Hypothesis Testing for Autocorrelated Short Climate Time Series*

VIRGINIE GUEMAS

Institut Català de Ciencies del Clima, Barcelona, Spain, and Centre National de Recherches Météorologiques/ Groupe d'Etude de l'Atmosphère Météorologique, Météo-France, CNRS, UMR 3589, Toulouse, France

LUDOVIC AUGER

Centre National de Recherches Météorologiques/Groupe d'Etude de l'Atmosphère Météorologique, Météo-France, CNRS, UMR 3589, Toulouse, France

FRANCISCO J. DOBLAS-REYES

Institut Català de Ciencies del Clima, and Instituciò Catalana de Recerca i Estudis Avançats, Barcelona, Spain

(Manuscript received 14 February 2013, in final form 17 September 2013)

ABSTRACT

Commonly used statistical tests of hypothesis, also termed inferential tests, that are available to meteorologists and climatologists all require independent data in the time series to which they are applied. However, most of the time series that are usually handled are actually serially dependent. A common approach to handle such a serial dependence is to replace in those statistical tests the actual number of data by an estimated effective number of independent data that is computed from a classical and widely used formula that relies on the autocorrelation function. Despite being perfectly demonstrable under some hypotheses, this formula provides unreliable results on practical cases, for two different reasons. First, the formula has to be applied using the estimated autocorrelation function, which bears a large uncertainty because of the usual shortness of the available time series. After the impact of this uncertainty is illustrated, some recommendations of preliminary treatment of the time series prior to any application of this formula are made. Second, the derivation of this formula is done under the hypothesis of identically distributed data, which is often not valid in real climate or meteorological problems. It is shown how this issue is due to real physical processes that induce temporal coherence, and an illustration is given of how not respecting the hypotheses affects the results provided by the formula.

1. Introduction

Time series of climatological and meteorological variables are frequently serially dependent, reflecting the strong persistence of the meteorological and climatic phenomena. For example, blocking conditions can be associated with similar weather conditions for more than 1 week, and a given phase of the Atlantic multidecadal oscillation (Knight et al. 2005) is associated with similar surface conditions in the Atlantic Ocean for a decade or more. This condition stands as an obstacle to the use of statistical tests of hypothesis or to the estimation of confidence intervals for which a required property of the time series is the independence of the data (von Storch and Zwiers 2001). This problem was highlighted early on by Bayley and Hammersley (1946), Leith (1973), and Jones (1975), who suggested various definitions of a persistence time T_p, which corresponds to the time between effectively independent data. The effective number of independent data in a time series is then given by

$$N_{\mathrm{eq}} = N\Delta t/T_p, \qquad (1)$$

where N is the actual sample size—that is, the actual number of data in the time series—and Δt is the sampling

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JAMC-D-13-064.s1.

Corresponding author address: V. Guemas, Institut Català de Ciencies del Clima, Carrer del Doctor Trueta, 203, 08005 Barcelona, Spain.
E-mail: vguemas@ic3.cat

DOI: 10.1175/JAMC-D-13-064.1

interval. The actual number of data N can then be replaced in any statistical test by the estimated effective number of independent data N_{eq}. This formulation is valid only if the persistence time T_p is larger than the sampling interval Δt; otherwise, the N data can be considered as independent and $N_{\mathrm{eq}} = N$. The problem of estimating confidence intervals for serially dependent data therefore reduces to a problem of estimating the persistence time T_p given the formulation in Eq. (1).

In the research fields of weather forecasting and climate prediction, the benefits of a given physical parameterization or a given data assimilation technique on the forecast quality are assessed by comparing the prediction scores of two retrospective sets of forecasts with and without this new feature. Those prediction scores can be, for example, the anomaly correlation coefficient or the root-mean-square error, which provide information on the accuracy of an ensemble-mean prediction, that is, how far from the observations the ensemble-mean prediction is on average. Those prediction scores can also be, for example, a measure of the spread of the members around the ensemble mean of a deterministic forecast that provides information on the reliability of the probabilistic forecast that can be derived from this deterministic forecast (Palmer et al. 2005). Any robust comparison of two scores requires their confidence intervals. The wide variety of scores used to assess the forecast quality leads currently to the use of a wide variety of parametric tests relying on the χ^2, Fisher, or Student distributions, which require an estimated N_{eq}, and the use of nonparametric block bootstrap tests, which require an estimated T_p for the serial dependence of the data to be accounted for. An alternative approach commonly used in weather forecasting is to consider the data to be independent from one day to the next, although the same meteorological conditions often persist for a few days (Wilks and Wilby 1999).

A few decades ago, the detection of climate change from past observational records was in its early stages. Estimating the significance level of differences in means between two observational periods taking into account the serial dependence of the data lay at the heart of this scientific question of detection. To estimate the effective number of independent data for tests of the mean, Anderson (1971) and Trenberth (1984) suggested the use of the autocorrelation function to estimate the time scales of persistence and developed the following formula:

$$N_{\mathrm{eq}} = \frac{N}{1 + 2\sum_{\tau=1}^{N-1} \frac{N-\tau}{N}\rho(\tau)}, \qquad (2)$$

where $\rho(\tau)$ is the autocorrelation function as a function of the lag τ. In a similar way, positive autocorrelations model a process that has a persistence time T_p larger than the sampling interval Δt while negative autocorrelations, though rare, would correspond to cyclical phenomena for which half of a period is equal to the sampling interval Δt, in which case we should consider the N data to be independent. Estimating the serial dependence through the autocorrelation function relies on the assumption of linearity of the dependencies, which stands as a common assumption. The derivation of Eq. (2), together with its underlying hypotheses, is provided as a document in the supplementary material. This method to deal with the serial dependence of meteorological or climatic time series is now widely taught in statistic courses and can be found, for example, in the book by von Storch and Zwiers (2001), which teaches statistics applied to climate sciences. The volatility of such a formula was highlighted early on by Thiébaux and Zwiers (1984). Zwiers and von Storch (1995) later proposed a more robust alternative method for the particular case of tests of the mean that is now widely adopted. Because of the lack of alternative methods for other types of inferential tests and confidence interval estimation such as the ones used in climate and weather predictions, Eq. (2) is still widely used in its original form in those research fields.

The objectives of this paper are the following:

1) We will illustrate the substantial uncertainty on the N_{eq} provided by the standard Eq. (2) from von Storch and Zwiers (2001), even when the hypotheses for its use are valid, which is an issue mentioned earlier by Thiébaux and Zwiers (1984), although this formula remains widely used currently, in particular when estimating confidence intervals for skill scores in weather and climate forecasting. Section 2 focuses on this objective.
2) We will suggest a preliminary treatment of the time series before applying such a formula that allows for more robust estimates of the effective sample size. Some tools are made available in the supplementary materials to apply this alternative approach, which aims at increasing the robustness of the N_{eq} estimation. Such a method is extensively described in section 3.
3) We will highlight a second issue never before reported in the literature that is raised by the application of such a formula to problems of climate and weather prediction because of the frequent violation of one of its hypotheses—namely, the hypothesis of "identically distributed data." We explain this issue in section 4.
4) We will suggest additional preliminary treatments that could prevent the consequences of violating such

a hypothesis. Those preliminary treatments are briefly described in section 4 and are validated in the supplementary material, and they are available as options to the tools that are contained in the supplemental material.

Sections 5 and 6 respectively provide a discussion and the conclusions. An appendix describes succinctly the tools provided in the supplementary material to apply the methods suggested in this article.

