Universidade de São Paulo
Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Departamento de Astronomia

Rafael Cesario de Abreu

Temperature variability and trends in Southeastern Brazil : From local to regional scale analysis

Variabilidade e tendências de temperatura no Sudeste do Brasil: Uma análise em escalas local e regional

São Paulo
2023

Rafael Cesario de Abreu

Temperature variability and trends in Southeastern Brazil : From local to regional scale analysis

Variabilidade e tendências de temperatura no Sudeste do Brasil: Uma análise em escalas local e regional

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Sciences in the Institute of Astronomy, Geophysics and Atmospheric Sciences.

Major field: Meteorology
Advisor: Humberto Ribeiro da Rocha
Co-advisor: Ricardo Hallak

Final version. The original copy is available in the library.

São Paulo

2023

dedicated to my family

Acknowledgements

I would first like to offer my sincere gratitude to my advisor, Humberto Ribeiro da Rocha and co-advisor, Ricardo Hallak, for the guidance during this period, friendship, the opportunity to develop a thesis as a team effort, and also the opportunities to collaborate with different people from Brazil and internationally. From this experience, I believe I have become a better scientist with more experience in the subject and more accurate critical thinking.

My family is an essential part of this process as well. In a country that struggles with education, they gave me the foundation to be the person I am today. All my contributions to the scientific community today and in the future are also a contribution from Imara, Mario Marcio, and Daniel.

I also acknowledge the significant contribution of Simon Tett from the University of Edinburgh, that helped me write part of my thesis during my short stay in Scotland, and for all collaborations that succeeded. One of those collaborations was Sarah Sparrow from the University of Oxford, that was kind enough to trust me to help tutor two workshops during this time. Those were life-changing experiences for me and gave me a new perspective on why we do this type of research and what type of scientist I want to become.

The support from my friends was also fundamental in helping me reach the end of this journey. Thanks to Yann, Matheus, Takao, Leo, Natalia, Luan, Rafaela and many others for sticking with me through all these years. I sincerely appreciate the effort.

I could only do this research with CAPES's support, which provided me with a scholarship during the thesis development. Still, I should acknowledge the cooperation from people in INMET, IAC, and ICEA that helped me with data acquisition and answering questions about data and metadata. Also, I truly appreciate all the workers and infras-

tructure from IAG, which gave me the fundamental blocks to develop my thesis.

"*The cake is a lie*"

Doug Rattmann (Portal)

Resumo

As projeções do aquecimento global e do aumento de eventos climáticos extremos (Painel Intergovernamental Sobre Mudanças Climáticas (IPCC) são especialmente importantes para as áreas mais populosas, que penalizarão os grupos mais vulneráveis. O Sudeste do Brasil é um exemplo, que contribui com mais de 50 % do Produto Interno Bruto nacional, e abriga mais de 40 % da população. No entanto, as tendências históricas de temperatura não são homogêneas ao redor do globo, e particularmente na região Sudeste, devido a efeitos de distribuição de aerossóis, cobertura de superfície e circulações locais. Este estudo visa atribuir os principais fatores que contribuem para a determinação da variabilidade espacial e das tendências de temperatura no Sudeste. Com o modelo estatístico de Ribes et al. (2017) estimou-se um aumento de 1.1 ^{o}C em 50 anos na temperatura média regional observada, que não pode ser explicado sem o aumento de gases de efeito estufa, e que mais de 50 % da incerteza na estimativa dos parâmetros ajustados vem da variabilidade dos modelos do Coupled Model Intercomparison Project (CMIP5). Na escala local, utilizando 52 estações meteorológicas no Sudeste, com um modelo estatístico Aditivo Generalizado (GAM), estimou-se a variabilidade espacial da média de temperautra mínima e máxima (Tmin e Tmax), respectivamente, que foram significativamente controladas pelos fatores de zonalidade e continentalidade (posição geográfica) e altitude ($\simeq$ 5.0 ^{o}C), pela cobertura de superfície segundo o Índice de Vegetação de Diferença Normalizada (NDVI) para Tmin ($\simeq$ 3.0 ^{o}C) e pela cobertura de nuvens para Tmax ($\simeq$ 3.5 ^{o}C). Ainda, a análise do controle do NDVI sugere uma resposta heterogênea de Tmin que deve levar em conta a distribuição da cobertura vegetal mais localizada. A variabilidade temporal de temperatura de longo prazo em cinco estações meteorológicas no estado de SP mostrou que a tendência ajustada pelo GAM traz informações mais acuradas da variabilidade se comparada com o

ajuste linear, revelando uma provável influência do efeito de urbanização nas tendências da temperatura mínima, coerente com o crescimento da população nas cidades estudadas.

Palavras-chave: tendências de temperatura do ar, modelo aditivo generalizado, atribuição de mudança climática, modelos estatísticos, mudança de uso da terra.

Abstract

The projected global warming and increase in extreme events reported by the Intergovernmental Panel on Climate Change (IPCC) are important concerns for the most populated areas, that will impact mainly the most vulnerable. Southeast Brazil (SEB), for example, contributes more than 50 % of the national Gross Domestic Product (GDP) and houses more than 40 % of the country's population. However, these trends are not homogeneous throughout the globe, because of internal variability, aerosol distribution, and changes in land use, for example. Therefore in this study, we aim to we attribute the main contributors to temperature spatial variability, and trends in Southeast Brazil. Using Ribes et al. (2017) statistical model, we have found a 1.1 ^{o}C increase in the regional average temperature in 50 years, that can not be explained without the increase in anthropogenic greenhouse gases, and that more than 50 % of the uncertainty in the parameters estimation of the statistical model comes from the climate model variability, that comes from Coupled Model Intercomparison Project (CMIP5). At a local level, using 52 weather stations in SEB, with the Generalized Additive Model (GAM), we have estimated the average minimum and maximum temperature spatial variability for minimum and maximum temperature (Tmin and Tmax, respectively), are mostly influenced by changes in geographical position and altitude ($\simeq$ 5.0 ^{o}C), with contribution from land cover with the Normalized Difference Vegetation Index (NDVI) for Tmin ($\simeq$ 3.0 ^{o}C) and cloud cover for Tmax ($\simeq$ 3.5 ^{o}C). Also, NDVI analysis suggests a heterogeneous response to Tmin that needs to account for regional and local vegetation. The long-range temperature variability in five selected weather stations in São Paulo show that the estimated GAM trend gives more accurate information about the variability of temperature anomalies, compared to the linear fit, revealing a probable influence of urbanization in minimum temperature

trends, which agrees with the changes in population in the selected sites.

Keywords: Air temperature trends, generalized additive models, climate change attribution, statistical models, land cover change.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Southeast Brazil is an area of great economic importance, contributing more than 50 % of the country's Gross Domestic Product (GDP) (IBGE, 2018b). It has various economic activities like agriculture, industry, and services, in addition to important water reservoirs used for hydric and energy supply. It is an area that is home to 40 % of the country's population (IBGE, 2018a), with large migration of people from other regions of the country (IBGE, 2012). It has a complex topography (Figure 1.1), with emphasis on Serra do Mar on the coast of the states of Paraná, São Paulo, and Rio de Janeiro, where it reaches elevations of 1,000 meters in a few tens of kilometers from the coastline. Also, Serra da Mantiqueira and Serra do Espinhaço reach up to 1,500 meters above sea level.

The climate in Southeast Brazil shows significant variability, with higher accumulated precipitation in the south portion and lower in the north, marked by hot and wet summers and cold and dry winters (Reboita et al., 2010). Summer is characterized by the manifestation of the South Atlantic Convergence Zone (SACZ), responsible for a significant part of the precipitation (Ambrizzi and Ferraz, 2015), while in winter, cold fronts are the most frequent cause of rainfall in the region, especially in the south (Foss et al., 2017). The South Atlantic Subtropical High-Pressure system also has an important role in the climate of the area, especially in winter when it is closest to the coast of Brazil, causing persistence of days with low cloud cover and increasing air temperature near the surface (Sun et al., 2017). Also, local scale effects should be mentioned, such as continentality with an important influence from the Atlantic Ocean sea surface temperature in the regimes of temperature and precipitation near the coast, caused by an atmospheric pattern of circulation known as sea breeze (Oliveira et al., 2003).

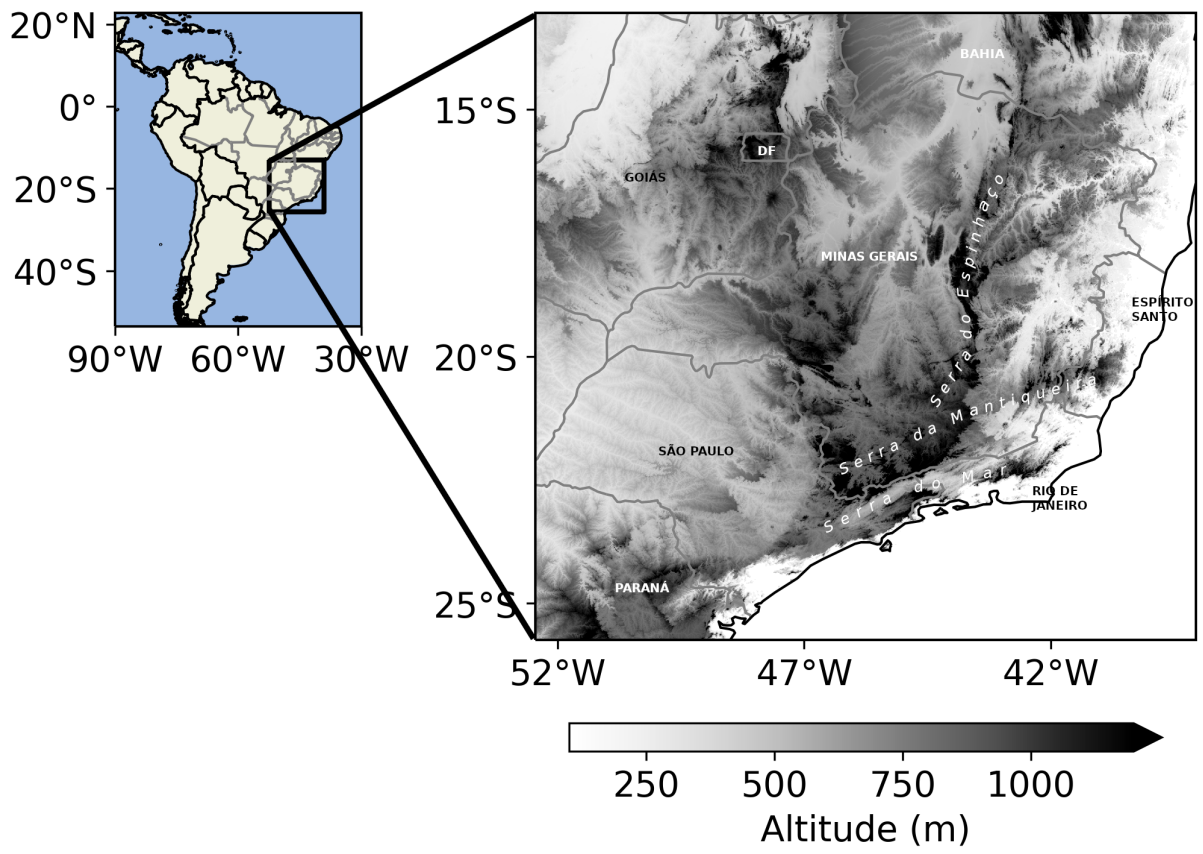

Figure 1.1: Map of Southeast Brazil highlighted by the black rectangle limited by the coordinates of 53.4^{o}W, 26.5^{o}S and 39^{o}W, 12.7^{o}S. The states are highlighted as well as the area of Serra do Mar, Serra da Mantiqueira, and Serra do Espinhaço, important regions of complex topography. Shaded is the altitude above sea level in meters.

Given the economic importance of the region, with activities that depend directly on the climate, as well as a large populational density, with a high number of vulnerable people, the impacts of anthropogenic climate change should be considered. The projections from the Sixth Assessment Report from the Intergovernmental Panel on Climate Change (IPCC) suggest an increase in global temperature between 2.7 e 5.7 ^{o}C until the end of the twentieth-first century, compared to the period of 1995 to 2014 for the highest emission scenario (Lee et al., 2021), as well as an increase in the frequency of events with extreme precipitation (Li et al., 2021). Currently, some of those events have already being observed, like drought in the state of São Paulo in 2014/2015, which was the lowest precipitation anomaly on record, based on measurements that started in 1961 (Coelho et al., 2015), but there is no evidence of being influenced by the increase in greenhouse gases (Otto et al., 2015). However, extreme precipitation events like the one that occurred in 2020 in Minas Gerais became 70 % more likely due to anthropogenic activities, with a loss of 56 lives and

millions of dollars (Dalagnol et al., 2022).

