Short Communication

São Paulo City is getting 2 °C hotter in the last 50 years: regression model with autoregressive errors or with lagged temperatures?

Airlane P. Alencar[1]

Received: 16 July 2020 / Accepted: 31 October 2020 / Published online: 11 November 2020

Abstract
This study presents two time series models to estimate the mean increase in the monthly mean temperature in São Paulo City from 1960 to 2017. The basic model consists of a linear regression model including trend and sine and cosine functions to consider seasonality. As the errors are supposed to be autocorrelated for time series, we can include in the regression model the lagged temperatures or autoregressive errors. The first approach is often used in practice, but the trend parameter estimator is biased to estimate the long-run trend effect. The unbiased trend effect estimator is presented with its variance and confidence interval. The second approach provides directly the unbiased trend estimator. Finally, there is evidence that the temperature trend effect is constant over time and both models lead to a significant increase of 1.9 °C in the last 50 years in São Paulo City. The 95% confidence interval is equal to [1.6; 2.2] for the model with autoregressive errors, which is beyond the limits announced in the Paris Agreement of 2015.

Keywords Global warming · Temperature · Autoregressive models with covariates · Regression model with autoregressive errors

Mathematics Subject Classification 62M10

1 Introduction

One of the main global concerns is global warming and its impact on our lives. Temperature is increasing worldwide [13], and the main goal of the Paris Agreement (2015) is to take efforts to maintain the rise of the global temperature in our century below 2 °C above pre-industrial levels and limit the temperature increase further to 1.5 °C.

There are many lines of evidence about global warming. In March 2020, there were 1.89 million published papers including the words global warming and temperature increase in the Google Scholar website. Some of them discuss the impact of global warming, and that evidence of the increase in temperature is undeniable. The global annual temperature has increased at an average rate of 0.18 °C per decade since 1981 [15].

This paper aims to present and discuss time series models to estimate the increase in temperature. In special, the main goal of this analysis consists of estimating the mean temperature increase in the last 50 years based on the monthly mean temperature from 1961 to 2017 in the city of São Paulo, Brazil.

São Paulo is a vast city with an estimated population of 12 million people in 2016 [10]. Another motivation for this analysis is that in the last ten years the number of vehicles increased by 39% in São Paulo, passing from 6,369,581 in 2008 to 8,861,208 in 2018 (Detran), whereas the population increased by 8% in the last two decades.

Linear regression models are convenient to estimate the linear trend and seasonality, but significant residual autocorrelations show that the usual assumption of error independence is violated when analyzing time series. To

✉ Airlane P. Alencar, lanealencar@usp.br | [1]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil.

estimate the mean temperature increase, we propose a linear regression model including linear trend and sine and cosine functions, but it is necessary to include also autoregressive errors, named AR error model. The alternative model is a linear regression model in which we include the lagged temperature (called ARX model) and consider the serial correlation between the temperatures in consecutive months. In this last model, the trend effect is accumulated in the long term and depends not only on the parameter of the linear trend, as some practitioners may think, but also on the parameters associated with the lagged temperatures.

In this paper, we propose, fit, and discuss both models. First, we check if all model assumptions are valid through a residual analysis. Then, the mean temperature increase is estimated for 50 years with the corresponding prediction interval. The method to estimate the mean annual increase is explained, and more complex calculations are required to estimate the long-term trend effect for the model with lagged temperatures.

Based on the observed monthly mean temperature from 1961 to 2017 in the city of São Paulo, it was observed a mean increase of 0.4 °C in each decade, and a mean increase of approximately 2 °C in the last 50 years. We also fitted a dynamic linear model only to confirm that the trend effect is not varying over time.

This paper unfolds as follows. Section 2 presents the method used. Section 3 presents all the estimation results. Section 4 discusses the advantages and disadvantages of each model. Finally, some conclusions are given in Sect. 5, along with some explanations on why the model with autoregressive errors is easier to estimate covariate effects, such as the trend effect.