2. A large uncertainty on the estimated effective number of independent data with the standard approach

When applied to practical examples, the true autocorrelation function $\rho(\tau)$ is unavailable and $\rho(\tau)$ has to be replaced by its estimate $\hat{\rho}(\tau)$:

$$N_{\mathrm{eq}} = \frac{N}{1 + 2\sum_{\tau=1}^{N-1} \frac{N-\tau}{N}\hat{\rho}(\tau)}. \qquad (3)$$

Replacement of $\rho(\tau)$ by its estimate $\hat{\rho}(\tau)$ introduces a large uncertainty: in some cases, negative estimated autocorrelations can lead to an effective number of independent data N_{eq} that is larger than the actual number of data N. Our first recommendation is bounding the result of the formula in the following way:

$$N_{\mathrm{eq}} = \min\left[\frac{N}{1 + 2\sum_{\tau=1}^{N-1} \frac{N-\tau}{N}\hat{\rho}(\tau)}, N\right]. \qquad (4)$$

Even when bounding the result to N, the effective number of independent data N_{eq} provided by this formula still bears a large uncertainty, as illustrated with the histograms shown in Fig. 1. Those histograms have been built by drawing 1000 stationary first-order autoregressive processes (AR1) with $N = 50$ and by computing the effective number of independent data N_{eq} by applying Eq. (4) to each one of those stationary AR1, defined as (von Storch and Zwiers 2001)

$$\phi(t) = \frac{\alpha_0}{1-\alpha_1} \quad \text{if} \quad t = 1$$
$$\phi(t) = \alpha_0 + \alpha_1\phi(t-1) + \varepsilon(t) \quad \text{if} \quad t > 1\,, \qquad (5)$$

where t is the time; α_0, set to 0 (without loss of generality since autocorrelations are independent on the mean of the process), and α_1, which takes different values from 0.1 to 0.8 in the various panels of Fig. 1, are the two parameters defining an AR1; and $\varepsilon(t)$ is white noise following the normal distribution with mean 0 and standard deviation $(1-\alpha_1^2)^{1/2}$. To make sure that our AR1 process is stationary, we draw 100 data and we keep only the last 50 ones. A stationary first-order autoregressive process has been selected as a benchmark to model any meteorological and climatic phenomena because it is the simplest known statistical model that allows for persistence of the information. Furthermore, AR1 processes mimic the dynamics of discretized physical processes that follow a first-order linear differential equation and are therefore commonly used to represent meteorological and climatic phenomena (von Storch and Zwiers 2001). The case of nonstationary processes will be dealt with later in this article. The true autocorrelation function of those AR1 is given by (von Storch and Zwiers 2001)

$$\rho(\tau) = \alpha_1^{|\tau|}. \qquad (6)$$

Using this true autocorrelation function instead of the estimated one, Eq. (4) provides a true N_{eq} equal to 41.1, 33.6, 27.3, 21.8, 17.1, 13.0, 9.3, and 6.1 for an AR1 with α_1 set respectively to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. The true N_{eq} is thus expected to decrease when α_1 increases. The distribution of the estimated N_{eq} from the 1000 AR1 we have drawn should be approximately centered on the true N_{eq}. Figure 1 shows that the estimator of N_{eq} provided by Eq. (4) is heavily positively biased, and the maximum of occurrence in the distribution of N_{eq} estimates is centered at $N_{\mathrm{eq}} = 50$, which corresponds to the actual sample size, the actual number of data in the sample drawn. Even with $\alpha_1 = 0.8$, $N_{\mathrm{eq}} = 50$ in about 25% of the cases and the other obtained N_{eq} are spread between 8 and 50 without any particular peak, whereas we should obtain $N_{\mathrm{eq}} = 6.1$ for this particular case. From Fig. 1, we conclude that the N_{eq} provided by Eq. (4) when applied to practical cases is not reliable since it bears far too much uncertainty, which is consistent with previous findings from Thiébaux and Zwiers (1984).

The large spread of the N_{eq} provided by Eq. (4) and the unrealistic results in a number of cases originate from the large uncertainty in the estimation of the autocorrelation function from the short time series that are usually available. This uncertainty is illustrated with Fig. 2, which shows the true autocorrelation function of an AR1 with $\alpha_1 = 0.3$ or $\alpha_1 = 0.8$ as a function of the lag as black lines together with the interval between various quantiles of the 1000 estimated autocorrelation functions of AR1 drawn randomly, as explained in the previous paragraph, as colored lines. The range of estimated autocorrelations increases quickly with the lag. For lags

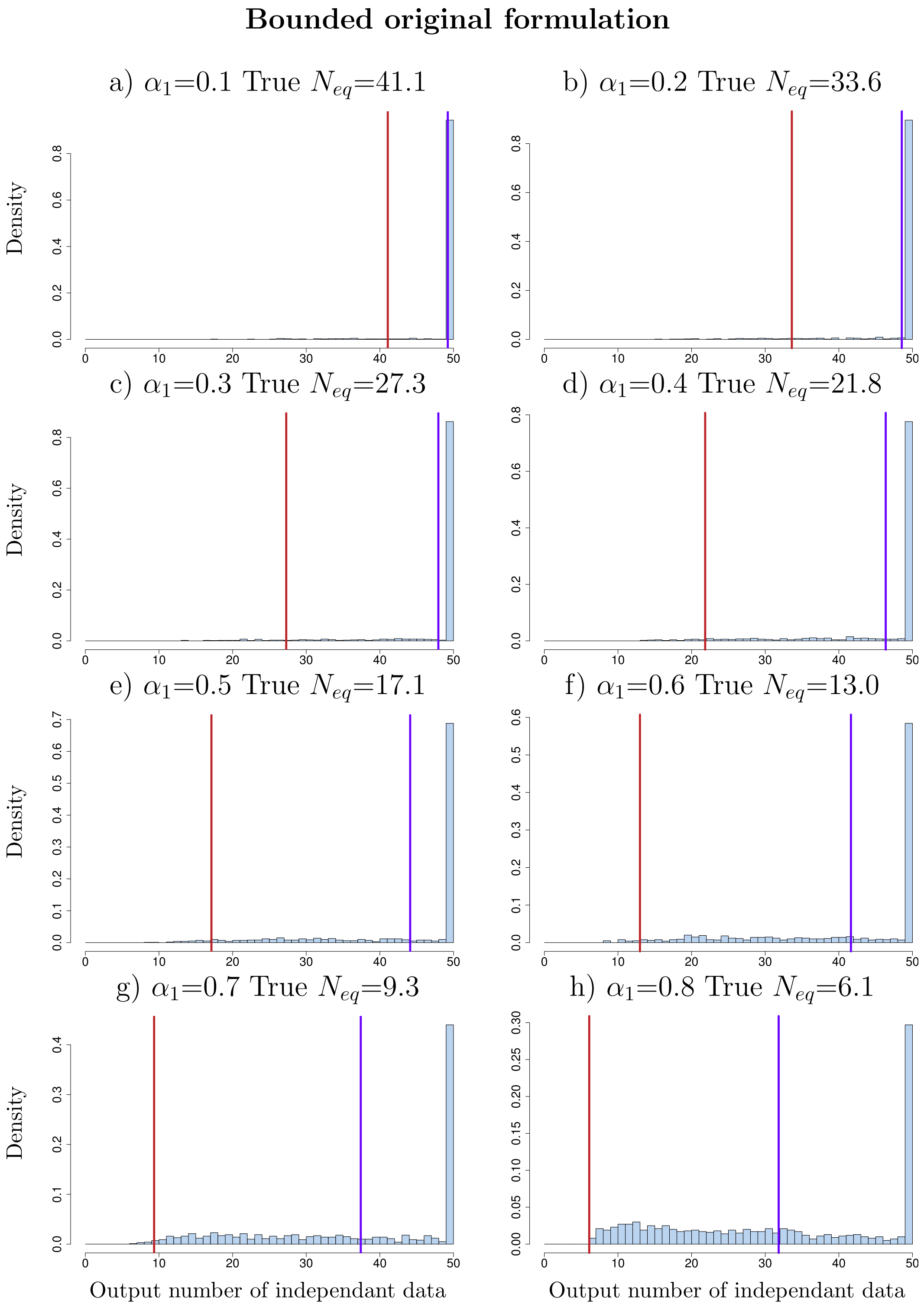


FIG. 1. Histograms of the number of effective independent data provided by Eq. (4) when applied to 1000 drawings of AR1 with α_1 set to (a) 0.1, (b) 0.2, (c) 0.3, (d) 0.4, (e) 0.5, (f) 0.6, (g) 0.7, and (h) 0.8. The mean of the distribution is provided as a vertical blue line. The true N_{eq} is provided in the title for each α_1 and as a vertical brown line.