Despite the increase in global temperature due to the increase in greenhouse gases, there is regional and local variability in the observed temperature trends due to internal variability, aerosol distribution, and changes in land use. The last one is difficult to estimate, with Lott et al. (2020) suggesting a contribution that decreases the global temperature by -0.06 ^{o}C 10 yr.$^{-1}$ due to the exposed surface in the poles, that reflects the incoming solar radiation. In China, Sun et al. (2016) show that urbanization contributed to a third of the 1.44 ^{o}C increase in temperature between 1961 e 2013. At the local scale, Sugahara et al. (2012), for example, suggests a heating effect of 0.16 ^{o}C 10 yr.$^{-1}$ e 0.17 ^{o}C 10 yr.$^{-1}$ for maximum and minimum temperature, respectively, for the city of São Paulo in Brazil, due to urbanization since the 1960s.

1.1 Objectives

From the context given above, this study aims to attribute the main contributors to temperature spatial variability, and trends in Southeast Brazil over recent years, and the main questions to be answered are:

- Does the increase in greenhouse gas concentration contribute to the observed temperature trends in Southeast Brazil between 1955 and 2004?
- What are the main geographical controls of spatial variability of mean temperature in Southeastern Brazil?
- Does land cover change influence the variability of the observed trends in local temperature?

As specific objectives we highlight:

- Use a novel methodology for detection and attribution of temperature trends;
- Separate the main sources of uncertainty in the attribution of trends;
- Calculate the influence of vegetation on the spatial variability of the average temperature in Southeast Brazil

- Verify the difference from linear fit to estimate temperature trends and other non-linear methods

The document is divided into four chapters: Chapter 2 is a study of Detection & Attribution of temperature trends in Southeast Brazil using the methodology proposed by Ribes et al. (2017). Chapter 3 used a non-linear additive model to determine the contribution of different factors to the spatial variability of the average temperature in Southeast Brazil. Chapter 4 evaluates temperature trends using linear and non-linear methods for local weather stations and possible sources. Finally, Chapter 5 brings conclusions of the study based on the previous chapters and also gives suggestions for future studies.

Chapter 2

Attribution of detected trends in Southeast Brazil

The results that are presented in this chapter were published in **Geophysical Research Letters** (de Abreu et al., 2019).

2.1 Introduction

The Detection and Attribution (D&A) problem consists in demonstrating that observed trends are significant and different, in a statistical sense, from what can be explained by internal variability, which are caused by the interaction of low and high frequency climate components capable of producing long time-scale variations causing the impression of an identifiable trend, without any external forcings. After the identification of the trend, the causal attribution to different types of forcings can be made, which could be from natural sources like volcanic aerosols or changes in incoming solar radiation, or anthropogenic sources like increase in greenhouse gases, aerosols, or changes in land use (Mitchell et al., 2001).

Experimentation with introducing different kinds of forcings in the climate system and comparing them is not feasible, so we rely on model simulations and statistical analysis for the attribution task. For example, Stott et al. (2000) uses simulations with either natural or anthropogenic forcings to attribute the increase in global temperature. The standard approach uses linear regression models called optimal fingerprint (Hegerl et al., 1996; Allen and Tett, 1999; Allen and Stott, 2003) where scaling factors on simulated signals are estimated with a range of uncertainty. The magnitude of the scaling factor and its confidence interval is then used to make inferences about the causation of a particular forcing or a set of forcings that are statistically significant to explain the observed changes.

Using the methodology indicated above, Bindoff et al. (2013) shows attributable global warming due to anthropogenic forcings between 0.6 and 0.8 ^{o}C between 1951 to 2010 from the global temperature, which is consistent with the observed trends. In the sub-national scale, we could only find three recent studies (Wang et al., 2017; Wan et al., 2019; Wang et al., 2018). These studies found a human influence in temperature trends in Western China, regional Canadian change, and in extreme temperature indices in 17 subcontinent regions worldwide. The study of Karoly and Stott (2006) also detected a human influence on Central England temperature.

In Brazil, there are no studies focusing on the attribution of long-term observed trends, only attribution of weather events, which compare the probabilities of occurring a selected event in the actual world with a natural world, without human influence (Otto et al., 2015; Abreu et al., 2018). Studies have shown an increase of more than 3 ^{o}C in the city of São Paulo between 1940 and 2010 (Silva Dias et al., 2013). Also, the frequency of warm nights (minimum temperature above 90 % percentile) have increased while cold nights (minimum temperature below 10 % percentile) have decreased with statistically significant trends over Southeast Brazil (Vincent et al., 2005). Other observational studies have been made in South America and on individual cities in the state of São Paulo, to analyze whether a change in temperature could be detected and if these were due to increasing greenhouse gases and other forcings (Marengo, 2001; Blain et al., 2009). Although these studies find statistically significant trends in observations, the authors suggest these changes could be either due to increase in greenhouse gases from climate change or local factors like urbanization and land use changes from agricultural production, that could have a significant impact on observed trends.

Therefore, this study aims to answer the question of whether observed temperature changes in Southeast Brazil can be attributed to human and natural forcings. Southeast Brazil is the geopolitical region in Brazil that comprises the states of São Paulo, Rio de Janeiro, Minas Gerais and Espírito Santo and is responsible for more than 50 % of Brazil Gross Domestic Product (GDP) with a broad range of economic activities that includes agriculture, mineral extraction, automobile industries and others (IBGE, 2018b). More than 40 % of Brazil's population live in this region and it contains two of the most important cities of the country, São Paulo and Rio de Janeiro (IBGE, 2018a). The high exposure makes the region vulnerable to changes in climate, like droughts in major cities (Coelho

et al., 2015) and impacts on agricultural production due to an increase of temperature (Marengo, 2001; Camargo, 2010).

2.2 Material and Methods

2.2.1 Attribution model

In this study we use the attribution model presented in Ribes et al. (2017), hereafter referenced as R17. We changed some of the notation from R17 to keep the notation consistent throughout the thesis. Let's assume that the true observed climate response for a particular variable, like temperature, can be expressed as a vector $\mathbf{y}^*$ with n time steps that is the sum of the $i = 1, \ldots, n_f$ true responses from the individual forcings $\mathbf{x}_i^*$ which are also vectors with n time steps:

$$\mathbf{y}^* = \sum_{i=1}^{n_f} \mathbf{x}_i^* \tag{2.1}$$

$$\mathbf{y} = \mathbf{y}^* + \boldsymbol{\epsilon}_\mathbf{y} \tag{2.2}$$

$$\mathbf{x}_i = \mathbf{x}_i^* + \boldsymbol{\epsilon}_{\mathbf{x}_i} \tag{2.3}$$

This means that the observed response $\mathbf{y}$ is a sum of the true response with a random noise term ($\boldsymbol{\epsilon}_\mathbf{y} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{y})$), which arises from internal variability and observational error, while the observed response from the individual forcings $\mathbf{x}_i$ are a sum of the true response for that particular forcing and a random noise term ($\boldsymbol{\epsilon}_{\mathbf{x}_i} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{x}_i})$) that is a composition of internal variability and model error.

Maximum likelihood estimation is then used to obtain estimates for $\mathbf{y}^*$ and $\mathbf{x}_i^*$ which gives a -2 log-likelihood function of the form:

$$l(\mathbf{x}_1^*, \ldots, \mathbf{x}_{n_f}^*) = \left(\mathbf{y} - \sum_{i=1}^{n_f} \mathbf{x}_i^*\right)^T \boldsymbol{\Sigma}_\mathbf{y}^{-1} \left(\mathbf{y} - \sum_{i=1}^{n_f} \mathbf{x}_i^*\right) + \sum_{i=1}^{n_f} (\mathbf{x}_i - \mathbf{x}_i^*)^T \boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1} (\mathbf{x}_i - \mathbf{x}_i^*) + \text{cte.} \tag{2.4}$$

That, when maximized, can be used to find the estimates $\hat{\mathbf{y}}^*$ and $\hat{\mathbf{x}}_i^*$:

$$\hat{\mathbf{y}}^* = \mathbf{y} + \boldsymbol{\Sigma}_\mathbf{y}(\boldsymbol{\Sigma}_\mathbf{y} + \boldsymbol{\Sigma}_\mathbf{x})^{-1}(\mathbf{x} - \mathbf{y}) \tag{2.5}$$

$$\hat{\mathbf{x}}_i^* = \mathbf{x}_i + \boldsymbol{\Sigma}_{\mathbf{x}_i}(\boldsymbol{\Sigma}_\mathbf{y} + \boldsymbol{\Sigma}_\mathbf{x})^{-1}(\mathbf{y} - \mathbf{x}) \tag{2.6}$$

Where $\mathbf{\Sigma_x} = \sum_{i=1}^{n_f} \mathbf{\Sigma}_{\mathbf{x}_i}$ and $\mathbf{x} = \sum_{i=1}^{n_f} \mathbf{x}_i$. $\hat{\mathbf{y}}^*$ and $\hat{\mathbf{x}}_i^*$ are unbiased estimators for $\mathbf{y}^*$ and $\mathbf{x}_i^*$ with the following distributions:

$$\hat{\mathbf{y}}^* \sim N\left(\mathbf{y}^*, (\mathbf{\Sigma}_\mathbf{y}^{-1} + \mathbf{\Sigma}_\mathbf{x}^{-1})^{-1}\right) \tag{2.7}$$

$$\hat{\mathbf{x}}_i^* \sim N\left(\mathbf{x}_i^*, \left(\mathbf{\Sigma}_{\mathbf{x}_i}^{-1} + \left(\mathbf{\Sigma_y} + \sum_{j\neq i} \mathbf{\Sigma}_{\mathbf{x}_j}\right)^{-1}\right)^{-1}\right) \tag{2.8}$$

The estimates $\hat{\mathbf{y}}^*$ and $\hat{\mathbf{x}}_i^*$ can be used to calculate the true influence of each individual forcing. Also, a series of statistical tests can be used to detect their contribution. For example, the null hypothesis that $\mathbf{y}^* = \mathbf{0}$ means that the detected trend is due to internal variability only and can be tested using as a basis the likelihood ratio test where $\mathbf{y}^T\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{y} \overset{\mathrm{H_0}}{\sim} \chi_n^2$. The consistency test for any subset of individual forcings can also be tested using $\mathrm{H_0}$ as follows:

$$(\mathbf{y} - \mathbf{x}_I)^T(\mathbf{\Sigma_y} + \mathbf{\Sigma}_{\mathbf{x}_I})^{-1}(\mathbf{y} - \mathbf{x}_I) \overset{\mathrm{H_0}}{\sim} \chi_n^2 \tag{2.9}$$

Where I is a subset of forcings from 1 to n_f, $\mathbf{\Sigma}_{\mathbf{x}_I} = \sum_{i\in I} \mathbf{\Sigma}_{\mathbf{x}_i}$ and $\mathbf{x}_I = \sum_{i\in I} \mathbf{x}_i$. As in R17, Ordinary Least Square (OLS; Allen and Tett, 1999) is also used to compare the results found with the proposed methodology. The OLS method assumes a linear dependency between the simulated responses $\mathbf{x}_i$ and the observed one $\mathbf{y}$. A scaling factor $\boldsymbol{\beta}$ is then estimated by $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ which can be used for inference:

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^T\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{y} \tag{2.10}$$

Where $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_{n_f}}]$. The main differences between the two methods is that the proposed one does not include a scaling factor and explicitly include observational and model uncertainty. Therefore, $\mathbf{\Sigma_y}$ in OLS is calculated using only internal variability.

2.2.2 Estimation of the covariance matrices

One difficulty of the proposed method arises from the estimation of the covariance matrices $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_y}$ which are assumed to be known. The approach suggested in R17 is to use the "models are statistically indistinguishable from the truth"paradigm which, from a Bayesian perspective, where the true value is considered to be a non-deterministic

quantity that is part of a underlying distribution, assumes that the models and the truth are taken from the same distribution. Under this assumption the covariance matrices can be estimated by:

$$\mathbf{\Sigma_y} = \underbrace{\mathbf{\Sigma}_{obs}}_{\text{Observational error}} + \underbrace{\mathbf{\Sigma}_v,}_{\text{Internal variability}} \tag{2.11}$$

$$\mathbf{\Sigma}_{\mathbf{x}_i} = \underbrace{\left(1 + \frac{1}{n_m}\right)\hat{\mathbf{\Sigma}}_m}_{\text{Model error}} + \underbrace{\frac{1}{n_m^2}\sum_{j=1}^{n_m}\frac{\mathbf{\Sigma}_v}{n_j}}_{\text{Internal variability}} \tag{2.12}$$

Where n_j is the number of ensemble members for the jth model, n_m is the number of available models. The matrix $\mathbf{\Sigma}_v$ is the covariance matrix due to internal variability only, while $\mathbf{\Sigma}_{obs}$ is the observational error. The model error $\hat{\mathbf{\Sigma}}_m$ is given by:

$$\hat{\mathbf{\Sigma}}_m = \frac{1}{n_m - 1}\left(SSM - \frac{n_m - 1}{n_m}\sum_{j=1}^{n_m}\frac{\mathbf{\Sigma}_v}{n_j}\right) \tag{2.13}$$

Where $SSM = \sum_{j=1}^{n_m}(w_j - \overline{w})^2$ which is the squared difference between the jth model ensemble mean w_j and the multi model mean $\overline{w}$.