2 Models, estimation methods, and residual analysis

The monthly mean temperatures in São Paulo, measured at the Mirante de Santana station from 1961 to 2017, were obtained from the Instituto Nacional de Pesquisas Espaciais (INPE) [11]. This time series consists of 684 monthly observations, corresponding to 57 years. This time series presents 20 missing values, mainly in the beginning of the 80's. No imputation method was applied and the models proposed in this paper may be fitted with few missing values.

The basic model includes a linear trend and harmonic components to take into account the seasonality as in [1]. Then, the regression model for the mean temperature in the tth month, y_t, is given by

$$y_t = \beta_0 + \beta_1 t/12 + \sum_{j=1}^{6}\left[\gamma_{cj}\cos\left(\frac{2\pi jt}{12}\right)\right] + \sum_{j=1}^{5}\left[\gamma_{sj}\sin\left(\frac{2\pi jt}{12}\right)\right] + e_t, \quad (1)$$

for $t = 1, \ldots, 684$.

Note that all 11 harmonics, $\cos(2\pi jt/12)$ and $\sin(2\pi jt/12)$, were included, except the sine function for $j = 6$ since $\sin(2\pi 6t/12) = \sin(\pi t) = 0$. Beyond the orthogonality of these trigonometric functions, another important feature is that for any integer k we have

$$\sum_{t=k+1}^{k+12}\cos\left(\frac{2\pi jt}{12}\right) = 0, \quad j = 1, \ldots, 6$$
$$\sum_{t=k+1}^{k+12}\sin\left(\frac{2\pi jt}{12}\right) = 0, \quad j = 1, \ldots, 5.$$

Then, the seasonal component oscillates around zero and its inclusion in (1) implies that the mean of y_t oscillates around the linear trend $\beta_0 + \beta_1 t/12$.

The main concern is that the error process is probably autocorrelated. Then, it was proposed that the error may be modeled as an autoregressive (AR error) process, as suggested by the residual autocorrelation and partial autocorrelation functions.

After removing the nonsignificant harmonic terms, based on the likelihood ratio test to verify the null hypothesis that 7 coefficients are all simultaneously null (p=0.5521), the model with AR(1) errors is

$$y_t = \beta_0 + \beta_1(t/12) + \sum_{j=1,2,6}\gamma_{cj}\cos\left(\frac{2\pi jt}{12}\right) + \gamma_{s1}\sin\left(\frac{2\pi t}{12}\right) + e_t, \quad (2)$$
$$e_t = \phi e_{t-1} + a_t,$$

$t = 1, \ldots, 684$, where a_t are zero-mean, independent, Gaussian, and homoscedastic random errors. Also, after a sequence of replacements, the error can be written as $e_t = \sum_i \phi^i a_{t-i}$ and the process e_t is stationary assuming that $|\phi| < 1$.

This model was estimated using the maximum conditional likelihood method. The likelihood is the product of the Gaussian conditional density functions of y_t given all the available information up to $t-1$ [19]. The asymptotic distribution of these maximum likelihood estimators is Gaussian and their variance are obtained from the inverse of Fisher information matrix. More details are, for example, in [3, 19].

In this analysis, the model was estimated using the Arima command of the `forecast` library in [8] using the `R programming language`.

As an alternative, we also proposed a model that includes the last observation y_{t-1} in the regression model to take into account the autocorrelation. This proposal is simpler to practitioners since it is a regression model where the lagged observation is another covariate. We call this model ARX(1) model because it may be understood as an AR(1) model with covariates (X = trend and seasonal components), as the ARMAX model in [19]. The ARX(1) model is defined for $t = 1, \ldots, 684$ as

$$y_t = \beta_0 + \beta_1(t/12) + \sum_{j=1,2,6} \gamma_{cj} \cos\left(\frac{2\pi jt}{12}\right) + \gamma_{s1} \sin\left(\frac{2\pi t}{12}\right) + \phi y_{t-1} + e_t, \tag{3}$$

where e_t are independent Gaussian errors with standard deviation σ_e and $|\phi| < 1$.