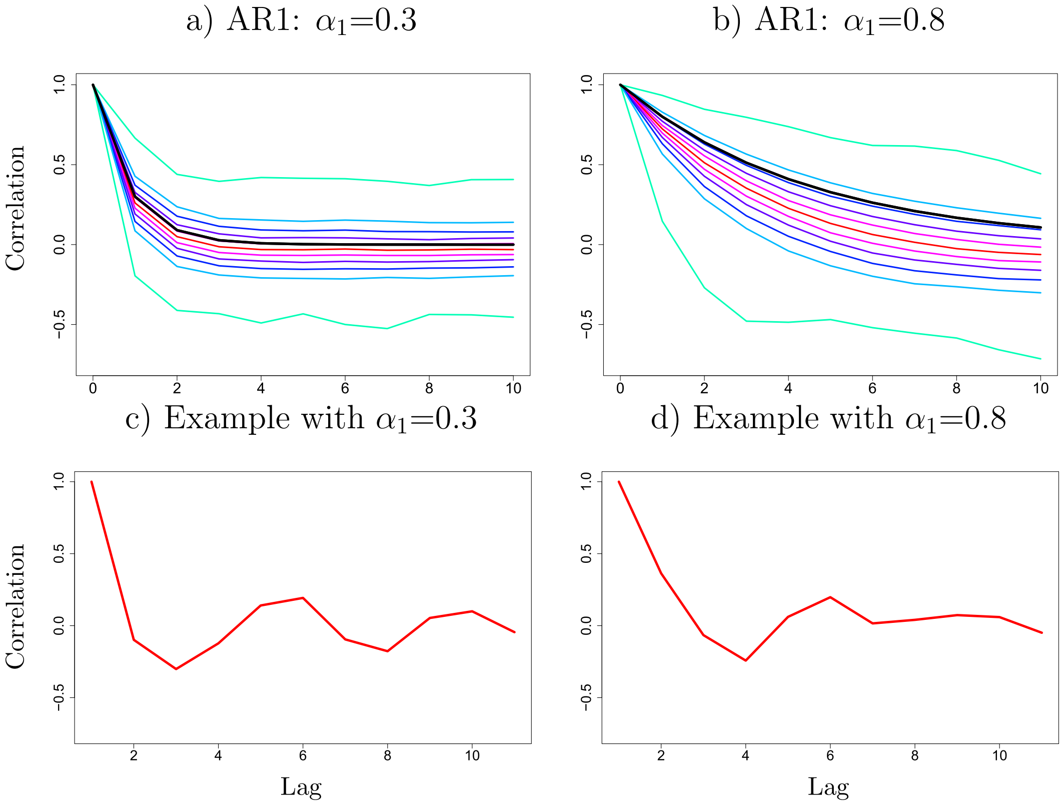


FIG. 2. True autocorrelation function of an AR1 (black lines) and quantiles of the estimated autocorrelation functions of 1000 drawings of AR1 (colored lines) with α_1 set to (a) 0.3 and (b) 0.8: 50th in red, 40th and 60th in light purple, 30th and 70th in dark purple, 20th and 80th in dark blue, 10th and 90th in light blue, and envelope (maximum and minimum) in green. Example of estimated autocorrelation function of an AR1 with α_1 set to (c) 0.3 and (d) 0.8 and for which the autocorrelation is negative at lag 2 in (c) and lag 3 in (d).

larger than $\tau = 5$, the estimated autocorrelation ranges between approximately −0.5 and 0.5. This uncertainty strongly affects the estimated $N_{\rm eq}$ shown in Fig. 1.

3. How can one reduce the uncertainty on the estimated effective number of independent data?

We successively consider two possible options and assess their added value in terms of reliability of the estimated $N_{\rm eq}$. The first method, solution 1, consists of adding the autocorrelations only until they drop below the 95% level of significance using a one-sided t test [H_o: $\rho(\tau) = 0$; H′: $\rho(\tau) > 0$] that neglects the dependence of the data:

$$N_{\rm eq} = \frac{N}{1 + 2\sum_{\tau=1}^{T_{\rm max}} \frac{N-\tau}{N}\hat{\rho}(\tau)}, \qquad (7)$$

with $T_{\rm max}$ given by

$$T_{\rm max} = \min\left[\tau \,|\, \hat{\rho}(\tau)\sqrt{\frac{N-2}{1-\hat{\rho}(\tau)^2}} \le \mathrm{qt}(0.95, N-2)\right], \qquad (8)$$

where the vertical bar (|) stands for ‘‘given that’’ and qt is a function providing the 95th quantile of the t distribution for the number of degrees of freedom $N - 2$. In this method, we aim at accounting only for the autocorrelation values which are above the level that can be considered as sampling noise corresponding to an actual null correlation.

In solution 2, we consider that our sample can be represented by a mathematical model such as an AR1 for which we need to estimate α_1. We fit the equation for the true autocorrelation function of this mathematical model [Eq. (6)] to a first guess of the estimated autocorrelation function computed from the time series in a classical way. To perform this fit, we seek for $\hat{\alpha}_1$ by minimizing the mean-square distance between the first guess of the estimated autocorrelation function and the true autocorrelation function of an AR1 for a given $\hat{\alpha}_1$:

$$\hat{\alpha}_1 = \alpha_1 | \min_{\alpha_1} \left(\sum_{\tau=1}^{N-3} \left\{ \frac{[\hat{\rho}(\tau) - \alpha_1^\tau]}{\tau} \right\}^2 \right). \qquad (9)$$

We give less and less weight, though, to the first guess of the estimated autocorrelation function as the lag increases by dividing the distance between both autocorrelation functions by τ because the uncertainty on this first guess tends to increase with the lag. This approach for estimating α_1 is slightly different from the "usual one." Indeed, in classic terms, fitting an AR1 implies estimating the autocorrelation at lag 1 and using the true relation $\alpha_1 = \rho(1)$ to estimate $\hat{\alpha}_1 = \hat{\rho}(1)$. The added value of our fitting method over the classical one is illustrated in Fig. S1 of the supplemental material. We then apply the standard formula using the true autocorrelation function of an AR1 in which α_1 would be $\hat{\alpha}_1$:

$$N_{\rm eq} = \frac{N}{1 + 2\sum_{\tau=1}^{N-1} \frac{N-\tau}{N} \hat{\alpha}_1^\tau}. \qquad (10)$$

The histograms of $N_{\rm eq}$ obtained by applying solutions 1 and 2 to 1000 AR1 are respectively shown in Figs. 3 and 4. Both solutions provide a distribution that peaks around the true $N_{\rm eq}$, indicated by a vertical brown line, whereas the original formulation provided a distribution that peaked at $N_{\rm eq} = 50$. The distance between the mean of the distribution, indicated by a vertical blue line, and the true $N_{\rm eq}$ ranges between 7.74 and 14.33 depending on α_1 for solution 1 and between 0.45 and 3.66 only for solution 2. The peak in distribution is closer to the true $N_{\rm eq}$ with solution 2 (Fig. 4) than with solution 1 (Fig. 3), and the spurious secondary peak at $N_{\rm eq} = 50$ disappears much more quickly when α_1 increases with solution 2. We thus conclude that solution 2 reduces to a larger extent, relative to solution 1, the uncertainty on the estimated number of independent data.

We have assessed the sensitivity of the $N_{\rm eq}$ provided by solution 2 to the number of lags τ used from the first guess of the estimated autocorrelation function to fit the true autocorrelation function of an AR1 to minimize the computation time. We concluded from those sensitivity tests (not shown) that the optimal number of lags to increase the reliability of the obtained $N_{\rm eq}$ is 2:

$$\hat{\alpha}_1 = \alpha_1 | \min_{\alpha_1} \left(\sum_{\tau=1}^{2} \left\{ \frac{[\hat{\rho}(\tau) - \alpha_1^\tau]}{\tau} \right\}^2 \right). \qquad (11)$$

A function CFU_eno, coded using the R language from the Comprehensive R Archive Network, is provided in the supplementary material. This function applies solution 2 to any input time series to compute its number of independent data $N_{\rm eq}$.