Another difficulty comes from the inversion of the covariance matrices, required to compute the desired quantities. For example, a natural candidate for the estimation of $\mathbf{\Sigma}_v$ is the sample covariance matrix $\hat{\mathbf{\Sigma}}_v = \mathbf{Z}\mathbf{Z}^T/m$, where $\mathbf{Z}$ is a $n \times m$ matrix with m vectors of "pseudo-observations"that comes from the control simulation, or in our case, from the within-ensemble difference, used to represent internal variability. However, in climate sciences the size of the covariance matrices is usually large and therefore, the number m of realizations needed to compute the covariance matrix accurately is also large. For example, if $n/m > 1$ the rank of $\hat{\mathbf{\Sigma}}_v$ is at most m and the matrix is therefore non invertible. However, even when $n/m \leq 1$ but the ratio is non negligible this lead to a numerically ill-conditioned matrix that when inverted amplifies the error substantially (Figure 2.1; Ledoit and Wolf, 2004). A usual approach is to use the Moore-Penrose pseudo-inverse, which implies the truncation of the matrix to the k leading Empirical Orthogonal Functions (EOFs) where $k << m$ focusing on the main patterns of variability (Allen and Tett, 1999; Hegerl et al., 1996). Another approach, which is used in this study, is to regularize the matrix:

$$\hat{\mathbf{\Sigma}}_{v_r} = \rho\mathbf{I}_n + \lambda\hat{\mathbf{\Sigma}}_v \tag{2.14}$$

The values ρ and λ are obtained using the method described in Ledoit and Wolf (2004) and used in Ribes et al. (2009) and Ribes et al. (2013). An example of the effect of the regularization is shown in Figure 2.1 for different ratios of n/m using the identity matrix as the true covariance matrix. As the number of realization m increases in respect to the number of features n, the eigenvalues of the sample covariance matrix converge to 1s which are the theoretical values. In Figure 2.1a the number of realizations is ten times higher than the number of features and there is still a spread of the estimated eigenvalues, with an overestimation of the highest eigenvalues and an underestimation of the smallest ones. This error is amplified by the inversion of the sample covariance matrix, which is more evident in Figure 2.1b when m is only double the number of features. With the regularization, even in cases when $n/m > 1$ the eigenvalues are closer to the theoretical ones.

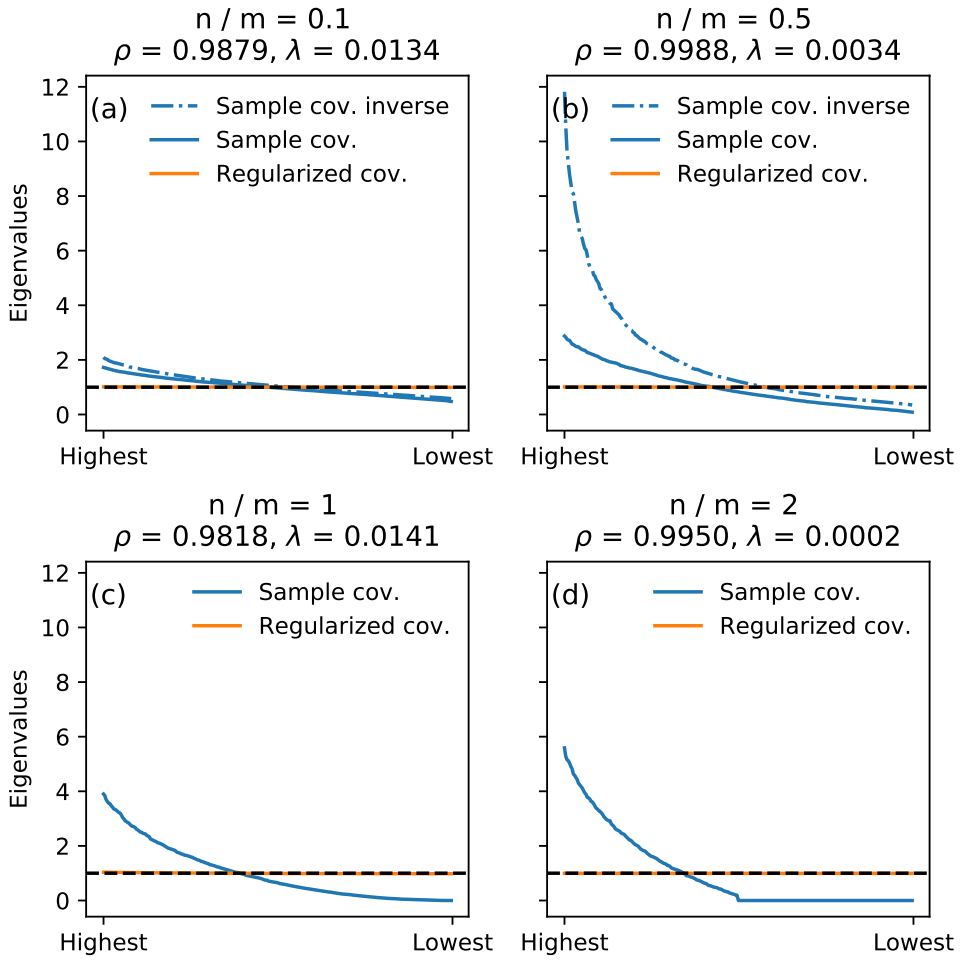

Figure 2.1: True eigenvalues (dashed black line) from a covariance matrix given by the identity matrix **I**, compared with the ones calculated from the sample covariance matrix and the regularized estimate for different ratios of n/m. The eigenvalues are ordered from the highest to the lowest and the estimated values for ρ and λ for the regularization from Eq. 2.14 are given in the title. The eigenvalues for the inverse of the sample covariance matrix are also shown whenever possible. Adapted from Ledoit and Wolf (2004).

2.2.3 Data and preprocessing

Gridded temperature observations from the Climatic Research Unit Temperature, version 4 dataset (CRUTEM4) are used in this study to estimate the observed trends in temperature for Southeast Brazil. This dataset uses homogenized weather stations, has been corrected for urbanization effects, and provides monthly anomalies on a $5^o \times 5^o$ latitude/ longitude grid from 1850 to present (Jones et al., 2012). The area selected for this study comprises all of the land in Southeast Brazil, bounded by 53.4^{o}W, 26.5^{o}S and 39^{o}W, 12.7^{o}S (Figure 2.2a).

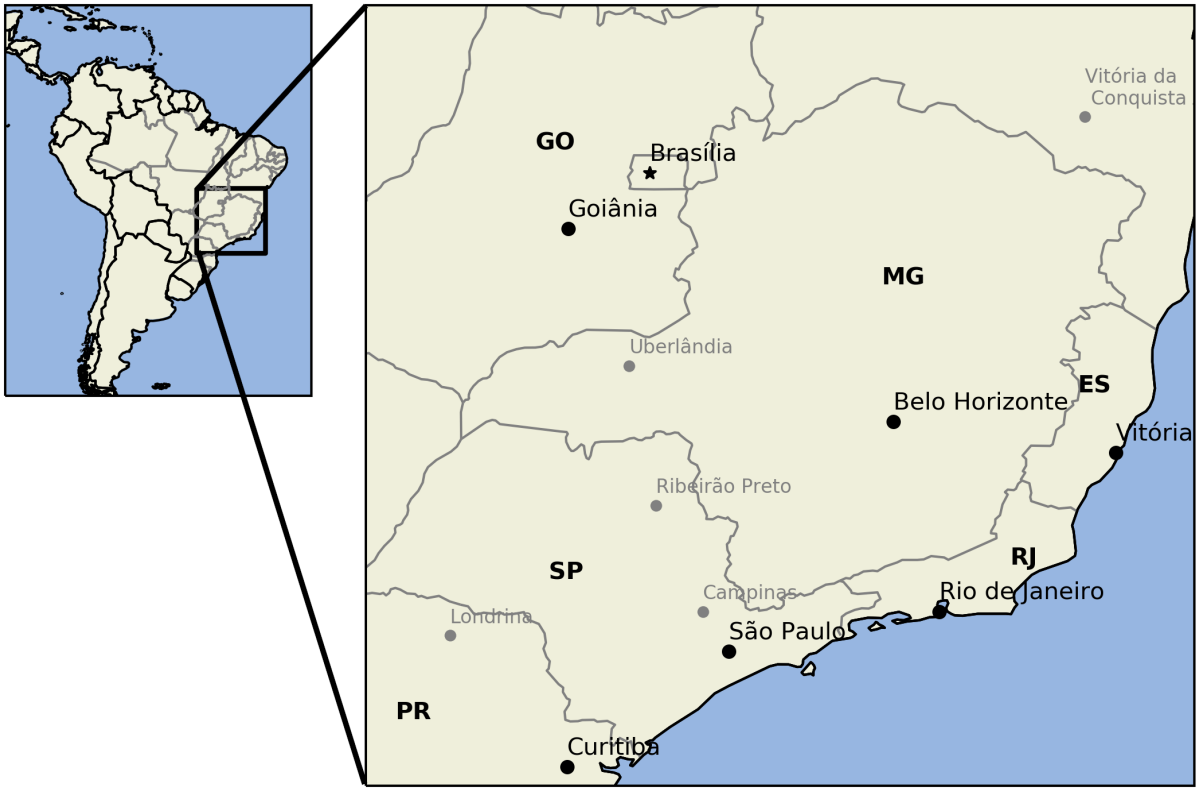

Figure 2.2: Region of interest comprising all states in Southeast Brazil highlighted by the black box bounded by 53.4^{o}W, 26.5^{o}S and 39^{o}W, 12.7^{o}S. The states and capitals of each states are highlighted in the figure (black points), as well as some populous regions (grey points).

The area-averaged anomalies were calculated using the procedure described in Morice et al. (2012), in which a weight is attributed to each grid cell that is proportional to its area. After that the annual averages are calculated. We focus on three distinct periods: 1955 to 2004, 1935 to 2004 and 1955 to 2014. The first period was selected because it is when the trend is most significant, observations are more reliable, and simulated signals for different forcings are available. The second period was selected as a sensitivity test with a longer period of data. The third period expands the analysis for ten years to include a clear signal of the detected trend using only the all forcings simulation, which includes both anthropogenic and natural forcings. For 2005 to 2014 simulations using Representative Concentration Pathway 8.5 (RCP8.5) from the models of the Coupled Model Intercomparison Project Phase 5 (CMIP5; Table A.1) were used. Ten-year averages were then computed to increase the signal to noise ratio and the temporal mean subtracted from the data in order to focus only on the anomalies following Ribes et al. (2013). Therefore, the size n of the vectors $\mathbf{x}_i$ and $\mathbf{y}$ are five for the 1955-2004 period, seven for 1935-2004

and six for 1955-2014.

In this study we use simulations from the Community Earth System Model (Hurrell et al., 2013) to understand which of the various uncertainties (internal variability, observational error and model error) are most important. We use 34 members from the large ensemble (CESM-LE) (Kay et al., 2015) driven with both natural and anthropogenic forcings (ALL), a 3-member ensemble with solar and volcanic forcings (NAT) and a 3-member ensemble driven only with greenhouse gases (GHG). Simulated data are interpolated to the CRUTEM4 $5^o \times 5^o$ grid and masked by the observational monthly mean dataset. The CMIP5 models in Table A.1 are also used to compute model error and the multi-model ensemble mean is also used to attribute changes in temperature due to one or a subset of the listed forcings.