The conditional mean of the temperature is

$$E(y_t|y_{t-1}) = \beta_0 + \beta_1(t/12) + \sum_{j=1,2,6} \gamma_{cj} \cos\left(\frac{2\pi jt}{12}\right) + \gamma_{s1} \sin\left(\frac{2\pi t}{12}\right) + \phi y_{t-1},$$

but the annual increase in the mean temperature does not directly correspond to β_1 since y_{t-1} also depends on t. By increasing one month, the covariate $x_t = t$ goes to $t + 1$, and in the first month y_{t+1} increases on average β_1, in the second month y_{t+2} increases $\beta_1 + \beta_1\phi$ and after h months y_{t+h} increases $\beta_1(1 + \phi + \cdots + \phi^h) = \beta_1 \frac{1-\phi^{h+1}}{1-\phi}$ [7, 9]. As the process is a stationary autoregressive process, $|\phi| < 1$. Then, in the long term (as $n \to \infty$), the mean annual increase corresponds to

$$\beta_1^* = \frac{\beta_1}{1-\phi}. \tag{4}$$

Using the Delta method [18] and assuming that $|\phi| < 1$, the asymptotic variance of the maximum likelihood estimator $\widehat{\beta}_1^* = \widehat{\beta}_1/(1-\widehat{\phi})$ is

$$Var(\widehat{\beta}_1^*) = \frac{1}{(1-\phi)^2}\left[Var(\widehat{\beta}_1) - 2\frac{\widehat{\beta}_1 Cov(\widehat{\beta}_1, \widehat{\phi})}{1-\phi} + \frac{\widehat{\beta}_1^2 Var(\widehat{\phi})}{(1-\phi)^2}\right]. \tag{5}$$

The estimated variance is obtained by replacing the parameters β_1 and ϕ by their respective estimates.

The residual analysis of both models included residual time series plots, residual autocorrelations and residual quantile–quantile (QQ) plots. After concluding that all model assumptions are valid, the predicted temperatures were calculated for both models. The predicted temperatures for the AR(1) error model in (2) are

$$\widehat{y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t/12 + \sum_{j=1,2,6} \widehat{\gamma}_{cj} \cos\left(\frac{2\pi jt}{12}\right) + \widehat{\gamma}_{s1} \sin\left(\frac{2\pi t}{12}\right), \tag{6}$$

and, for the ARX(1) model in (3), the prediction includes the lagged term

$$\widehat{y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t/12 + \sum_{j=1,2,6} \widehat{\gamma}_{cj} \cos\left(\frac{2\pi jt}{12}\right) + \widehat{\gamma}_{s1} \sin\left(\frac{2\pi t}{12}\right) + \widehat{\phi} y_{t-1}. \tag{7}$$

The temperature forecasts for the next 50 years are calculated with corresponding 95% prediction intervals using the AR(1) error model through the `forecast` library [8] in the R software [17]. The variance of forecasts using ARMA models are obtained, writing the model as an infinite order-MA model as explained in [3]. As we are considering an AR(1) error model, the coefficients of the MA model are $\theta_i = \phi^i$. Including the covariates at time h ($\mathbf{x}_h$), the variance of a forecast h steps ahead is

$$Var(\widehat{y}_{t+h}|\text{observed until } y_t) = \sigma^2\left(1 + \sum_{i=1}^{h-1} \theta_i^2\right) + \mathbf{x}_h' \mathbf{Var}(\widehat{\boldsymbol{\beta}})\mathbf{x}_h. \tag{8}$$

A state space model [16] with a time-varying trend and also seasonality was also proposed only to verify if the trend parameter is constant over time. The non-observed state variables to measure the time-varying trend β_t evolves as a random walk in

$$y_t = \beta_0 + \beta_t t + s_t + v_t,$$
$$s_t = -s_{t-1} - \cdots - s_{t-11} + w_t^S,$$
$$\beta_t = \beta_{t-1} + w_t^\beta,$$

where β_0 is the intercept, β_t is the time-varying trend parameter, each s_t is the monthly effect and the errors $v_t \sim N(0, V)$, $w_t^S \sim N(0, W^S)$ and $w_t^\beta \sim N(0, W^\beta)$ are independent. In state space models [16], it is usual to assume that the seasonal effects s_t are random variables such that $s_t + s_{t-1} + \cdots + s_{t-11} = w_t^S$ have zero-mean and the time-varying effect is a random walk.