Figure 5 further compares the performance of the original formulation (open circles), the solution 1 (filled circles), and the solution 2 (triangles) in terms of bias (left panel), standard deviation (center panel), and root-mean-square error (RMSE; right panel) of the estimated $N_{\rm eq}$, expressed in percentage of the sample size N, for samples of length N equal to 20 (black lines), 50 (red lines), and 100 (green lines) and with α_1 between 0.1 and 0.8 as given by the x axis. The bias is systematically reduced when using solution 1 over the original formulation and when using solution 2 over solution 1. The variance is increased for low α_1 but decreased for high α_1 when using solution 2 over the original formulation. Overall, the root-mean-square error, which integrates the bias and standard deviation information, is systematically reduced when using solution 1 over the original formulation and when using solution 2 over solution 1. Whereas the root-mean-square error increases with α_1—that is, with a decrease in the true $N_{\rm eq}$—when using the original formulation, it decreases with α_1 when using the solution 2.

We came to the conclusion, in section 2, that the large uncertainty in the estimation of the effective number of independent data when using Eq. (3) originates from the large uncertainty in the estimation of the autocorrelation function. We compare in Fig. 6 the performance of our solution 2 (triangles) with the first guess of the estimated autocorrelation function (ACF; open circles) and with the maximum likelihood estimator (MLE; closed circles) in terms of bias (left panel), standard deviation (center panel), and root-mean-square error (right panel) of the estimates of the autocorrelation at lag 1. These performances have been computed on 2000 AR1 processes drawn with α_1 in the x axis and with length N equal to 20 (black lines), 50 (red lines), and 100 (green lines). Our estimator following solution 2 has a lower bias than the other two estimators for low α_1 but a larger bias for large α_1 and it has a lower variance for any α_1 for the sample sizes considered in this study. Overall, its root-mean-square error is lower than for the other two estimators for $\alpha_1 < 0.6$ for $N = 20$, for $\alpha_1 < 0.5$ for $N = 50$, and $\alpha_1 < 0.4$ for $N = 100$. Since the most common case is the small sample size and small α_1 case in climate and weather prediction, we recommend the use of our solution 2 over the maximum likelihood estimator or over the first guess of the autocorrelation function.

Thiébaux and Zwiers (1984) had suggested seven different estimators for the effective number of independent data. We have reproduced parts of their Tables 3, 4, and 5, which compare the performance of their estimators. We

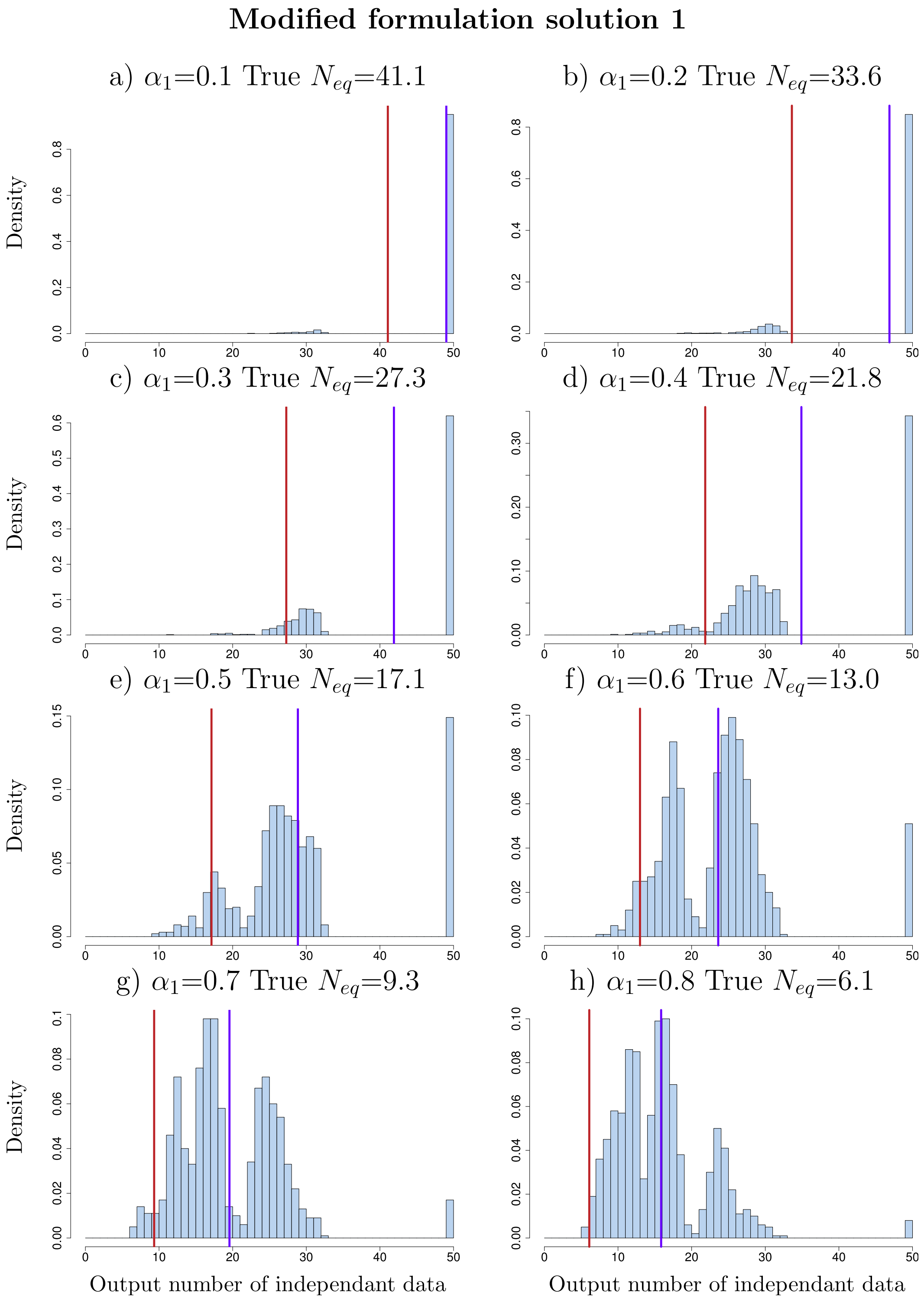


FIG. 3. Histograms of the N_{eq} provided by Eq. (7) following solution 1 (see details in the text) when applied to 1000 drawings of AR1 with α_1 set to (a) 0.1, (b) 0.2, (c) 0.3, (d) 0.4, (e) 0.5, (f) 0.6, (g) 0.7, and (h) 0.8. The mean of the distribution is provided as a vertical blue line. The true N_{eq} is provided in the title for each α_1 and as a vertical brown line.