We consider the effects of greenhouse gases (GHG), natural influences (solar and volcanic; NAT) and other anthropogenic forcings (OA, mostly aerosols and land use changes). Therefore, using the notation introduced in section 2.2.1, we have $\mathbf{x}_{\mathrm{GHG}}$, $\mathbf{x}_{\mathrm{NAT}}$ and $\mathbf{x}_{\mathrm{OA}}$ respectively, where the latter is calculated as: $\mathbf{x}_{\mathrm{OA}} = \mathbf{x}_{\mathrm{ALL}} - \mathbf{x}_{\mathrm{NAT}} - \mathbf{x}_{\mathrm{GHG}}$, where $\mathbf{x}_{\mathrm{ALL}}$ is the all forcings simulation. To calculate the covariance matrices $\mathbf{\Sigma_y}$, $\mathbf{\Sigma}_{\mathbf{x}_{\mathrm{GHG}}}$, $\mathbf{\Sigma}_{\mathbf{x}_{\mathrm{NAT}}}$ and $\mathbf{\Sigma}_{\mathbf{x}_{\mathrm{OA}}}$ the covariance matrix for internal variability ($\mathbf{\Sigma}_v$) is required. This is done by calculating the within-ensemble differences from the large CESM-LE ensemble. In order to be consistent with the OLS approach $\hat{\mathbf{\Sigma}}_v$ is split into two covariance matrices, one used to pre-whiten the data and other for uncertainty estimates in $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. This is achieved by splitting the members from the large ensemble into two subsets of 17 members (34 members form CESM-LE divided by two) and then calculating $\hat{\mathbf{\Sigma}}_{v_1}$ and $\hat{\mathbf{\Sigma}}_{v_2}$ using this subset of simulations, used in equations 2.11 and 2.12, respectively.

To determine which of the different errors is dominant we carry out three main analyses:

- We use only internal variability ($\mathbf{\Sigma}_v$) to compute $\mathbf{\Sigma_y}$ and $\mathbf{\Sigma}_{\mathbf{x}_i}$;
- We include observational error ($\hat{\mathbf{\Sigma}}_{\mathrm{obs}}$) from CRUTEM4 for the estimation of $\mathbf{\Sigma_y}$;
- We include model errors ($\hat{\mathbf{\Sigma}}_m$) for the estimation of $\mathbf{x}_{\mathrm{GHG}}$, $\mathbf{x}_{\mathrm{NAT}}$ and $\mathbf{x}_{\mathrm{OA}}$ and their respective covariance matrices. The covariance matrices for internal variability ($\hat{\mathbf{\Sigma}}_{v_1}$ and $\hat{\mathbf{\Sigma}}_{v_2}$) are calculated using the regularization approach described in section 2.2.2

In order to calculate the observational uncertainty $\hat{\mathbf{\Sigma}}_{\mathrm{obs}}$ we consider the correlated error ($\hat{\mathbf{\Sigma}}_{\mathrm{corr}}$) by using 100 ensemble members of the land only component of HadCRUT4 and the

uncorrelated errors ($\hat{\mathbf{\Sigma}}_{\text{uncorr}}$) from the same dataset. The ensemble members are generated based on the spatial and temporal uncertainties that are correlated (Morice et al., 2012). Therefore, $\hat{\mathbf{\Sigma}}_{\text{obs}} = \hat{\mathbf{\Sigma}}_{\text{corr}} + \hat{\mathbf{\Sigma}}_{\text{uncorr}}$. The model error covariance matrix ($\hat{\mathbf{\Sigma}}_m$) is conservatively estimated using the CMIP5 models that are indicated in Table A.1 using equation 2.13. In all cases we assume that the mean values for $\mathbf{x}_i^*$ comes from the CESM-LE ensemble, which means that $n_m = 1$ for calculating internal variability in equation 2.12. This is done also on the third step, when the CMIP5 models are included to calculate $\hat{\mathbf{\Sigma}}_m$, to make the analysis consistent with the previous steps. We carry out a final analysis where we estimate the $\mathbf{x}_i$ from the CMIP5 multi-model average ($n_m > 1$). Throughout this study, internal variability was computed from CESM-LE.

From the best estimates of the true signal ($\hat{\mathbf{y}}^*$ and $\hat{\mathbf{x}}_i^*$) calculated using R17 method, trends are estimated using linear regression. We used 1000 random samples with replacement generated from the covariance matrices from equations 2.7 and 2.8 to estimate the uncertainty from the estimated trends. The 5 % and 95 % percentiles are considered as the lower and upper threshold, respectively. For OLS, the model response is scaled by $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ to calculate the trend by linear regression and estimate the warming/cooling rate to be compared with R17 best estimates trends. We also show 5-95 % ranges for OLS.

2.3 Results and Discussion

The observed anomalies for Southeast Brazil from CRUTEM4 (Figure 2.3) shows a warming trend, from the decadal averages, of 0.22 [0.15 to 0.31] ^{o}C per decade between 1955 and 2004, which is equivalent to a 1.1 [0.7 to 1.5] ^{o}C over this period for the average temperature. The ALL simulation, which contains both natural and anthropogenic forcings, captures the observed warming for the period with a correlation of 0.60, which suggests that about 40 % of the observed inter-annual to multi-decadal variability is forced by natural and anthropogenic forcings. This happens because each ensemble member has its own internal variability that is not necessarily in phase with the observation, so when we calculate the average, the internal variability tends to cancel out, and the signal is mostly from the external forcings (natural + anthropogenic).

The GHG simulation has a similar trend compared to CRUTEM4 and ALL between 1955 and 2004, with a correlation of 0.39. This suggests much of the observed temperature

increase in Southeast Brazil could be due to greenhouse gases only. The NAT simulation suggests a slight cooling of about 0.25 ^{o}C in the early 1990s from the 1991 Pinatubo eruption, which is also apparent in ALL. The estimated OA signal, which is calculated from the difference of ALL to GHG and NAT, cools until about 1980 and warms after that, with a linear correlation of 0.25 with observations in the 1955-2004 time window. This result might be due to changes in emissions of sulphur dioxide from Europe and North America, which had rapidly increased starting at the beginning of the 20th century and then declined from the 1970s due to emission control policies (Hoesly et al., 2018).

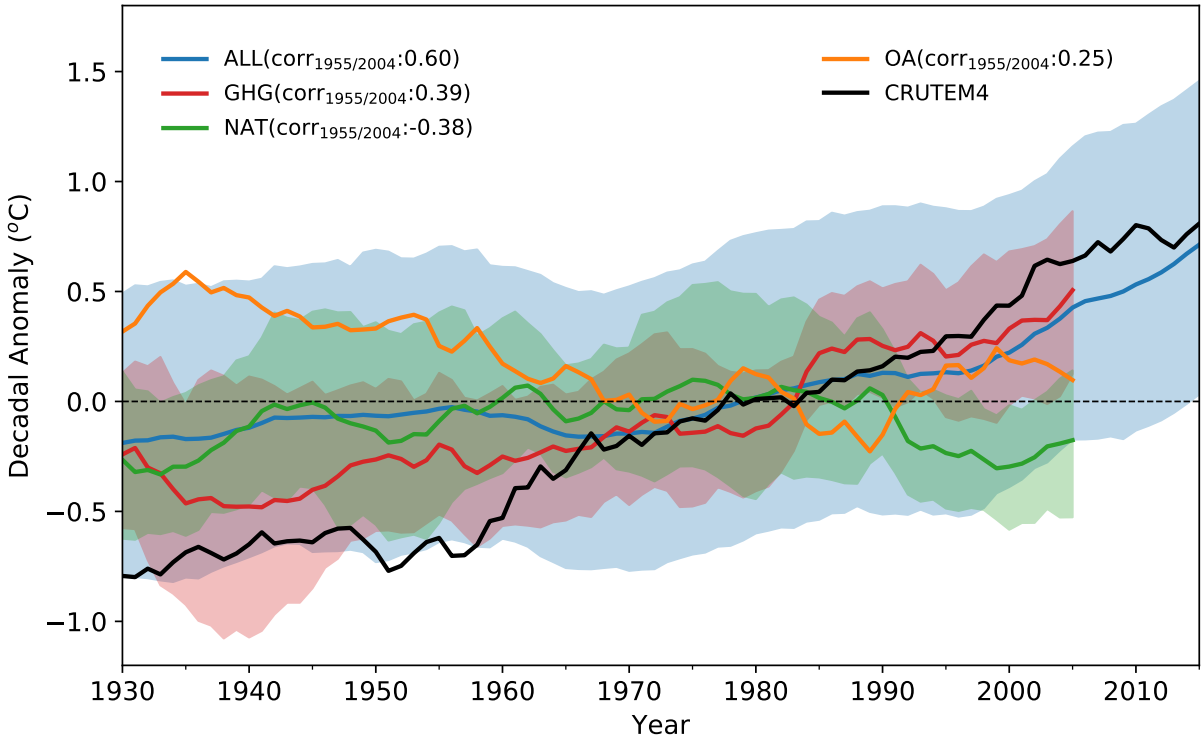

Figure 2.3: Ten-years moving average of annual temperature anomalies, between 1920 and 2017 for CRUTEM4 (black line), ALL (blue line), GHG (red line) and NAT (green line) simulations. Other Anthropogenic (OA; orange line) is ALL minus GHG and NAT ensemble means. Shading indicates the model spread (5 to 95 % range). Correlations between CRUTEM4 annual anomalies and 1955-2004 ensemble means are displayed in the labels. The anomalies are calculated with respect to 1961 to 1990 climatology.

First we calculate the best estimates of the OLS scaling factors ($\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$) which are shown in Figure 2.4. We can see that the observed signal is underestimated for the GHG signal, specially for the 1955-2004 period with a value of 1.81 [1.11 to 2.51], being approximately -44 % [$= 100 \times (1/\beta_{\mathrm{GHG}} - 1)$] lower than the observation. However, the signal is consistent with unit in the 1935-2014 period, which means that it is clearly detected by the method.

On the other hand, OA and NAT have large uncertainties, being consistent with zero, which means that their signal is undetected in the OLS estimation.

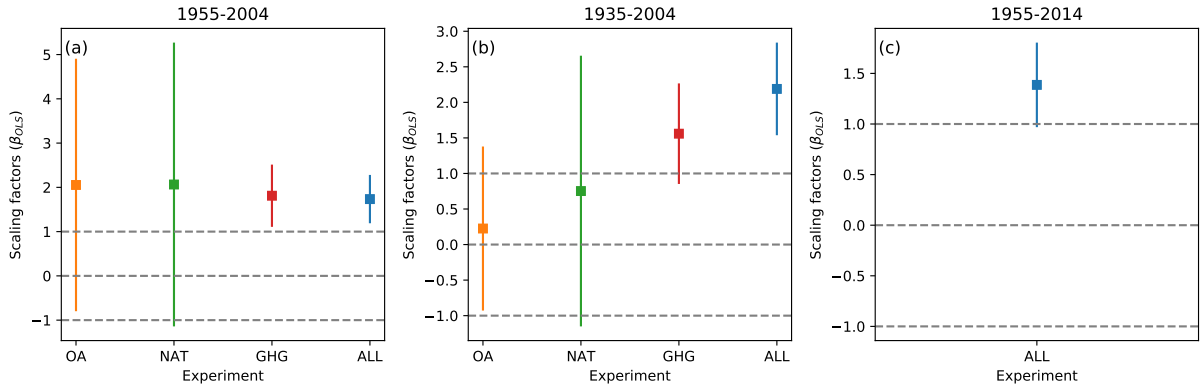

Figure 2.4: Best estimate of the OLS scaling factors and their 5-95 % confidence interval for other anthropogenic (OA, orange), natural (NAT, green), greenhouse gases (GHG, red) and all forcings (ALL, blue) for the periods between: (a) 1955-2004; (b) 1935-2004 and (c) 1955-2014.

The best estimates of $\hat{\mathbf{y}}^*$ and $\hat{\mathbf{x}}_i^*$ are calculated using the R17 statistical model, using CESM as reference in three steps to understand the importance of the different types of uncertainties, as described in section 2.2.3:

- Using only internal variability to estimate the error (R17 iv only);
- Including observational error (R17 iv + obs)
- Including model error (R17 iv + obs + model)

After that, the linear trends in temperature are computed using the raw model data and the R17 estimates using each of the different errors and they are compared with each other.