This state space model was fitted using the Kalman Filter and the maximum likelihood method through the dlm library [16] in the R software [17]. The predicted time-varying effect $\hat{\beta}_t$ is obtained with the smoothed equations

of the Kalman filter and is plotted to verify if it may be considered a constant or time-varying effect.

The computing code used and the database is available upon request.

3 Results

In this section, we shall present the results of the present temperature data analysis.

Figure 1 shows the monthly mean temperature in São Paulo. We note a linear increase from 1960, and the mean temperature increase is around 0.4 °C per decade as presented in Table 1. This corresponds to a mean increase of almost 2 °C in the last 50 years ($0.39 \times 5 = 1.95$ °C).

All the estimates and corresponding standard errors for the proposed models are presented in Table 2. For both models, the estimated annual mean increase in temperature is 0.038 °C, which corresponds to an increase of 1.9 °C in 50 years with 95% confidence intervals equal to $[1.91 \mp 0.29] = [1.62;2.20]$ for the AR(1) error model and $[1.91 \mp 0.43] = [1.48;2.34]$ for the ARX(1) model.

Both models, the AR(1) error model and the ARX(1) model, presented nonsignificant residual autocorrelations (Fig. 2) and also there is no significant residual autocorrelation until lag 10 according to the Ljung–Box test (p value $= 0.5161$ for the AR(1) error model and p value $= 0.2424$ for the ARX(1) model) [2]. Also, Fig. 2 shows that the residuals are homoscedastic and normally distributed. Then, all model assumptions are valid for the AR(1) error and ARX(1) models and the prediction intervals and forecast may be used for inference.

Figure 3 presents the predicted values only from 2010 for better visualization. The predictions for the AR(1) error model in (6) and ARX(1) model in (7) are similar and close to the observed temperature values.

Notice that the weather will be warmer in São Paulo if the current trend remains the same in the coming years, as observed in Fig. 4. These forecasts are calculated assuming that the trend will remain the same for a long time and it is shown only to visualize the future temperatures.

Based on the dynamic model with time-varying trend effect, the predicted time-varying effect $\hat{\beta}_t$ oscillates close to 0.0383 as shown in Fig. 5. The variance of the time-varying error is too small (< 0.0001), indicating that the linear trend effect does not change over time.

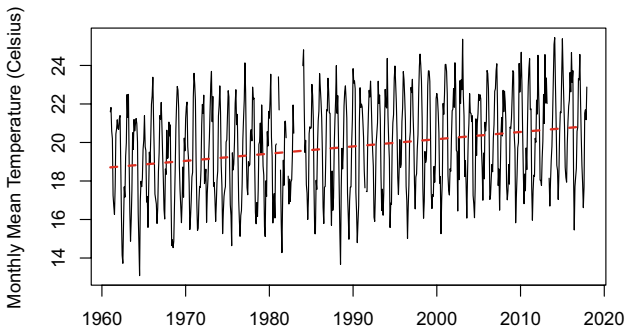


Fig. 1 Monthly mean temperature in São Paulo City from 1960 to 2017

Table 1 Descriptive statistics of monthly mean temperature by decade in São Paulo

Years	Average	Difference	Minimum	Maximum	Range
61–69	18.8		13.1	23.4	10.3
70–79	19.2	0.4	14.6	24.1	9.5
80–89	19.5	0.3	13.7	24.8	11.1
90–99	20.1	0.6	14.8	24.6	9.8
00–09	20.4	0.3	15.3	25.4	10.1
00–17	20.8	0.4	15.5	25.4	10.0
Mean	19.8	0.4			