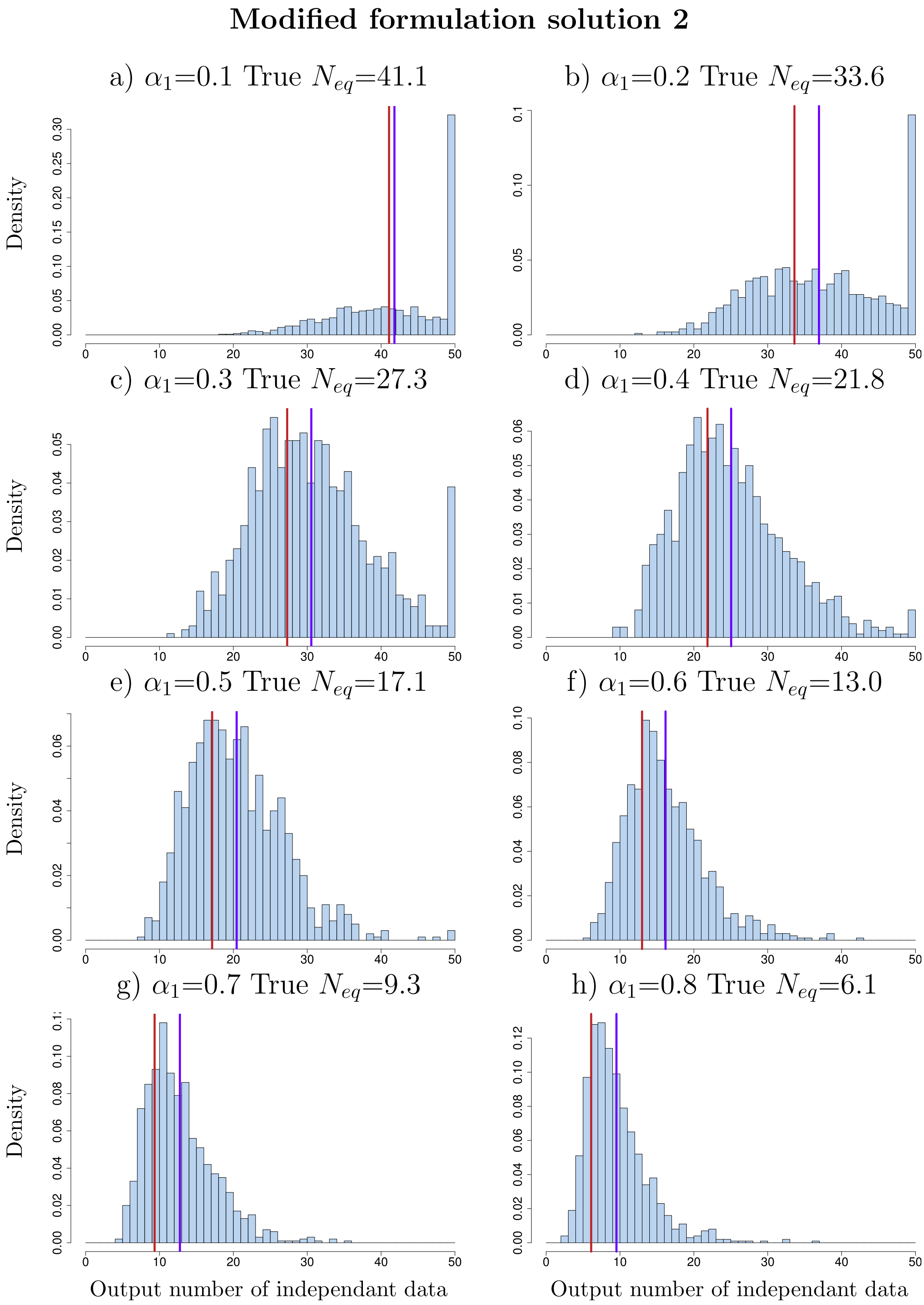


FIG. 4. As in Fig. 3, but as provided by Eq. (10) following solution 2.

have only considered the results for a sample size below 60 since one of the issues also faced in climate or weather prediction is the small sample size. Furthermore, we have not considered their example of a second-order autoregressive process that is affected by a large peak in its power spectrum and therefore does not seem representative of our typical observed data. The methods considered by Thiébaux and Zwiers (1984) consist of the following:

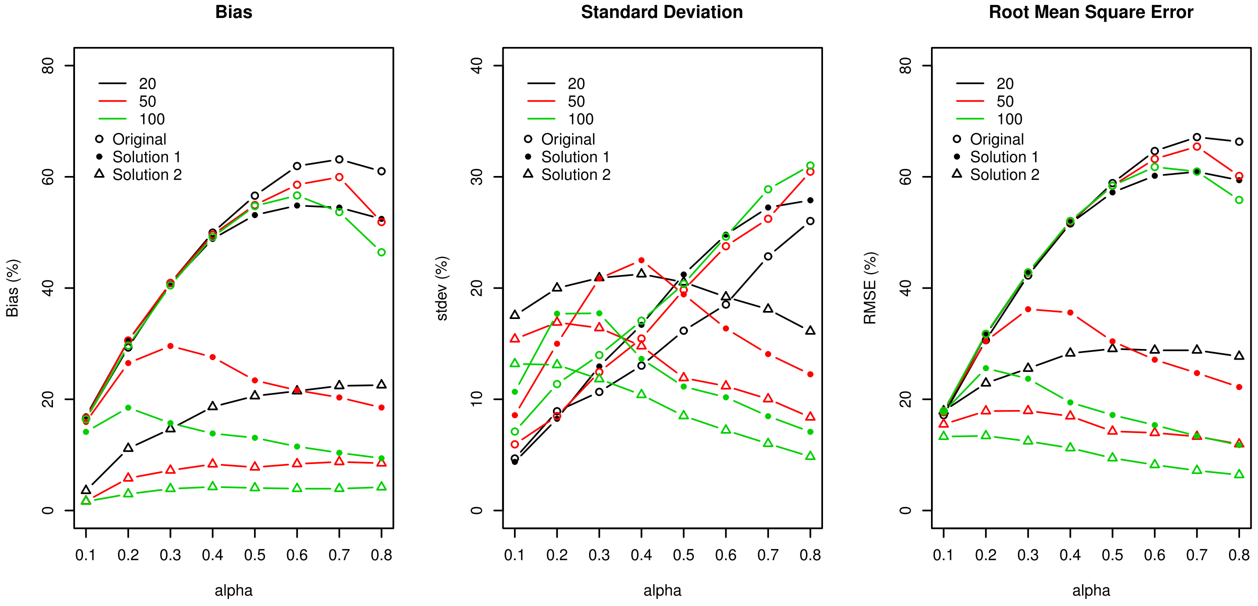


FIG. 5. The (left) bias, (center) standard deviation, and (right) RMSE of the estimates of the N_{eq} computed on 2000 drawings of AR1 with α_1 given by the x axis and length 20 (black lines), 50 (red lines), and 1000 (green lines) when using the original formulation (open circles), our solution 1 (closed circles), and our solution 2 (triangles) as estimators.

1) "DIR": applying the original Eq. (3),
2) "DIR2": truncating the original Eq. (3) at lag 10,
3) "ARMA": fitting an autoregressive moving-average process after determining its order by Alaike's information criterion and then using its theoretical autocorrelation function,
4) "SPEC1": estimating the spectrum at the origin by using the first ordinate of the smoothed periodogram (or Daniel estimator),
5) "SPEC5": estimating the spectrum at the origin by using the first five ordinates of the smoothed periodogram (or Daniel estimator),

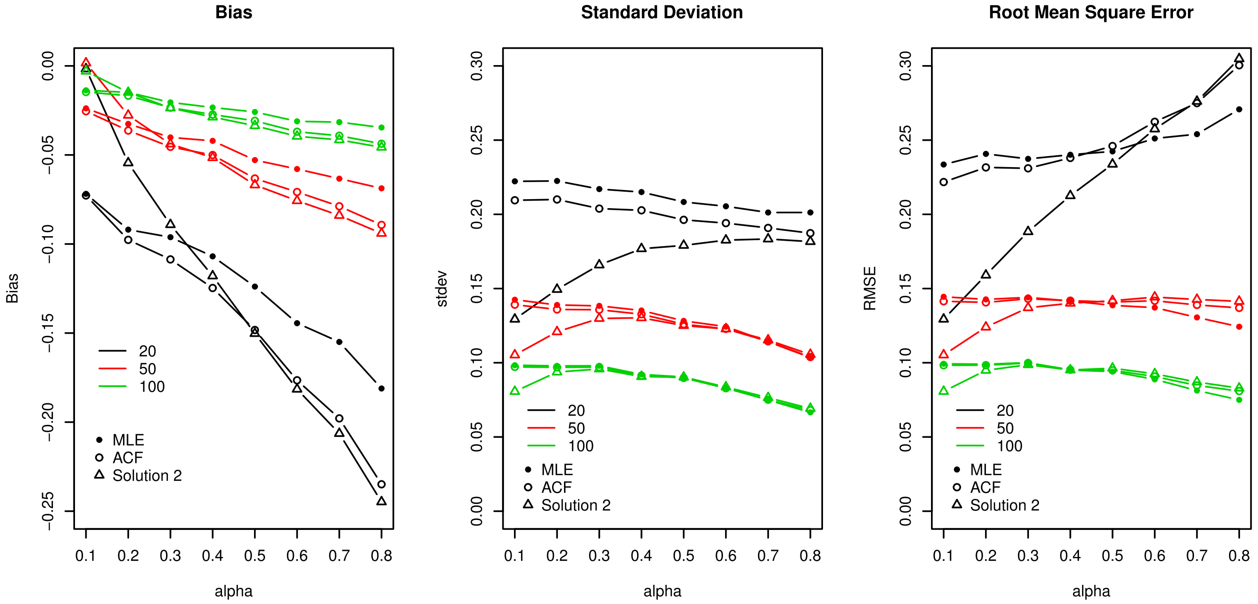


FIG. 6. The (left) bias, (center) standard deviation, and (right) RMSE of the estimates of the autocorrelation at lag 1 computed on 2000 drawings of AR1 with α_1 given by the x axis and length 20 (black lines), 50 (red lines), and 1000 (green lines) when using our solution 2 (triangles), the MLE (filled circles), and the first guess of ACF (open circles) as estimators.