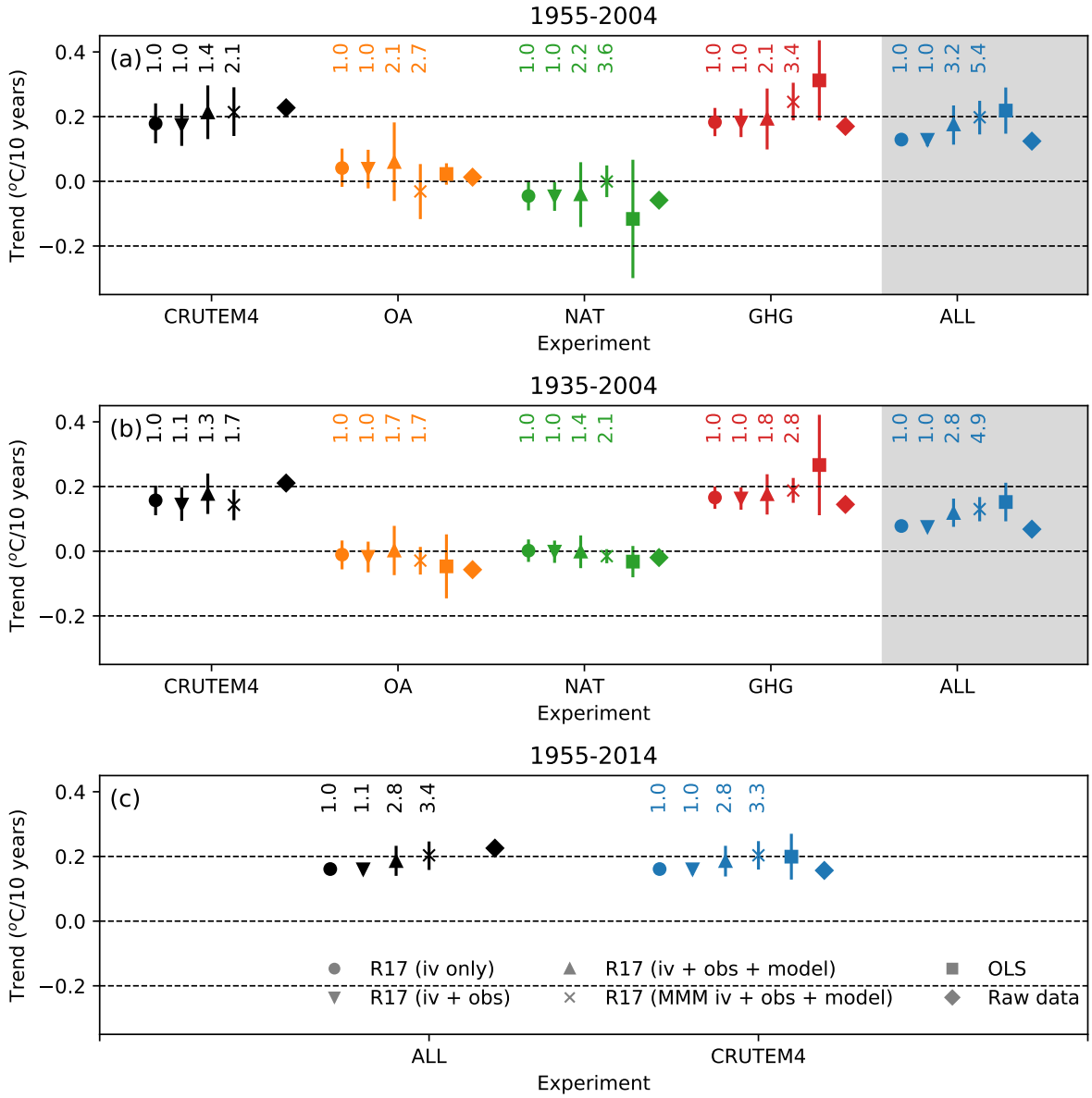

Figure 2.5: Temperature trends calculated from decadal averages for the observations (CRUTEM4) and each individual forcing (OA, NAT and GHG) using R17 three signal models best estimates ($\hat{\mathbf{x}}_i^*$ and $\hat{\mathbf{y}}^*$) for the different steps of analysis that include: (1) Internal variability only to estimate the covariance matrices (iv only, circle); (2) Inclusion of observational error (iv + obs, triangle down) and (3) inclusion of observational error and model error (iv + obs + model, triangle up). The estimated trend using the multi model ensemble mean (MMM iv + obs + model, x symbol) and the CESM/CRUTEM4 raw data (diamond) are also included (Raw data; $\mathbf{x}$ and $\mathbf{y}$ from R17 notation) as well as the OLS estimate after scaling by $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ (squares). The trends for ALL are based on R17 one signal model best estimates and is displayed in the shaded area in left of Figures (a) and (b). Figure (a) shows the trends between 1955 and 2004 and (b) for 1935 to 2004 (c) for 1955 to 2014 using RCP8.5 to extend the simulations after 2005. The numbers above the marker shows the ratio between the uncertainty relative to the best estimate ($\hat{\mathbf{x}}_i^*$ and $\hat{\mathbf{y}}^*$) of the iv only case calculated as in equation 2.12.

Our results, using the different types of error, consistently find a detectable impact of

greenhouse gases on Southeast Brazil temperature as seen by the estimated linear trends in Figure 2.5a and b. The inclusion of observational error (iv + obs) did not cause a significant increase in uncertainty, suggesting that observational error is not a relevant source of error for estimating the true response of the temperature trends. On the other hand, model error (iv + obs + model) is a major source of uncertainty for calculating R17 best estimates. When separating the errors that comes from iv + obs from the ones that come from model, we see that roughly 50 % of the total uncertainty comes from model error. As an example, the uncertainties in the CESM GHG forced trend for the CESM model are 2.1 times larger when including model uncertainty compared with the estimates when just internal variability is considered, for the 1955-2004 period.

In the case where all sources of error are used (iv + obs + model) the CESM OA signal indicates a small warming and large uncertainty -0.06 ^{o}C to 0.18 ^{o}C, mostly due to model error which is consistent with the observed warming according to the chi-squared test defined in Eq. 2.9 (Table 2.1). However, for the 1935 to 2004 time window OA makes no statistically significant contribution to the observed trends. Given this and the conservative estimates of model error for R17, we think that OA alone does not explain changes in Southeast Brazil

In the analysis that we mentioned previously we used CESM as the mean to define the forced signals and all CMIP5 models from Table A.1 to estimate $\hat{\mathbf{\Sigma}}_m$, which may not be ideal since we assume a Gaussian uncertainty centered around CESM which lies towards the tail of the CMIP5 model distribution, with a trend that corresponds to the tenth lowest trend from the 35 models available for the ALL experiment. Therefore we estimate the true response signals, including all uncertainties, using the CMIP5 multi-model ensemble mean (R17 MMM iv + obs + model). Between 1955-2004 (Figure 2.5a) we also find a significant contribution to the observed warming from GHG. A trend of 0.19 ^{o}C to 0.30 ^{o}C per decade is found that is equivalent to a 0.95 ^{o}C to 1.50 ^{o}C warming in this 50 year period. NAT and OA are small and have significant uncertainties which makes it difficult to draw any conclusion regarding the impact of those forcings for this time scale and for the study region. Contrary to using CESM OA signal, CMIP5 OA multi-model ensemble mean does not show consistency with the observed warming in the 1955 to 2004 time window, with a trend of -0.07 ^{o}C to 0.01 ^{o}C per decade.

Table 2.1 - Hypothesis testing χ^2 p-value for individual forcings from R17 model in cases where (1) Internal variability only was used to estimate the covariance matrices (iv only); (2) Inclusion of observational error (iv + obs); (3) inclusion of observational error and model error (iv + obs + model) and (4) considering the multi model mean as the ensemble mean instead of CESM large ensemble (MMM iv + obs + model). The results presented here are for the 1955-2004 and 1935-2004 time window.

Forcing/Error	iv only	iv + obs	iv + obs + model	MMM iv + obs + model
		1955-2004		
Internal-Variability	0.00	0.00	0.00	0.00
OA	0.01	0.02	0.36	0.03
NAT	0.00	0.00	0.02	0.01
GHG	0.66	0.70	0.84	0.98
All forcings (GHG+NAT+OA)	0.70	0.73	0.96	0.96
ALL	0.35	0.43	0.74	0.86
		1935-2004		
Internal-Variability	0.00	0.00	0.00	0.00
OA	0.00	0.00	0.00	0.00
NAT	0.00	0.00	0.01	0.00
GHG	0.73	0.78	0.92	0.95
All forcings (GHG+NAT+OA)	0.36	0.47	0.82	0.48
ALL	0.01	0.06	0.21	0.36

Changing the time window of the analysis to begin in 1935 (Figure 2.5b) reduces the uncertainty bars but results remain consistent with the 1955-2004 analysis. For the 1935-2004 time window the GHG trend is 0.15 ^{o}C to 0.23 ^{o}C per decade that is equal to a 1.05 ^{o}C to 1.61 ^{o}C warming in 70 years, which is also consistent with the observed trend. Our results are consistent with OLS even though this estimate shows a higher positive trend and uncertainty for GHG. Using data from 1955 to 2014 from the ALL simulation increases the signal to noise ratio, reducing the uncertainty bars when considering model error, which implies the simulated warming signal is more consistent across the different models. The results are also compatible with the observed trends which continues to imply a forced component.

2.4 Conclusions

The current study has used a novel Detection & Attribution method from Ribes et al. (2017) to attribute temperature changes of approximately 1.1 ^{o}C per 50 years for Southeast

Brazil, and answer the question of whether or not the observed trends can be attributed to the increase in greenhouse gases. Using the CMIP5 multi-model ensemble mean gave a trend of 0.95 ^{o}C to 1.50 ^{o}C per 50 years from GHG which suggests anthropogenic activities made a substantial contribution to the observed trend with no significant contribution from natural or non-greenhouse gases anthropogenic sources. The results seem to be robust to change in time window of the analysis and by taking account of both observational and model errors. Using CESM as the model mean to investigate which error is dominant in this analysis showed that more than half of the error may come from model uncertainty. It might be possible to reduce this uncertainty by rejecting some models that are very different from the observations. The inclusion of model error had a significant impact in the uncertainty of CESM OA warming signal for 1955-2004 which was not supported by the multi-model mean and by changes in the time window that did not reveal any contribution from other anthropogenic sources.

When other attribution studies are considered, we see that warming trend in Southeast Brazil due to anthropogenic activities is consistent with other regions. A trend of 0.19 ^{o}C to 0.30 ^{o}C per decade due to GHG was found in this study, with a small contribution from other anthropogenics, of -0.07 ^{o}C to 0.01 ^{o}C per decade. The IPCC AR5 reported an attributable warming trend of 0.08 ^{o}C to 0.21 ^{o}C, for global temperature, per decade due to GHG (Bindoff et al., 2013). GHG trends, that are comparable to the overall anthropogenic trends for SE Brazil, are consistent with regional studies for Western China, Canada and Central England that showed attributable decadal warming trends due to anthropogenic activities of 0.19 ^{o}C to 0.30 ^{o}C, 0.07 ^{o}C to 0.23 ^{o}C and 0.14 ^{o}C to 0.26 ^{o}C, respectively (Wang et al., 2018; Wan et al., 2019; Karoly and Stott, 2006). However, unlike these regional studies we are able to calculate contributions from three different forcings. The results shown in this study, of a significant anthropogenic induced warming in a regional scale, also suggests that human induced climate change is becoming very strong at human relevant scales.

Chapter 3

Effects of local vegetation and geographical regional controls in near-surface air temperature for Southeastern Brazil

The results that are presented in this chapter were published in **Atmosphere** (de Abreu et al., 2022).

3.1 Introduction

The average global near-surface air temperature increased 0.85 ^{o}C [0.69 ^{o}C to 0.95 ^{o}C] between 1995-2014 relative to 1850-1900, dominated by increasing anthropogenic greenhouse gases and a minor contribution from aerosols and other natural sources that range from months to decades as solar radiation, volcanism, sea salt, and mineral dust (Gulev et al., 2021). However, this warming rate showed a large spatial variability across continental areas due to other anthropogenic factors like land cover change, increasing urbanization, and aerosols from industrial processes that contribute to either warming or cooling at the regional scale (Doblas-Reyes et al., 2021). For Southeastern Brazil (SEB), a large and populated region larger than 900,000 km^2, with more than 89 million inhabitants, de Abreu et al. (2019) reported warming of 1.1 ^{o}C in 50 years (1955 to 2004) using CRUTEM4 data (Jones et al., 2012), well above the global average, but also attributed mostly to greenhouse gases and in agreement with the global scale effect (Bindoff et al., 2013). However, at SEB there is a lack of information about the spatial variability of temperature trends, and only local studies are available (Blain et al., 2009; Marengo, 2001). We suspected even more about the spatial variability by analyzing historical trends

from 52 weather stations in SEB. They showed a mean of 0.25 ^{o}C 10 yr^{-1} across stations for daily averaged temperature, which compared well with the regional mean of de Abreu et al. (2019), but with a high range of values showing both cooling and warming trends, of +0.02 to +0.51 ^{o}C 10 yr^{-1} for minimum daily temperature (Figure 3.1a), from -0.01 to +0.46 ^{o}C 10 yr^{-1} for average temperature (Figure 3.1b), and from -0.1 to +0.60 ^{o}C 10 yr^{-1} for maximum daily temperature (Figure 3.1c).

