complementary tests that take as the null hypothesis
difference stationarity instead of trend stationarity. The Phillips–Perron (PP) test (Phillips and Perron 1988) has been designed to test the null hypothesis of a difference-stationary random walk process against a trend-stationary alternative. The PP test is based on the model \( X_t = \eta + \beta_t + \pi X_{t-1} + \psi_t \), where \( \psi_t \) is a stationary noise process and \( \eta \) and \( \beta \) are the parameters of a first-order polynomial regression. The difference-stationary null hypothesis is expressed by \( H_0: \pi = 1 \) against \( H_1: \pi < 1 \). The advantage of this test compared to previous difference-stationary tests, such as the Dickey–Fuller test and its variants (Dickey and Fuller 1979; Said and Dickey 1984), is that the PP test (as the KPSS test) does not assume the stationary noise process to be white noise, allowing for serial correlation and heteroskedasticity (handled directly in the test statistic).

The joint application of the KPSS and PP tests yields additional information than when applying a single test, because of the inherent design of statistical tests, which privileges rejection of the null as the “strong” conclusion. If the null hypothesis of the KPSS test is rejected, but the null of the PP test is not rejected, one can infer that a time series is difference stationary. Conversely, if the trend-stationary null of the KPSS test is not rejected, and the null of the PP test is rejected, one can infer that a time series is trend stationary. Two additional possibilities remain. If both tests reject the respective null hypothesis, alternative behaviors (such as long-range dependence) need to be considered. Finally, if both tests fail to reject the respective null hypothesis, then the tests and/or the time series are not informative enough for discriminating the kind of stationary behavior.

In the present study the KPSS and PP tests are applied to the SST time series using the implementation in R-package t-series (Trapletti and Hornik 2009).

3. Data

Sea surface temperature observations are taken from the Extended Reconstructed Sea Surface Temperature dataset version 3b (ERSST.v3b) of Smith et al. (2008). This is an improved version of the dataset of Smith and Reynolds (2005) resulting from improved low-frequency (LF) tuning that reduces the SST anomaly damping before 1930. This version of the dataset does not include satellite measurements in order to avoid a cold bias in the radiometer SST retrievals. Monthly SST values on a 2° regular grid over the global ocean are extracted from the ERSST.v3b dataset for the period from January 1880 to December 2009. Reconstructed data are available since 1854 but the signal is damped before 1880 because of very sparse data in the early years. The mean seasonal cycle, computed from the mean monthly averages over the complete 130-yr period, is removed from each time series. Only these (de-seasoned) time series of SST anomalies are considered hereafter. The global mean SST time series (Fig. 1) is obtained by spatially averaging over the global ocean the values for each month.

4. Results

Trend assessment is carried out first for the global mean time series of SST anomalies (section 4a) and then performed on the individual time series of monthly SST anomalies at each point of the 2° global grid (section 4b).

a. Global mean SST

The global mean SST time series shown in Fig. 1 displays an apparent increasing trend. The fit of a linear regression model by ordinary least squares gives a positive slope of 0.046°C decade\(^{-1}\) (with a standard error of 8.90 \times 10^{-4} \text{ °C decade}^{-1} ), which is statistically significant for a 95% confidence level (\( p \) value < 2 \times 10^{-16} ). However, it is well known that the assumption of independent errors implicit in ordinary least squares regression is not realistic and that it can significantly affect the significance of the derived trends (e.g., Lee and Lund 2004; Barbosa et al. 2008). Serial correlation is addressed here using two different approaches: generalized least squares (e.g., Hamilton 1994) and maximum entropy bootstrap (e.g., Vinod and de Lacalle 2009).

Generalized least squares is a generalization of the ordinary linear regression model for the case of a non-diagonal covariance matrix (i.e., autocorrelated errors). The method is based on the maximization of the likelihood given the autocorrelation in the data. Here, a first-order autoregression has been assumed for the structure of the covariance matrix, yielding a trend estimate of 0.045°C decade\(^{-1}\) (with a standard error of 0.0078°C decade\(^{-1}\) ). Ordinary and generalized least squares slopes are very similar, differing only in the magnitude of the corresponding standard errors. Generalized least squares gives a more realistic error estimate since it takes into account the positive autocorrelation in the series, but
although increasing considerably the estimation error, the resulting slope is still statistically significant for a 95% confidence level.

Bootstrapping is a nonparametric alternative for assessing the uncertainty of a linear trend in the presence of autocorrelation. Maximum entropy bootstrap allows us to construct a set of replicates of the original series to be used for inference while retaining the temporal dependence structure of the original series in the resampling process. Maximum entropy bootstrap as implemented in the R language (Vinod and de Lacalle 2009) is applied to the global mean SST series. The resulting 10,000 replicates yield a 95% confidence interval for the slope of \( [0.045, 0.048] \) °C decade\(^{-1}\).

All the three different approaches used to compute and assess the uncertainty of a deterministic linear trend in global mean SST yield a statistically significant positive slope, suggesting an increasing linear trend. However, when testing the hypothesis of whether this assumption of a deterministic linear trend is reasonable (i.e., assessing the assumed model itself rather than its parameter’s estimates, a different picture emerges). The KPSS test gives a \( p \) value < 0.01, indicating rejection of the null hypothesis of a deterministic linear trend. The PP test also gives a \( p \) value < 0.01, indicating rejection of the null hypothesis of a difference stationary series. The joint results of the KPSS and PP tests suggest the need to consider alternative forms of temporal dependence, such as long-range dependence, since neither difference-stationary nor trend-stationary behaviors seem to be appropriate for the global mean SST time series. This is supported by the form of the temporal structure of the detrended global mean SST series (Fig. 2). The autocorrelation function still exhibits nonzero values up to lags of 5 yr, even after subtraction of a deterministic linear trend, indicating temporal persistence.