TABLE 1. Partial reproduction of Table 3 from Thiébaux and Zwiers (1984) with an additional column, in boldface roman font, that provides the performance of our estimator. Shown are the true median and medians of 1000 estimates of the effective sample size made with our estimator and the seven estimators from Thiébaux and Zwiers (1984). The first two columns give the properties of the samples on which we estimate the effective sample size. Boldface italics indicate that our method is outperformed.

Model	N	True	**Ours**	DIR	DIR2	ARMA	SPEC1	SPEC5	SPECT5	BART
AR(1) $\rho = 0.30$	30	17	**19**	63	36	15	22	19	19	29
	60	33	**35**	114	49	33	47	35	***34***	41
AR(1) $\rho = 0.45$	30	12	**14**	39	26	***13***	14	16	16	18
	60	23	**26**	80	37	26	32	26	26	31
AR(1) $\rho = 0.60$	30	8	**11**	25	19	11	***10***	14	14	15
	60	15	**18**	55	26	18	21	20	19	24
AR(1) $\rho = 0.75$	30	5	**8**	19	12	8	***7***	13	13	12
	60	9	**11**	33	14	12	11	15	15	18

6) "SPECT5": estimating the spectrum at the origin by using the Daniel estimator with 20% of the data tapered with a cosine taper, and
7) "BART": using a weighted covariance estimator with a Bartlett lag window.

According to Table 1, our solution 2 provides a median of the distribution that is closer to the "true" value than do any of the estimators from Thiébaux and Zwiers (1984), except in four cases in which (α_1, N, method) = (0.30, 60, SPECT5), (0.45, 30, ARMA), (0.60, 30, SPEC1), and (0.75, 30, SPEC1). According to Table 2, our solution 2 systematically provides a mean of the distribution that is closer to the true value than do any of the estimators from Thiébaux and Zwiers (1984). According to Table 3, our solution 2 provides a standard deviation of the distribution that is lower than that of any of the estimators from Thiébaux and Zwiers (1984) except in five cases in which (α_1, N, method) = (0.45, 30, SPEC5), (0.6, 30, SPEC5), (0.6, 30, SPECT5), (0.75, 30, SPEC5), and (0.75, 30, SPECT5) although the lower standard deviations in those five cases do not compensate for the larger bias. We therefore conclude that our solution 2 outperforms the methods suggested by Thiébaux and Zwiers (1984). The better performance of our method is most probably attributable to the fact that we give a larger weight [see Eq. (9)] to the information from the first guess of the autocorrelation function at shorter lags than at longer lags, taking into account in such a way the rapid increase in uncertainty with the lag (see Fig. 2).

Last, we would like to warn the reader about a drawback of this method. Whereas the original formulation tends to provide N_{eq} that is larger than N when applied to random samples of independent data, solution 2 tends to provide N_{eq} that is slightly lower than N when applied to such samples. Using this solution might thus make the statistical tests of hypothesis too conservative.

4. Violation of the "identically distributed" hypothesis

The formula developed by Anderson (1971) and revisited in von Storch and Zwiers (2001) is demonstrable under a range of hypotheses (see the supplementary material). Each datum of the time series is considered as a particular drawing of an X random variable. The process of drawing a time series is then modeled as a list of $X_1, X_2, X_3, X_4, \ldots, X_N$ identically distributed random variables. This mathematical framework to model the sample is a classical one that has been used to develop many commonly used parametric statistical tests, except that the independence of the data is required in addition in those parametric tests (e.g., Student's test, or the Fisher test). The identically distributed hypothesis implies that the data in the sample are equivalent; that is,

TABLE 2. As in Table 1, but for Table 4 of Thiébaux and Zwiers (1984) and for the mean.

Model	N	True	**Ours**	DIR	DIR2	ARMA	SPEC1	SPEC5	SPECT5	BART
AR(1) $\rho = 0.30$	30	17	**20**	107	88	30	99	21	22	24
	60	33	**36**	185	96	42	251	41	39	44
AR(1) $\rho = 0.45$	30	12	**15**	76	60	20	71	18	18	20
	60	23	**27**	143	73	29	131	29	29	34
AR(1) $\rho = 0.60$	30	8	**12**	54	56	15	41	15	15	16
	60	15	**19**	100	47	21	154	22	21	25
AR(1) $\rho = 0.75$	30	5	**8**	48	36	15	80	13	13	13
	60	9	**12**	73	24	14	80	16	16	19

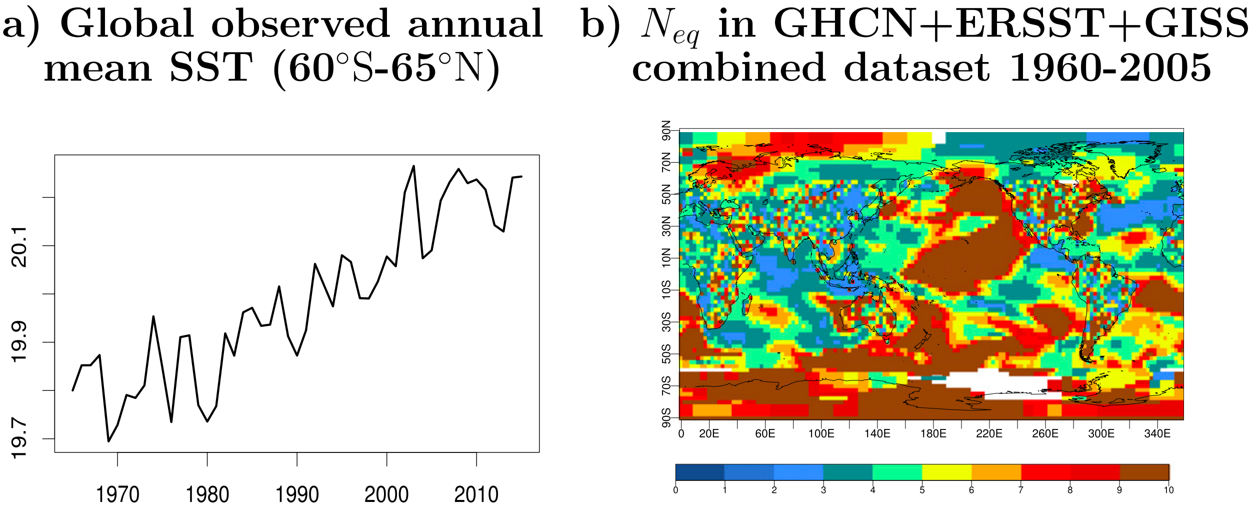
TABLE 3. TABLE 2. As in Table 1, but for Table 5 of Thiébaux and Zwiers (1984) and for the standard deviation.

Model	N	True	**Ours**	DIR	DIR2	ARMA	SPEC1	SPEC5	SPECT5	BART
AR(1) $\rho = 0.30$	30	17	**5.9**	125.9	270.3	61.2	453.0	7.5	9.3	8.5
	60	33	**9.5**	215.8	219.6	34.5	1247	21.9	20.4	15.2
AR(1) $\rho = 0.45$	30	12	**5.3**	99.9	151.0	24.8	481.5	***4.9***	5.7	7.1
	60	23	**7.4**	188.0	153.9	14.6	742.2	11.0	14.2	12.0
AR(1) $\rho = 0.60$	30	8	**4.6**	79.9	225.0	15.6	195.4	***3.6***	***4.3***	4.8
	60	15	**5.7**	130.7	129.5	11.3	2093	7.3	7.5	6.8
AR(1) $\rho = 0.75$	30	5	**3.5**	82.3	120.1	63.6	1402	***2.7***	3.1	4.0
	60	9	**4.3**	117.5	43.9	9.3	1342	4.4	5.2	4.8

they are not influenced by any physical underlying process that would make the distribution of two X_i different.