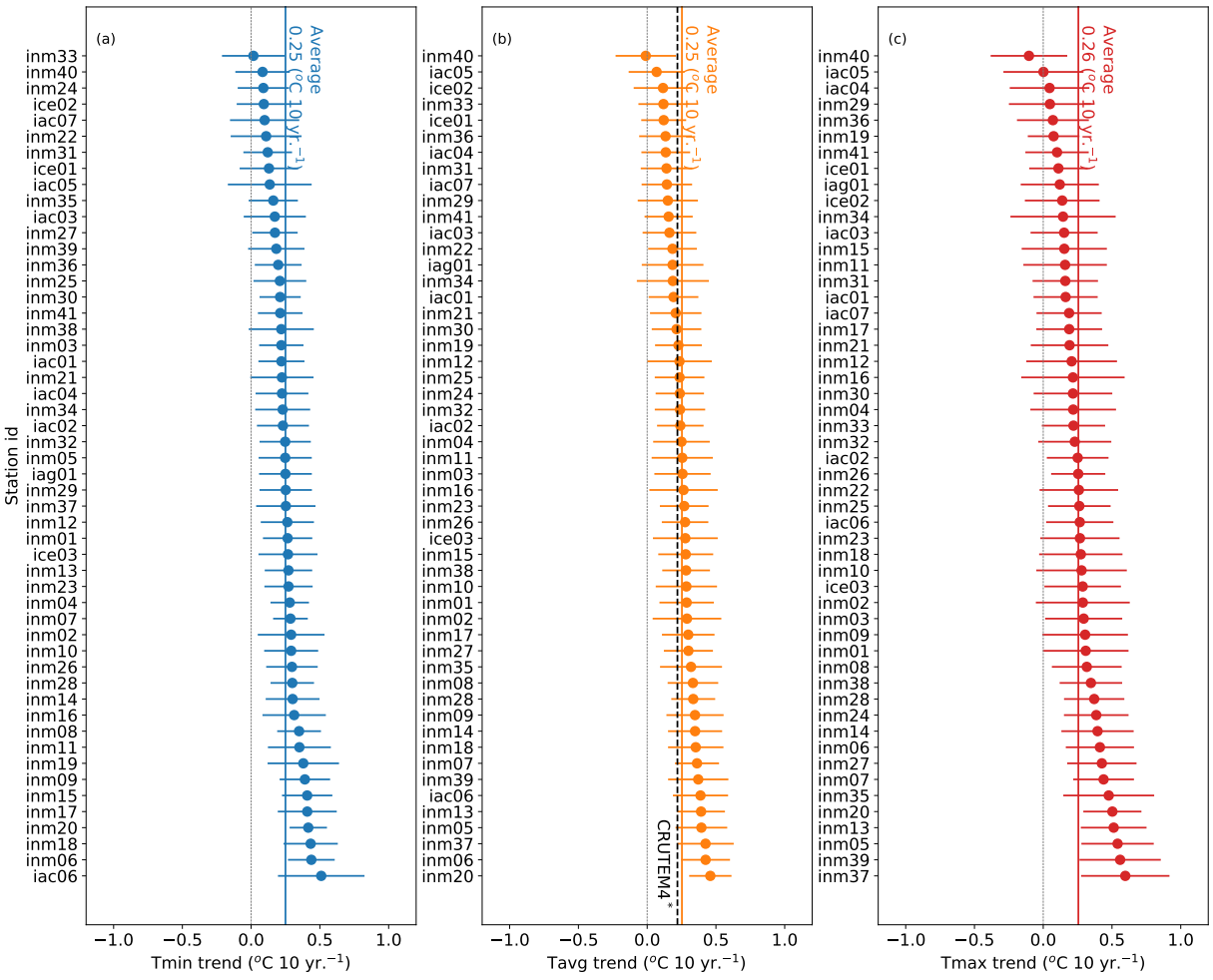

Figure 3.1: (a) Minimum daily temperature (Tmin) linear trend for the stations in southeastern Brazil between 1985 and 2010, in ^{o}C 10 yr^{-1}, ordered from the lowest trend to the highest; (b) same as (a) but for daily average temperature (Tavg); (c) same as (a) but for maximum daily temperature (Tmax). Solid lines are the 95 % confidence interval, vertical solid line is the average of all stations. * The dotted vertical line in (b) is the southeastern Brazil average temperature trend estimated from CRUTEM4 calculated in (de Abreu et al., 2019).

The pattern with a pronounced spatial heterogeneity in temporal temperature trends (Figure 3.1) motivated us to investigate in depth the geographical spatial controls of the long-term mean temperature. Various authors reported connections between the temporal

trends and the controls affecting the mean near air-surface temperature, for example, the land cover (Camilloni and Barros, 1997; Kalnay and Cai, 2003; Wang and Yan, 2016) and topography (Ceppi et al., 2012; Kagawa-Viviani and Giambelluca, 2020), that can possibly enhance or diminish how the large-scale controls, like the greenhouse gases contribution, affect the local response of temporal changes. Therefore, to identify the likely controls of local-regional variability in temporal trends, we have first to guarantee the physical consistency of the measurements, and then obtain the controls of variability in the mean state of temperature.

The SEB contributes over 50 % of the national Gross Domestic Product (GDP) with services, agricultural and industrial goods, as well as headwaters for hydroelectricity and human water supply, where the impacts of climate change are risky due to significant exposure and vulnerability (Hunt et al., 2018; Nobre et al., 2011; Pereira et al., 2017). The region is spatially complex in topography and land cover, where latitudinal variation explains up to 50 % of average temperature range, and longitude, which is a proxy for continentality, has a lower impact (Alvares et al., 2013). Rodríguez-Lado et al. (2007) used a linear regression model to study the patterns of spatial variability of temperature in the state of São Paulo, which is part of the SEB, with altitude and latitude as independent variables, with no significant impact on longitude. The influence of continentality is confounded with the effects of two large mountain ranges placed parallel to the coastline (Serra do Mar and Serra da Mantiqueira), wherein a combined effect of sea breeze and orographic circulation both contribute to cloud cover and cold air advection (Silva Dias et al., 1995). In SEB the land cover in rural areas is dominated by pastureland and sugarcane plantations, and an increasing urbanization that enhances local warming (Oke et al., 2017). For example, the urban heat island effect in the city of São Paulo was characterized by accelerated expansion of urban areas since the early twentieth century, and is partly responsible for increasing local air temperature (Silva Dias et al., 2013). However, the urban heating is not homogeneous over it spatial extent and depends on many different physical factors such as diurnal cycle, atmospheric turbulence, thermal properties of constructed materials, and urban morphology as well as the background temperature (Camilloni and Barros, 1997; Zhao et al., 2014; Manoli et al., 2019). On this subject, Kagawa-Viviani and Giambelluca (2020) used multiple linear regression at the regional scale in Hawaii and showed a dependency of the spatial distribution of minimum temperature with vegetation

and wind speed, which was not observed for maximum temperature, that had a dependency with precipitation and cloud cover. Also, other features might explain the effects at the local scale with complex topography, like slope and aspect (Sun and Zhang, 2016).

In general, the studies of attribution of spatial variability of temperature in SEB are modest, using a limited set of independent variables that did not include likely effects of land cover and cloud cover caused by mesoscale circulation (Alvares et al., 2013; Rodríguez-Lado et al., 2007). To contribute to this understanding, multiple linear regression (MLR) is advantageous, as a simple and parsimonious model for explicit quantification of different independent variables. However, MLR has intrinsic restrictions, like assuming a linear relationship between the dependent and independent variables. The Generalized Additive Model (GAM) was developed to overcome this limitation, as a generalization of the Generalized Linear Model (GLM). GAM fits smoothing functions to build relationships between a set of predictors and the predicted variable instead of assuming a linear dependency (Wood, 2017). Still, relationships in GAM can be easily interpreted graphically for each independent variable, and provide inferences about individual contributions, differently from more complex non-linear methods like neural networks and other "black box"types of models.

Our objective is to assess the influence of local vegetation and regional scale geophysical controls on the spatial variability of near-surface temperature in Southeastern Brazil. Using a wide network of weather stations, and with the attempt to understand and describe a simple way of showing the dependencies of temperature with the geophysical features of zonality and continentality, altitude, cloud cover, and as a novelty the influence of local vegetation, with parsimonious linear (MLR) and non-linear (GAM) models.

3.2 Materials and Methods

3.2.1 Weather station information

We used the average minimum and maximum air temperature data (Tmin, and Tmax, respectively) calculated from daily data, from a network of weather stations in Southeast Brazil (INMET, IAC, ICEA, and IAG institutes) (Figure 3.2). We only used conventional weather stations that are installed according to standards about site selection and exposure, and used the longest available period to compute the climate means (WMO, 1967, 2018).

We selected the years from 1985 to 2010, which was optimum because of satellite data availability to calculate the Normalized Difference Vegetation Index (NDVI) values and by maximizing the number of available stations from IAC network, which was available until 2010. We quality controlled the available data by first checking metadata for the exact location of each station to calculate NDVI and fill in missing information about altitude. The second step included comparing the timeseries with predefined thresholds of physically plausible values, persistency of repeated values, and discarding data outside the boundaries defined by three times the standard deviation. We also compared it with nearby stations when more than six of them were available in a 100 km radius of the target stations, according to Meek and Hatfield (1994) and Shafer et al. (2000). Weather stations that were moved during the selected period were not used in the analysis, resulting in 52 stations that met all requirements (Table B.1). A total of 15 stations have between 10-30 % of missing data for Tmin and 20 stations for Tmax, and only five have between 20-30 % for both Tmin and Tmax.

To interpolate this large number of missing data we used an EOF based method described in Beckers and Rixen (2003). Henn et al. (2013) showed that this method is more suitable to interpolate longer periods of data. We start with a first guess for the missing variables based on the average of each station and then apply the Singular Value Decomposition (SVD) to the matrix $\mathbf{X}$ ($n \times s$ dimensions, where n = the number of days, s = number of stations) whose columns are each station and the rows are each day in the selected time window:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{3.1}$$

$\mathbf{U}$ is the matrix of eigenvectors for $\mathbf{X}\mathbf{X}^T$, $\mathbf{D}$ is the diagonal matrix with the singular values and $\mathbf{V}$ is the matrix of eigenvectors for $\mathbf{X}^T\mathbf{X}$. A number m of components are retained to reconstruct the original matrix and produce a new estimate of the missing data.

$$\mathbf{X}' = \mathbf{U}[:, 1{:}m]\mathbf{D}[1{:}m, 1{:}m]\mathbf{V}[:, 1{:}m]^T \tag{3.2}$$

The new values of $\mathbf{X}'$ where missing data were located are the new estimated values for $\mathbf{X}$, while the non missing values are kept unchanged. This process is repeated iteratively

until convergence is reached. To calculate the optimal number m of EOFs to be used in the interpolation we randomly introduce 2 % of missing data that is used to calculate the root mean squared error (RMSE). The m with the lowest RMSE is then chosen to perform the interpolation.