4 Discussion

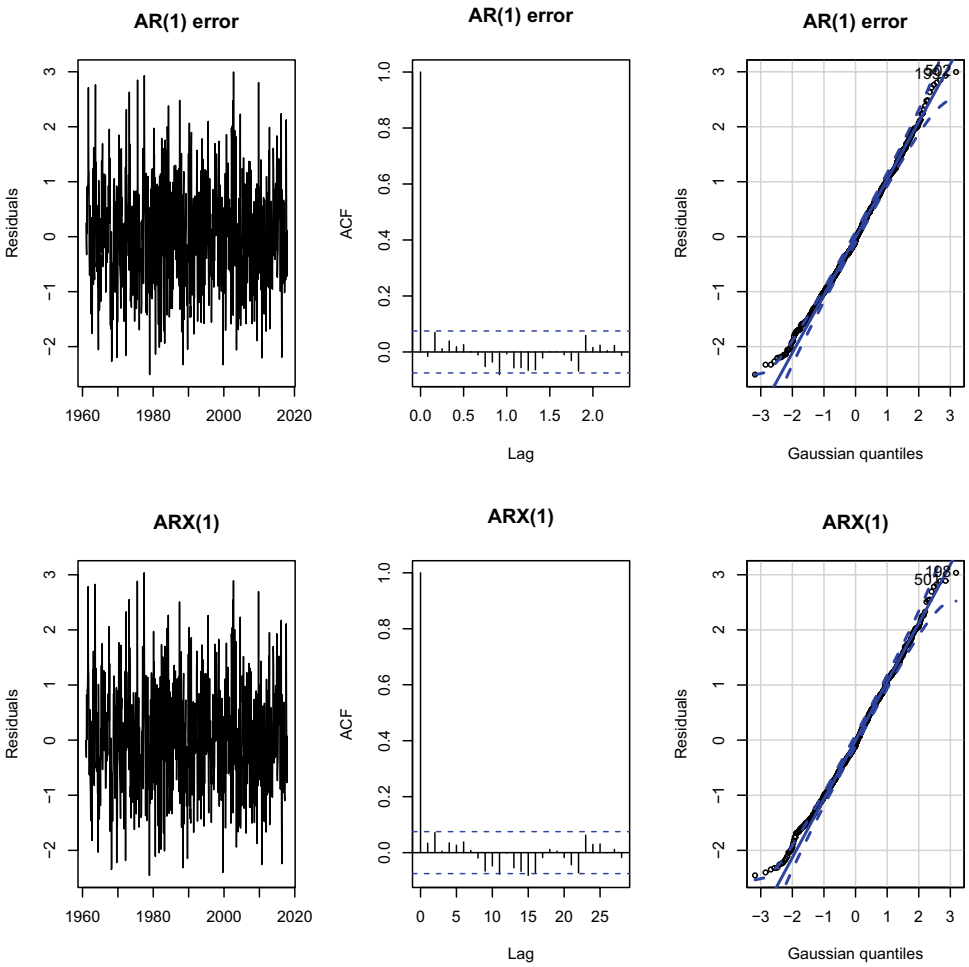
Both the regression model with AR(1) errors in (2) and the ARX(1) model in (3) provide similar estimates and predicted values.

For several practitioners, it is easier to include the lagged observation in the regression model, as in the ARX(1) model (3), since it can be estimated by the least squares method using any spreadsheet program. If the main goal is to calculate predictions or forecasts for the temperature, this model may be chosen or even a SARIMA model. However, considering the trend coefficient as the

Table 2 Estimates and corresponding standard errors, first residual autocorrelation, and estimated residual standard deviation

Effects	AR(1) Error model		ARX(1) model	
	Estimate	Std. Error	Estimate	Std. Error
Intercept	18.685	0.098	15.475	0.681
Trend	0.038	0.003	0.032	0.003
$cos(2\pi t/12)$	2.505	0.066	2.283	0.072
$cos(4\pi t/12)$	−0.583	0.060	−0.518	0.057
$cos(6\pi t/12)$	0.105	0.032	0.124	0.039
$sin(2\pi t/12)$	1.766	0.066	1.288	0.115
ar(1)	0.214	0.037	0.172	0.036
1 year	0.038	0.003	0.038	0.004
50 year	1.910	0.148	1.912	0.220
1st Autocorrel	−0.015		0.034	
Res. SD	1.027		1.021	