### b. Individual SST series

To obtain a spatial perspective, linear trends are computed for the individual series of monthly SST anomalies by ordinary and generalized least squares. The resulting slopes (which are very similar for ordinary and generalized least squares) are displayed in Fig. 3. Most trends are positive, indicating a general increase in sea surface temperature for the period 1880–2009. Negative slopes are found in the northern part of the North Atlantic Ocean, south of Greenland. The standard errors (s.e.) corresponding to these trend estimates are shown in Figs. 4 and 5 for the ordinary and generalized least squares case, respectively. As expected, the standard errors from generalized least squares that take into account the autocorrelation in the data are considerably higher than the ordinary least squares errors. The spatial pattern of the formal standard errors shows higher uncertainty in the equatorial Pacific, characterized by a marked ENSO-like pattern. As a measure of the significance of the estimated trends, the \( p \) values corresponding to the generalized least squares estimates are shown in Fig. 6. The \( p \) values lower than 0.05 indicate statistically significant trends at the 95% confidence level. Figure 6 shows that even considering the more realistic standard errors computed by generalized least squares, trends in sea surface temperature are statistically significant at most locations except in the equatorial Pacific, where the derived trends are not statistically significant at a 95% confidence level. Stationarity tests allow to test whether a deterministic linear model is a reasonable assumption to
describe long-term variability in SST time series. The PP test gives a p value < 0.01 for all locations, indicating rejection of the null hypothesis of difference stationary behavior. The results of the KPSS test are shown in Fig. 7. For most locations the p value resulting from the KPSS test is below 0.05, indicating rejection of the null hypothesis of a deterministic linear trend. For some regions, including the Arabian Sea and the eastern equatorial Pacific, the p values indicate that the null hypothesis of a deterministic linear trend cannot be rejected.

5. Conclusions

The characterization of long-term variability in sea surface temperature is often based on linear slopes derived by fitting a linear deterministic model to the series of in situ and/or satellite observations (e.g., Casey and Cornillon 2001; Andersen et al. 2002; Barbosa and Andersen 2009). Although attention is increasingly being paid on how to assess statistical significance, particularly taking into account the autocorrelation in the data (e.g., Lee and Lund 2004), the adequacy of the linear model itself is seldom considered. In this work, a well-established statistical test, the KPSS parametric test, is applied to assess whether SST time series can be described by a deterministic linear trend. Alternatives to deterministic linear processes that could generate trend-like features include long-range-dependent processes, but although there is some evidence of scaling behavior in SST (Fraedrich and Blender 2003; Fraedrich et al. 2004), long-range dependence is a more subtle concept that is not addressed in this work. Instead, the present study focuses on whether a deterministic trend can be assumed, specifically on whether the null hypothesis of a linear deterministic trend is rejected (the strong conclusion in statistical testing).

The results for global mean SST indicate that although the series can be characterized by a statistically significant linear slope, the assumption of a linear deterministic trend is not justified. A spatially consistent behavior is found when analyzing individual time series, with the KPSS test indicating for most locations, similarly to the global mean, rejection of a deterministic linear trend. The equatorial Pacific appears in terms of long-term variability as a well-marked ENSO pattern, with the largest errors associated to linear slopes. Furthermore, the KPSS test results do not preclude a deterministic linear trend in some well-delimited regions in the equatorial Pacific and Arabian Sea. In the equatorial Pacific, oscillations associated with the ENSO phenomenon (Ghil and Vautard 1991) are dominant over interdecadal variability and the test is not able to reject the trend-stationary hypothesis. For the global

Fig. 4. Standard errors of slopes estimated by ordinary least squares: (top) spatial map and (bottom) histogram.

Fig. 5. As in Fig. 4, but estimated by generalized least squares.
mean SST as well as for all the time series at individual locations, long-term variability is not characterized by difference stationary behavior. Thus, although the integration of stationary white noise processes could lead to variability in the form of random walks, this kind of stochastic nonstationary behavior is rejected by the parametric statistical tests.

The main conclusion of this work is that a linear deterministic trend is not adequate for describing the long-term variability of sea surface temperature over most of the earth’s surface. The rejection of trend stationary behavior does not mean that global mean SST is not increasing, rather that the increase is not well characterized by a deterministic linear trend. This has implications for forecasting and understanding the temporal evolution of global sea surface temperature. Characterizing long-term variability by a deterministic linear trend can then give misleading results in terms of uncertainty, for example, the actual increase could be considerably larger than the one predicted by a deterministic linear model.

It is well known that a straight line is often not adequate to describe trends in climate data, particularly temperature (e.g., Miranda and Tomé 2009; Thompson et al. 2010). This work shows how to test and quantitatively assess whether a straight line is reasonable and if yes demonstrates the importance of estimating (e.g., by generalized least squares) realistic error bars for the derived slope.

Acknowledgments. This work is supported by FCT (contract under Programme Ciencia 2008).

REFERENCES