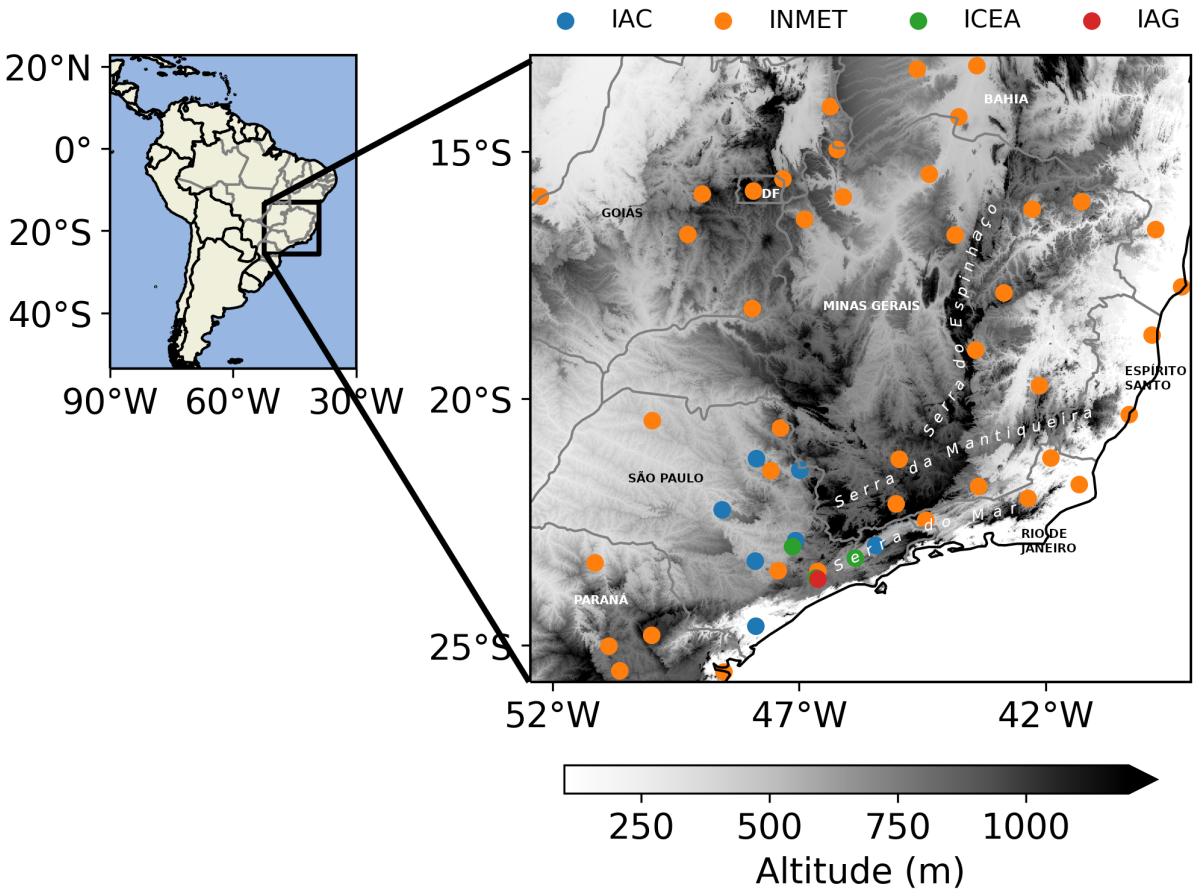

Figure 3.2: Geographical position of the weather stations used in this research. Each station is identified by a colored point according to the different networks they belong (red: Instituto de Astronomia, Geofísica e Ciências Atmosféricas/Universidade de São Paulo (IAG); blue: Instituto Agronômico de Campinas/Secretaria de Agricultura e Abastecimento de São Paulo (IAC); orange: Instituto Nacional de Meteorologia (INMET); green: Instituto de Controle do Espaço Aéreo/Ministério da Aeronáutica (ICEA)). Shading represents the altitude in meters. Upper case and bold letters are the federal states delimited by the grey solid lines, and italic highlight the location of three important mountain chains: Serra do Mar, Serra da Mantiqueira, and Serra do Espinhaço.

We used $m = 5$, which had an RMSE of 1.4 ^{o}C for minimum temperature. As a comparison, using inverse distance weighting (IDW) resulted in an RMSE of 2.1 ^{o}C. The lower RMSE is in part resulted by the fact that the SVD decomposition incorporates both spatial and temporal information about the station that is being interpolated, differently than IDW which only uses spatial information. When dealing with sparse networks like

the one we are using this is particularly useful to interpolate large gaps as shown in Figure 3.3. IDW is able to capture the seasonal variability in the gap between 1985 and 1990 from the nearby stations, however it has an offset of about 5 ^{o}C, differently than the EOF method that preserves the temporal average. Also, the interpolated values with the EOF method are closer to the observed range while with IDW it is not uncommon to see artificial minimums. After the interpolation of missing values, we homogenized the data (described in Section B.2), which is important to reduce the influence of random errors, detected through breakpoints in the timeseries, and caused by changes in instruments, drift, transcribing errors, that are not available in the metadata. We kept, however, stations where changes in NDVI and land use possibly occurred.

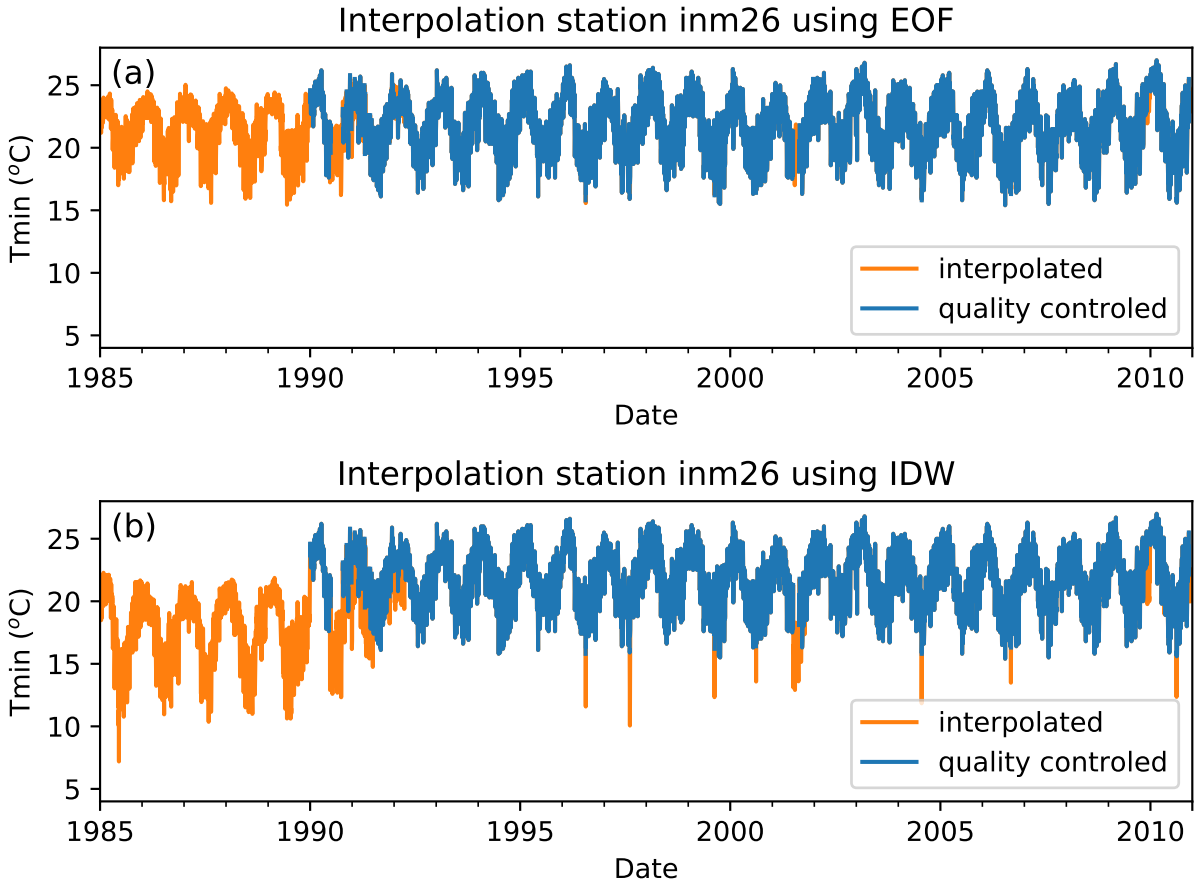

Figure 3.3: Daily average temperature in ^{o}C for the station inm10 after the quality control (blue) and after the interpolation (orange) for: (a) The EOF method described in Beckers and Rixen (2003); (b) Inverse Distance Weighting (IDW).

We used NDVI estimated by Landsat 5 satellite data (with 30 m of spatial resolution), using all available images for each of the 26 years of data. Then, the temporal average was

calculated (Figure B.4). We computed the spatial average using a circle with a variable radius with the origin at the station coordinates. After a sensitivity analysis with the length of the radius, we selected 300 m for the consolidated averages, based on the highest correlation with Tmin in the annual aggregation (Figure B.5a). Altitude was selected from metadata when available, and we used the estimates from Shuttle Radar Topography Mission (SRTM,Farr et al., 2007) when not. Cloud cover was estimated from a 15-year average (2000-2014) from the MODIS satellite with 1 km of spatial resolution (Wilson and Jetz, 2016).

3.2.2 Statistical model

We used two models to represent the spatial variability of air temperature: Multiple Linear Regression and Generalized Additive Model (described in more detail in Section B.3), which can be expressed as the sum of smoothing functions f_j (Hastie et al., 2009):

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon \tag{3.3}$$

Where Y is the dependent variable (Tmin, Tmax), α is the intercept term, X_j is the jth independent variable ($j = 1, ..., p$; p = number of regressors), and ϵ is the residual error which has a normal distribution with zero mean and constant variance. The smoothing functions in Equation 3.3 have the ability to fit non-linear relations between the independent and dependent variables, however they are fitted following a penalized least squares algorithm to penalize too wiggly functions preventing an overfit of the model. The Generalized Cross Validation (GCV) is used to estimate the penalization for each variable (Wood, 2017). By doing so, if the penalization is large we can even fit linear relations between the predictor and predicted variable. We used the concept of estimated degrees of freedom (edf) to determine whether or not the relations are linear since if edf ≥ 2, there is strong evidence that a nonlinear function might be the best fit (Wood, 2017). We used the following source equation to fit the GAM:

$$\hat{T} = \hat{\alpha} + \hat{f}_1(\text{altitude}) + \hat{f}_2(\text{NDVI}) + \hat{f}_3(\text{cloud cover}) + \hat{f}_4(\text{lon, lat}) \tag{3.4}$$

Every function $\hat{f}$ is fitted using the `mgcv` package for R (Wood, 2017), and the estimator $\hat{\alpha}$ is the spatial sample mean of the dependent variable T. The model was fitted individually

using the average from daily data for each of the following time aggregations: summer (December, January, and February), winter (June, July, and August), and annual (January to December). We highlight the usage of a single term for representing the zonal and continental effects through $f_4(lon, lat)$ that will be discussed in more detail in the results.

GAM is based on smoothing functions that require many different parameters to represent non-linear dependencies and therefore have a greater degree of complexity when compared with MLR. In the case of MLR, the model that we used is analogous to Equation 3.4, where $f_j(X_j) = \beta_j X_j$ and, therefore:

$$\hat{T} = \hat{\alpha} + \hat{\beta}_1 \text{altitude} + \hat{\beta}_2 \text{NDVI} + \hat{\beta}_3 \text{cloud cover} + \hat{\beta}_4 \text{lon} + \hat{\beta}_5 \text{lat} \tag{3.5}$$

Equation 3.5 has a maximum of six degrees of freedom, one for each regressor, while a GAM model can have more estimated degrees of freedom, depending on the nonlinearities found in the data. Therefore we compare the results for Equations 3.4 and 3.5 with the same regressors.

The models were evaluated according to the Bayesian Information Criteria (BIC), $\text{BIC} = n\text{SSE} - n \log n + p \log n$, where SSE is the sum of squared residuals and n is the number of data samples. The lower the BIC, the lower the error; however, a penalization term accounts for the number of parameters in the model, prioritizing more parsimonious models with fewer parameters. To ensure that we are using the most conservative model, we tested all possible combinations of terms from Equations 3.4 and 3.5 (15 possibilities for GAM and 31 for MLR) and selected the model in which all terms were statistically significant with a 5 % significance level and with the lowest BIC. A flow chart of the main steps is presented in Figure 3.4.

To detect if collinearity effects would be a problem in the model, affecting the interpretation of the results, we calculate the variance inflation factor (VIF), which estimates how much of the variance of the parameters in the regression is inflated in comparison with the case where they are linearly independent. The maximum value of VIF was 1.64 for latitude, which is inside the range of recommended values and should not cause any significant problems due to the collinearity of the regressors (Kutner et al., 2005).