Fig. 2 Residual analysis



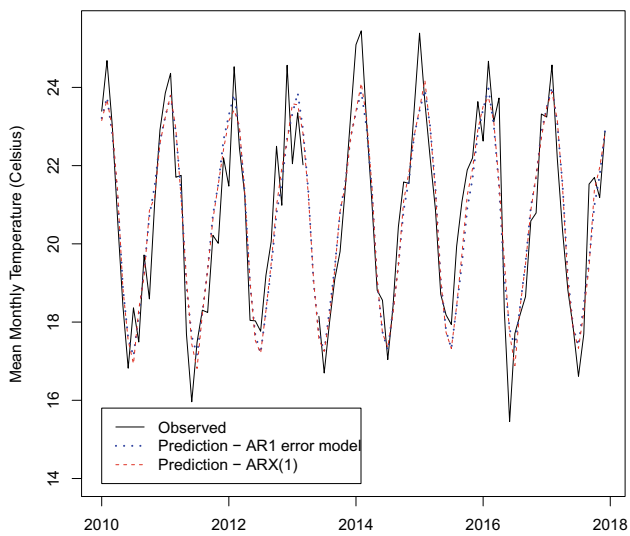


Fig. 3 Observed and predicted temperatures from 2010 to 2017 in São Paulo City

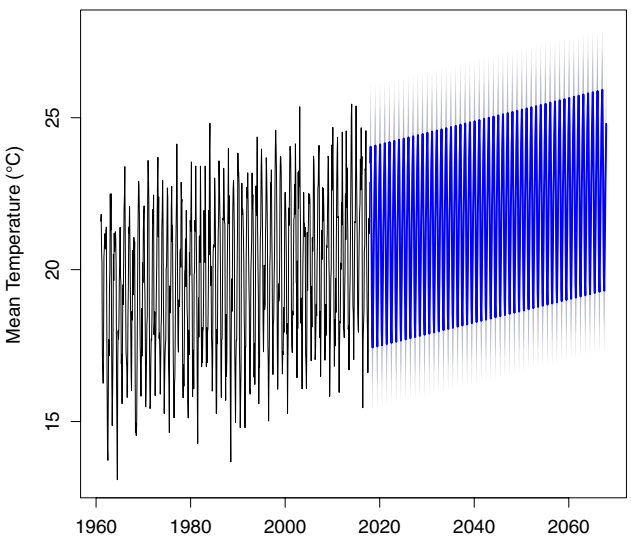


Fig. 4 Observed temperatures in São Paulo City from 1960 to 2017 and forecasts with 95% prediction intervals for the next 50 years using the AR(1) error model

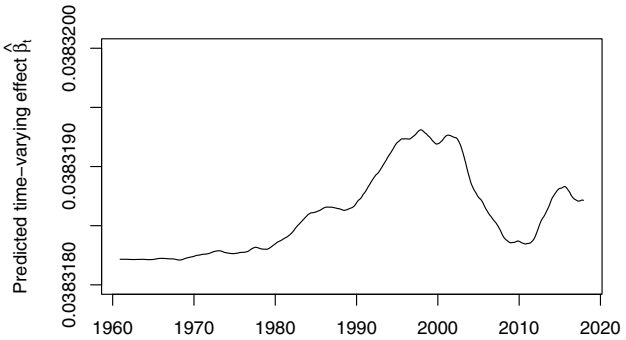


Fig. 5 Predicted time-varying trend effect for the dynamic linear model from 1960 to 2018

mean increase in temperature is not adequate since this would underestimate the trend. In Table 2, the trend parameter estimate is 0.032, but the mean increase estimator must be calculated using (4), providing an annual increase equals to 0.038 with corresponding asymptotic variance in (5).

On the other hand, to estimate and obtain confidence intervals or tests involving the model parameters, for example, for the secular trend, it is easier to consider the model with AR(1) errors. This occurs because its trend parameter directly measures the trend effect, and it is easier to obtain standard errors and confidence intervals.

It is worth noting that usually it is expected that the range of the forecast interval grows for larger horizons. However, the forecast interval are not getting wider in Fig. 4 and this may occur due to the very small variance of the trend estimator.

In [6], a regression model with trend and autoregressive error is fitted, but there is no seasonal component. This may have increased the estimated autoregressive coefficient, turning the error process similar to a random walk (unit root process). After considering differences, they also found significant temperature increases in Alaska, but the estimated parameters do not correspond to the mean annual increases presented here, which is easier to interpret.