The identically distributed hypothesis is a strong hypothesis that in many cases might not be verified. The annual cycle, driven by the insolation, makes an annual oscillation in the climate variable. The fact that summer and winter processes are not identically distributed is clear in the scientific community, and summer and winter values are never compared. Because of the El Niño–Southern Oscillation, the climate variable distribution might be different during warm and cold phases. In more general terms, in a time series that is affected by any climate oscillation, a datum in one phase cannot be represented by the same X as one in the other phase. Also, the role of climate change induces a slow change in the distribution of all of the climate variables. In any time series that is affected by a trend such as the one of climate change, we should consider that the mean of the random variable X is changing with time. Otherwise, a spurious correlation between the random variables X_i is introduced, and it affects the estimated equivalent number of observations.

To illustrate the impact of neglecting this identically distributed hypothesis, we consider the example of the global annual mean sea surface temperature (SST) over the 1960–2010 period (Fig. 7a), which is affected by a strong trend. The number of independent data provided by solution 2 applied to this example containing $N = 51$ actual data is $N_{\text{eq}} = 4.8$ because the trend stands as a major part of the signal; indeed, if we consider a sample drawn for a simple line with positive trend and containing $N = 51$ actual data, we obtain $N_{\text{eq}} = 2.3$, which is only slightly lower than the result obtained for the global mean SST. The autocorrelation function estimated from the global mean SST time series is only slightly lower than 1.0 because of the weak departures


a) Global observed annual mean SST (60°S-65°N)

b) N_{eq} in GHCN+ERSST+GISS combined dataset 1960-2005

FIG. 7. (a) Global (60°S–65°N) annual mean SST from the National Oceanic and Atmospheric Administration (NOAA) Extended Reconstructed SST (ERSST) v3b dataset (Smith et al. 2008) over the 1960–2005 period. (b) Effective number of independent observations for verification of the average 2–5-yr forecast temperature in the CMIP5 decadal prediction initialized every 5 yr over the 1960–2005 period (10 predictions). The number of independent observations is computed from the merged dataset using land air temperatures from the Global Historical Climatology Network/Climate Anomaly Monitoring System dataset (Fan and van den Dool 2008), SST from the NOAA ERSST V3b dataset (Smith et al. 2008), and the Goddard Institute for Space Studies Surface Temperature Analysis (GISTEMP) dataset with 1200-km decorrelation scale (Hansen et al. 2010) outside the band between 60°N and 60°S.

from the trend, thus resulting in a slightly larger N_{eq} than 2.3. In the mathematical framework under which the formula was developed, a sample drawn for a simple line corresponds to two phases of a long oscillation: a long phase of below-average anomalies followed by a long phase of above-average anomalies, hence leading to a number of independent data that is close to 2.

When verifying against observational datasets, the decadal temperature predictions produced within the framework of phase 5 of the Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012), which are initialized every 5 years from 1960 to 2005, 9–10 observational data are available depending on the forecast time considered. When applying solution 2 to the near-surface temperature averaged over forecast years 2–5 from this climate-prediction case, the number of independent data for verification is about 2 in the Indian and North Atlantic Oceans (Fig. 7b). If the physical mechanisms targeted by the analyses are contained in the departures from the long-term trend and if the trend has large amplitude as compared with the departures from the trend, the formula we discuss here should rather be applied to a detrended time series. The number of independent data provided by solution 2 applied to the linearly detrended global annual mean sea surface temperature shown in Fig. 7a is $N_{\mathrm{eq}} = 31.6$. A linear trend stands as a crude estimation of the climatic change response. However, the N_{eq} obtained with such a simple method seems already much closer to the true N_{eq} if one is interested in the climate signal contained in the departures from the trend. The added value of a linear detrending is illustrated in Fig. S2 of the supplemental material for various cases of α_1 and slope of trends.

5. Discussion

The use of Eq. (2) combined with classical parametric or nonparametric inferential tests suitable for each prediction skill score is standard practice in the research fields of climate and weather prediction. The mathematical theory that led to such a formula, however, initially considered only tests of the means. Its extension to other statistical inferential tests seems not to be as accurate as its original usage. On the one hand, from a physical point of view, the independence between two data depends solely on the underlying physical processes that induce (or not) a temporal coherence between two separate dates. On the other hand, from a statistical point of view, the concept of equivalent sample size depends on the statistical context—that is, the statistical parameter of interest, the particular inferential tests, or the distribution of the random variables that model the sample drawing process (see von Storch and Zwiers 2001), and the equivalent sample size is equal to the number of independent data that would provide the same amount of information about a given statistical parameter as the available sample of dependent data, that is, the same standard deviation of the estimator. The unification of both points of view is not straightforward, and the choice of which statistical model is the most suitable to represent a given physical process remains the responsibility of the physicist. As long as well-tested methods for the case of serial dependence of the data are not available for each one of the statistical inference tests, as is the case for tests of the means (Zwiers and von Storch 1995), an effective number of independent data has to be estimated using one of the currently available options. Because of the large uncertainties in the estimate obtained with the original formula following Eq. (2) and the better performance of our approach than any other one available to our knowledge, we recommend the use of our solution 2.

To improve the robustness of the equivalent sample size estimation, we have relied on the statistical model of first-order autoregressive process. More general statistical models such as higher-order autoregressive processes and autoregressive moving average processes (ARMA) could have been used. Those models might represent in a slightly more accurate way the physical processes of interest. However, fitting such models would require determining additional sets of parameters. When dealing with applicative purposes, one has to balance the benefits of using a more complex model with the uncertainties brought by the determination of additional parameters. As we have shown in Fig. 2, the information from the first guess of autocorrelation function is rapidly lost as the lag increases. A strength of our solution 2 lies in the weights we give to the different lags from the first guess of autocorrelation function to estimate a single parameter α_1, as explained in section 3. This would not be possible with statistical models of higher orders. Indeed, the solution proposed by Thiébaux and Zwiers (1984), which relies on fitting an ARMA statistical model to the sample to estimate its number of effective independent data, is outperformed by our solution 2 as mentioned in section 3.

6. Conclusions

Most climate and weather studies make use of statistical tests of hypotheses and confidence interval estimation to assess the robustness of their conclusions. The statistical methods involved in those assessments require independent data in the time series to which they are applied. Those time series are unfortunately usually affected by the strong persistence of meteorological and climatic phenomena. A common approach to cope with this obstacle is to replace in those statistical methods the actual number of data by an estimated effective number of independent data that is computed from the well-known formula of Anderson (1971) and von Storch and Zwiers (2001). Even though this formula is demonstrable under some hypotheses, it provides unreliable results on practical examples because an estimated autocorrelation function is used to apply this formula and this latter is affected by a large uncertainty due to the usual shortness of the available time series. Our recommendation is to fit the autocorrelation function of a first-order autoregessive process to the estimated one prior to the application of Anderson (1971) formula to reduce this uncertainty. An R software function called CFU_eno is made available to the reader to implement this method. In addition to not respecting the hypothesis of independence of the data, many meteorological and climatic time series also do not respect the hypothesis of stationarity of the time series—for example, when affected by climate change. The estimated number of independent data is heavily affected by the existence of a trend. The CFU_eno function we make available in the supplementary material contains an option to linearly detrend the time series prior to the computation of the number of independent data for cases in which the climate signal that is targeted by the analyses is contained in the departures from the trend. More details are provided in the appendix about the various R functions made available to the reader to apply the solution proposed in this article.

Acknowledgments. John Maxwell Halley is greatly acknowledged for the fruitful discussion about this article. This work was supported by the EU-funded QWeCI (FP7-ENV-2009-1-243964), CLIM-RUN projects (FP7-ENV-2010-265192), and SPECS (FP7-ENV-2012-308378) projects; the MINECO-funded RUCSS (CGL2010-20657) and PICA-ICE (CGL2012-31987) projects; and the Catalan government. The authors thank the two anonymous reviewers for their fruitful suggestions and their substantial contribution to the improvement of our article.

APPENDIX

Tools Provided as Supplementary Materials

A set of seven functions written in the R programming language is provided to the reader as part of the supplementary material. Those R functions are distributed to

1) enable the reader to apply the methods we suggest in our article on his or her own dataset by using the CFU_eno function described below, which is dependent on CFU_alpha and fitacfcoef also described below (and potentially dependent on CFU_spectrum and/or CFU_filter depending on the input arguments to CFU_eno);
2) reproduce our tests and results by using CFU_gen_series to generate autoregressive processes and by applying CFU_eno with different input arguments; and
3) have access to the exact implementation details of our method by reading our code.

A brief description of each function is provided below:

1) CFU_eno: This function estimates the effective number of independent data of the input array *xdata.* It has one compulsory argument (the *xdata* array of which the effective number of independent data has to be estimated) and two optional flag arguments [*detrend* (default: “FALSE”) to apply a linear detrending prior to the estimation of the effective sample size and *filter* (default: “FALSE”) to apply a filtering of any periodic signal prior to the estimation of the effective sample size]. This function calls CFU_alpha to obtain a refined estimate of the autocorrelation function at lag 1 following the solution 2 presented in our article, and then it applies Eq. (10) to obtain the effective number of independent data. The method used to apply the linear detrending and to filter any periodic signal is described below in the comments on the CFU_alpha function.
2) CFU_alpha: This function estimates the autocorrelation at lag 1 of the input array *xdata.* It has one compulsory argument (the *xdata* array of which the autocorrelation at lag 1 has to be estimated) and two optional flag arguments [*detrend* (default: “FALSE”) to apply a linear detrending prior to the estimation of the autocorrelation at lag 1 and *filter* (default: FALSE) to apply a filtering of any periodic signal prior to the estimation of the autocorrelation at lag 1] This function, after a potential linear detrending and periodic signal filtering on *xdata*, estimates its first guess of autocorrelation function and then calls fit_acfcoef to refine the estimate of the autocorrelation function at lag 1 following our solution 2. If *detrend* = “TRUE,” the linear trend of the *xdata* array is estimated together with its 95% confidence interval. If this confidence interval does not encompass 0, the minimum absolute value of the confidence interval limits is used as the slope of the linear trend to be removed. Indeed, a nonnull autocorrelation at lag 1 induces a trend in the *xdata* array so that the benefits of subtracting the linear trend for large slopes have to be balanced with the removal of part

of the autocorrelation signal at lag 1 for weak slopes. The results of our tests regarding the impact of a linear detrending are illustrated in Fig. S2 of the supplemental material. If *filter* = "TRUE," the frequency spectrum of *xdata* is estimated by calling CFU_spectrum. If any peak in the obtained frequency spectrum is significant at the 99% level, such peak is filtered out by calling CFU_filter.

3) CFU_gen_series: This function generates first-order autoregressive processes containing *n* data, with *alpha* as autocorrelation at lag 1 and mean and standard deviation provided by the *mean* and *std* arguments. It has four compulsory arguments: the *n*, *alpha*, *mean*, and *std* values. This function can be used by the reader to reproduce the various tests, the results of which are presented in this article.
4) fit_acfcoef: This function finds the minimum point of the fourth-order polynomial $(a - x)^2 + 0.25(b - x^2)^2$ written to fit the two autoregression coefficients *a* and *b* with the Cardan formula. It has two compulsory arguments: the *a* and *b* values. Provided that *a* and *b* are in the $[0, 1]$ interval, $\Delta > 0$ and there is only one solution to the minimum. This function can be used to minimize the mean-square differences between the true autocorrelation function of an AR1 and the first guess of estimated autocorrelation function using only the first two lags.
5) CFU_spectrum: This function estimates the frequency spectrum of the *xdata* array together with its 95% and 99% significance levels. It has one compulsory argument: the *xdata* array. Its output is provided as a matrix with dimensions (number of frequencies, 4). The second dimension contains the frequency values, the power, the 95% significance level, and the 99% one. The spectrum estimation relies on an R built-in function and the significance levels are estimated by a Monte Carlo method. This function can be used to detect the potential periodic signals in the *xdata* array that might need to be filtered out prior to the computation of the effective number of independent data to avoid a violation of the "identically distributed" hypothesis.
6) CFU_filter: This function filters from the *xdata* array, the signal of frequency *freq*. The filtering is performed by dichotomal seek for the frequency around *freq* and the phase that maximizes the signal to subtract from *xdata.* It has two compulsory arguments: an *xdata* array and the *freq* value. The maximization of the signal to subtract relies on a minimization of the mean-square differences between *xdata* and a cosine of given frequency and phase. As highlighted in section 4, the presence of a periodic signal induces a violation of the "identically distributed" hypothesis. Such a periodic signal can be filtered out prior to the computation of the effective number of independent data using this function.
7) CFU_fitautocor: This function can be used to minimize the mean-square differences between the true autocorrelation function of an AR1 and the first guess of estimated autocorrelation function using any range of lags, but it is less computationally efficient than fit_acfcoef for the particular case of two lags as advised in our article. This function has one compulsory argument (the first guess of estimated autocorrelation function, which can contain any range of lags) and two optional arguments [the window in which the output autocorrelation at lag 1 should lie (default: $[-1; 1]$) and the precision to which the output autocorrelation at lag 1 should be determined (default: 0.01)]. The estimation of the output autocorrelation at lag 1 relies on a dichotomial minimization of the mean-square differences between the true autocorrelation function of an AR1 and the first guess of the autocorrelation function.

REFERENCES

Anderson, T. W., 1971: *The Statistical Analysis of Time Series*. John Wiley and Sons, 704 pp.

Bayley, G. V., and J. M. Hammersley, 1946: The "effective" number of independent observations in autocorrelated time series. *J. Roy. Stat. Soc. Suppl.,* **8,** 184–197.

Fan, Y., and H. van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.,* **113,** D01103, doi:10.1029/2007JD008470.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.,* **48,** RG4004, doi:10.1029/2010RG000345.

Jones, R. H., 1975: Estimating the variance of time averages. *J. Appl. Meteor.,* **14,** 159–163, doi:10.1175/1520-0450(1975)014<0159:ETVOTA>2.0.CO;2.

Knight, J. R., R. J. Allan, C. K. Folland, M. Vellinga, and M. E. Mann, 2005: A signature of persistence natural thermohaline circulation cycles in observed climate. *Geophys. Res. Lett.,* **32,** L20708, doi:10.1029/2005GL024233.

Leith, C. E., 1973: The standard error of time-average estimates of climatic means. *J. Appl. Meteor.,* **12,** 1066–1069, doi:10.1175/1520-0450(1973)012<1066:TSEOTA>2.0.CO;2.

Palmer, T., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, and L. Smith, 2005: Ensemble prediction: A pedagogical perspective. *ECMWF Newsletter,* No. 106, ECMWF, Reading, United Kingdom, 10–17. [Available online at http://www.ecmwf.int/publications/newsletters/pdf/106.pdf.]

Smith, T. M., R. W. Reynolds, and T. C. P. H. Lawrimore, 2008: Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Climate,* **21,** 2283–2296, doi:10.1175/2007JCLI2100.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.,* **93,** 485–498, doi:10.1175/BAMS-D-11-00094.1.

Thiébaux, H. J., and F. W. Zwiers, 1984: The interpretation and estimation of effective sample size. *J. Climate Appl. Meteor.,* **23,** 800–811, doi:10.1175/1520-0450(1984)023<0800:TIAEOE>2.0.CO;2.

Trenberth, K. E., 1984: Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations. *Mon. Wea. Rev.,* **112,** 2359–2368, doi:10.1175/1520-0493(1984)112<2359:SEOFSS>2.0.CO;2.

von Storch, H., and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

Wilks, D. S., and R. L. Wilby, 1999: The weather generation game: A review of stochastic weather models. *Prog. Phys. Geogr.,* **23,** 329–357, doi:10.1177/030913339902300302.

Zwiers, F. W., and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *J. Climate,* **8,** 336–351, doi:10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2.