A dynamic linear model to estimate a time-varying trend effect was also fitted indicating that this effect does not vary over time and the temperature is always increasing with the same rate, as also concluded by [5] for the global temperature data until 1980.

Recent papers concluded that global warming was overestimated. For example, [4] indicated that, in the period 1993–2012, the mean global temperature increased 0.14 ℃(∓0.06℃) per decade. This increase is lower than the previously estimated 0.3 °C using simulations of complex models. They concluded that maybe the models did not reproduce the observed global warming over the past 20 years or the slowdown in global warming over the past fifteen years.

Also, the mean increase in the global temperature is 0.18° per decade since 1981 in the Global climate report [15] and, for example, it is 0.2° in Ghana [12]. All these increases are smaller than the estimated increase around 0.4 °C per decade in São Paulo City, as observed in Table 1 and estimated by the proposed models (0.38 °C in Table 2). Our data from São Paulo showed a constant increase in temperature, with no slowdown.

5 Conclusions

Whenever there is evidence that the temperature increase is not changing over time, it is recommended to fit a regression model including trend and seasonal components and autoregressive errors to estimate the mean temperature increase with the corresponding confidence interval. If a model with lagged temperatures is chosen, it is necessary to calculate the long-term mean increase as in (4) and its variance as in (5), because the estimator of the trend coefficient is a biased estimator of the mean annual increase in the temperature.

Focusing on the climate change issue, there is sufficient evidence that temperature is increasing and there is an increase of 1.9 °C in the last 50 years in the city of São Paulo, with a 95% confidence interval of [1.6; 2.2] for the AR(1) error model, including the upper limit, stated in the Paris Agreement [14] of 2 ℃-increase. Maintaining this increase, it is expected another increase of around 1.9 °C in the next 50 years.

Acknowledgements I also thank all the corrections and suggestions pointed out by the referees and I thank Francisco Marcelo M. Rocha for some corrections and Cristine Oliveira for English corrections. I would like to acknowledge the financial support from FAPESP (2018/04654-9), Brazil.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alpuim T, El-Shaarawi A (2009) Modeling monthly temperature data in Lisbon and Prague. Environmetrics 20:835–852
2. Box G, Jenkins G (1976) Time series analysis: forecasting and control. Holden-Day, San Francisco
3. Box GEP, Jenkins GM, Reinsel GC (2013) Time series analysis. Wiley, New York
4. Fyfe JC, Gillett NP, Zwiers FW (2013) Overestimated global warming over the past 20 years. Nat Clim Change 3:767–769

5. Galbraith JW, Green C (1992) Inference about trends in global temperature data. Clim Change 22(3):209–221
6. Gil-Alana LA (2012) Long memory, seasonality and time trends in the average monthly temperatures in Alaska. Theor Appl Climatol 108:385–396
7. Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
8. Hyndman R, Athanasopoulos G et al (2018) Forecast: forecasting functions for time series and linear models, R package version 8.4
9. Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice, 2nd edn. OText, Melbourne
10. IBGE. IBGE releases population estimates for municipalities in 2016, August 2016
11. INPE. Instituto nacional e pesquisas espaciais (2019). http://bancodedados.cptec.inpe.br/downloadBDM/
12. Klutse NAB, Owusu K, Boafo YA (2020) Projected temperature increases over northern Ghana. SN Appl Sci 2(8):1–14
13. NASA. Nasa, NOAA data show 2016 warmest year on record globally, January 2017
14. United Nations/Framework Convention on Climate Change (2015) Adoption of the Paris Agreement, 21st Conference of the Parties, Paris: United Nations
15. Oceanic NN, Administration A (2019) Global climate report—annual 2020
16. Petris G, Petrone S, Campagnoli P (2009) Dynamic linear models with R. Springer, New York
17. R Core Team (2013) R: a Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
18. Sen PK, Singer JM, Lima ACP (2010) From finite sample to asymptotic statistics. Cambridge University Press, Cambridge
19. Shumway RH, Stoffer DS (2016) Time series analysis and its applications with R examples. Springer, New York