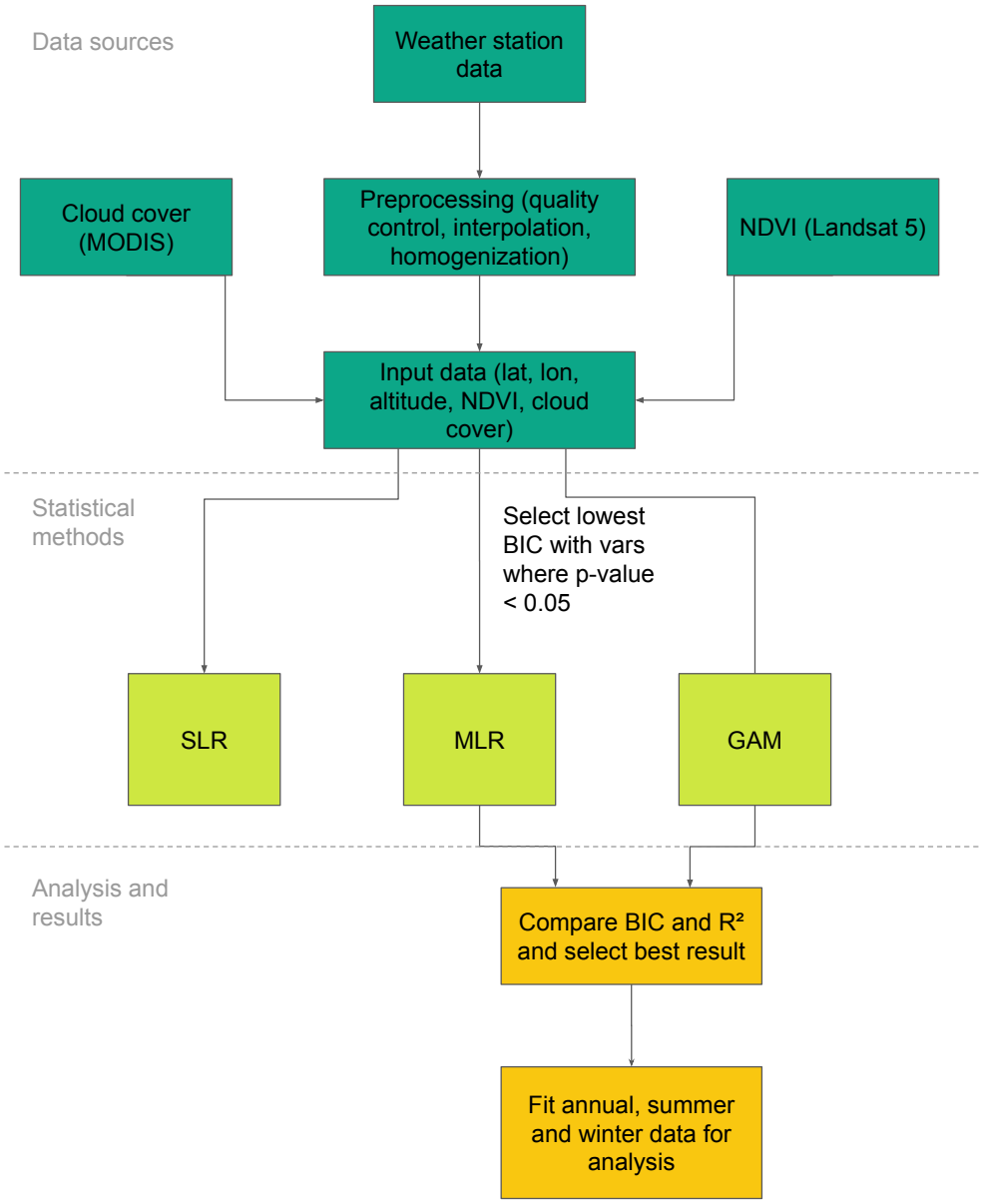

Figure 3.4: Flow chart of the main steps taken in the methods, including data source preprocessing, statistical methods (Simple Linear Regression (SLR), Multiple Linear Regression (MLR), and Generalized Additive Model (GAM)), and selecting the best model for the analysis.

3.3 Results and Discussion

To elucidate the possible dependency of the average temperature with the selected regressors (latitude, longitude, NDVI, and cloud cover), we show the scatter plot for the annual average Tmax and Tmin (Figure 3.5) with each independent variable for a visual inspection of the relationship between them. We notice a wide range of average temperatures of almost 10 ^{o}C, to be more precise, between 12.4 and 21.8 ^{o}C (17.3 $\pm$ 2.0 ^{o}C)

for Tmin and between 23.8 and 32.9 ^{o}C (28.5 $\pm$ 2.3 ^{o}C) for Tmax. We fitted the simple linear regression (SLR), and all scaling factors are statistically significant (p-value < 5 %), except for Tmax and longitude (p-value $= 0.28$), suggesting that there is a dependency between each pair of variables. Based solely on the scaling factor, Tmax has a greater sensitivity than Tmin with latitude and cloud cover, while Tmin is more sensitive to NDVI, longitude, and altitude. In particular, we notice an atypical pattern between Tmax and altitude, with a heterogeneous distribution of points close to altitudes of zero meters.

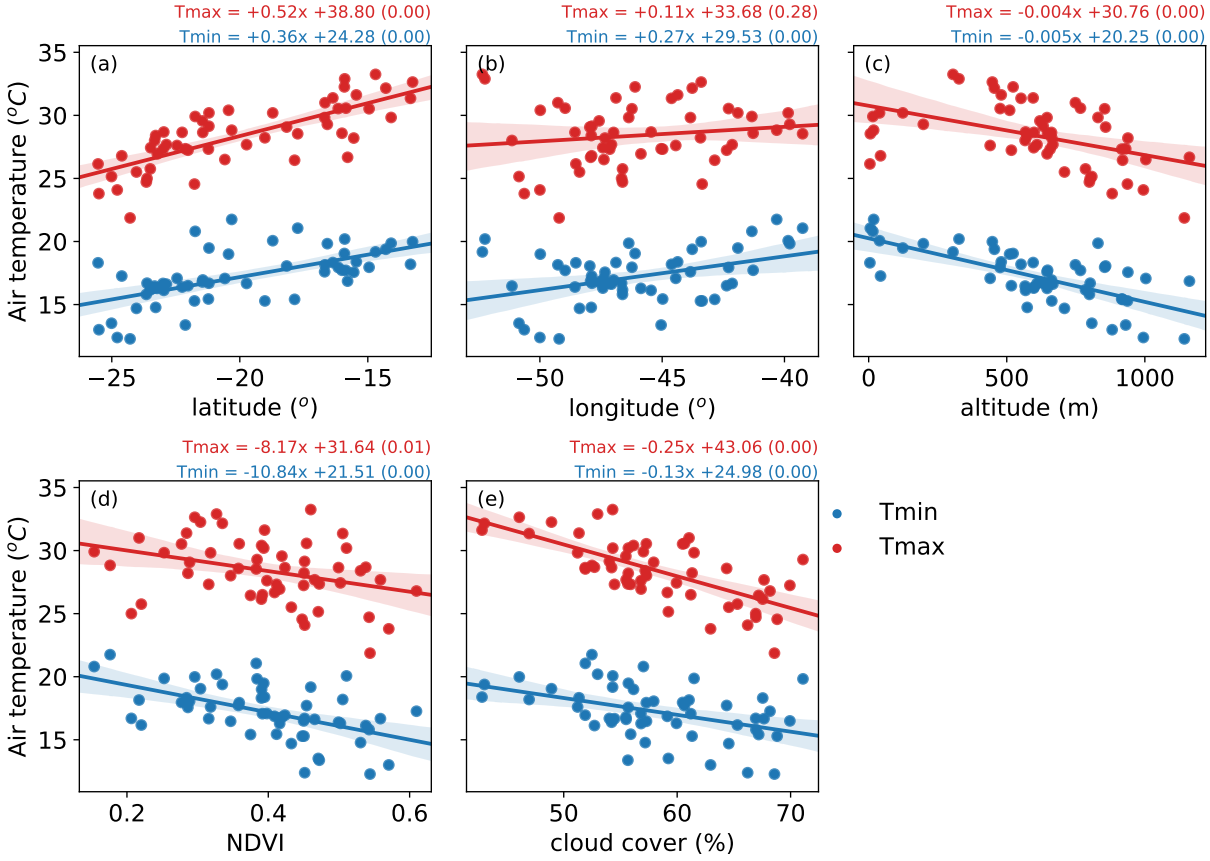

Figure 3.5: Scatter plot of the average annual maximum (red) and minimum (blue) temperature for each station, with the following independent variables: (a) latitude, (b) longitude, (c) altitude, (d) NDVI, (e) cloud cover. The solid line is the univariate linear regression fitted using ordinary least squares with the 95 % confidence interval in shading. The equation and estimated parameters are located above each scatter plot, with the p-value of the scaling factor in parenthesis.

3.3.1 Model fitting

The estimated model parameters for MLR ($\hat{\beta}_j$) and GAM (edf) are available in Table 3.1, with the respective R^2 and BIC for the fitted model. They show that both models have a high percentage of the explained variance for temperature, with R^2 between 86 %

and 94 %. GAM explains a higher percentage than MLR, with a difference of almost 9 % for Tmax and 7 % for Tmin, and a lower BIC. For Tmax, MLR had five degrees of freedom (number of statistically significant independent variables), while for GAM there were 8.9 edfs (sum of the individual edfs for each independent variable). For Tmin MLR showed four degrees of freedom, while GAM had 10.3 edfs.

We noticed that the fitted functions using GAM are significant for Tmax with (lon, lat), altitude, and cloud cover, and Tmin with (lon, lat), altitude, and NDVI. For MLR, the fitted linear functions were the same as GAM, except for longitude, which was not statistically significant for Tmin. The fitted function for NDVI using GAM has an edf close to 1, which suggests a linear relation with Tmin, with no significant gain compared to MLR. Still, for Tmin, the relationship with altitude was slightly non-linear (edf = 1.55). For Tmax, the edf of altitude and cloud cover is close to 2, suggesting a non-linear relationship. For both Tmin and Tmax, the function s(lon, lat) was the one with the highest edf, 7.71 and 5.40, respectively. The MLR fitted latitude for both Tmin and Tmax, but longitude only for Tmax, differently from SLR, in which both Tmin have a significant dependency on longitude (Figure 3.5b). There is a higher complexity when fitting the MLR with different variables leading to a non-significant relationship between Tmin and longitude. For both models, there was a consensus with NDVI only being significant for Tmin, and cloud cover only for Tmax.

Table 3.1 - Estimated degrees of freedom (edf), and scaling factors $\hat{\beta}_j$ of each independent variable. R^2 is the coefficient of determination, and BIC is the Bayesian Information Criteria. Only terms with p-value < 0.01 are displayed.

Generalized Additive Model (GAM)			Multiple Linear Regression (MLR)		
	Tmax (edf)	Tmin (edf)		Tmax $\hat{\beta}_j$	Tmin $\hat{\beta}_j$
Intercept	28.5	17.3	Intercept	37.5	27.5
-	-	-	lon	-0.17	-
s(lon,lat)	5.40	7.71	lat	0.44	0.26
s(altitude)	1.95	1.55	altitude	-0.003	-0.005
s(NDVI)	-	1.00	NDVI	-	-6.04
s(cloud cover)	1.56	-	cloud cover	-0.10	-
R^2	94.3 %	93.4 %	R^2	85.8 %	86.1 %
BIC	125.8	127.2	BIC	155.4	136.4

3.3.2 Regional range of GAM parameters

The absolute contribution in degrees Celsius for each of the individual functions of GAM, and their variability is presented in Figure 3.6. The contribution due to the geographical position, s(lon,lat), showed a zonal distribution with negative values to the south and positive to the north for both Tmin and Tmax (Figure 3.6a and d). This pattern was mainly expected from the differential radiative heating at the surface, but we noticed, especially for Tmin, a zonal deformation in the shape of an inverted "U", probably due to the complex terrain to the east of the inverted "U", which had a greater cooling for nighttime temperature (Figure 3.6a). In those areas of Serra da Mantiqueira and Serra do Espinhaço (Figure 3.2), the mountain-valley circulation contributes to creating areas of significant cooling, usually close to the valleys (Martin et al., 2019). For Tmax, the identified pattern of s(lon,lat) shows a meridional gradient parallel to the coast, consistent with the well-established thermal contrast between ocean and continent around noon (Oliveira and Silva Dias, 1982; Silva Dias et al., 1995). This variability of s(lon,lat) and its difference between Tmin and Tmax points to the importance of using the multiple parameter term with longitude and latitude, which were not evident when using SLR (Figure 3.5b), because of the simplification in the spatial dependency.

The near-surface air temperature variation with altitude (in ^{o}C per km of elevation) across a region is known as the terrestrial lapse-rate (TLR), which is firstly influenced by the vertical temperature profile in the atmosphere, which is represented by the ascension of a parcel of air under adiabatic process, that expands and cools down, according to the dry adiabatic lapse-rate Γ_d = -9.8 ^{o}C km^{-1}. In reality, Γ_d is summed with different effects that may affect Tmax and Tmin differently, from local to regional processes. For example, Li et al. (2015) showed a dependency of TLR increasing over warmer and wetter areas in China. Martin et al. (2019) showed lower absolute values of TLR for Tmin than Tmax in the mountain ranges of Southeastern Brazil, which was attributed to the nighttime thermal belt at about 200 m up the valley bottom. This tends to reduce the cooling rate with altitude, having significant seasonal variability that marks the observed lapse-rate, measured by radiosonde, with an average global value of γ = -6.5 ^{o}C km^{-1} (Wallace and Hobbs, 2006). The TLR is estimated based on γ but restricted to surface measurements that are grouped regionally, where microclimatic phenomenons are added, like the type

and hydrological state of the underlying vegetation, types of urbanization like, for example, local climate zones, and etc. (Kirchner et al., 2013; Li et al., 2015; Martin et al., 2019; Wanderley et al., 2019).

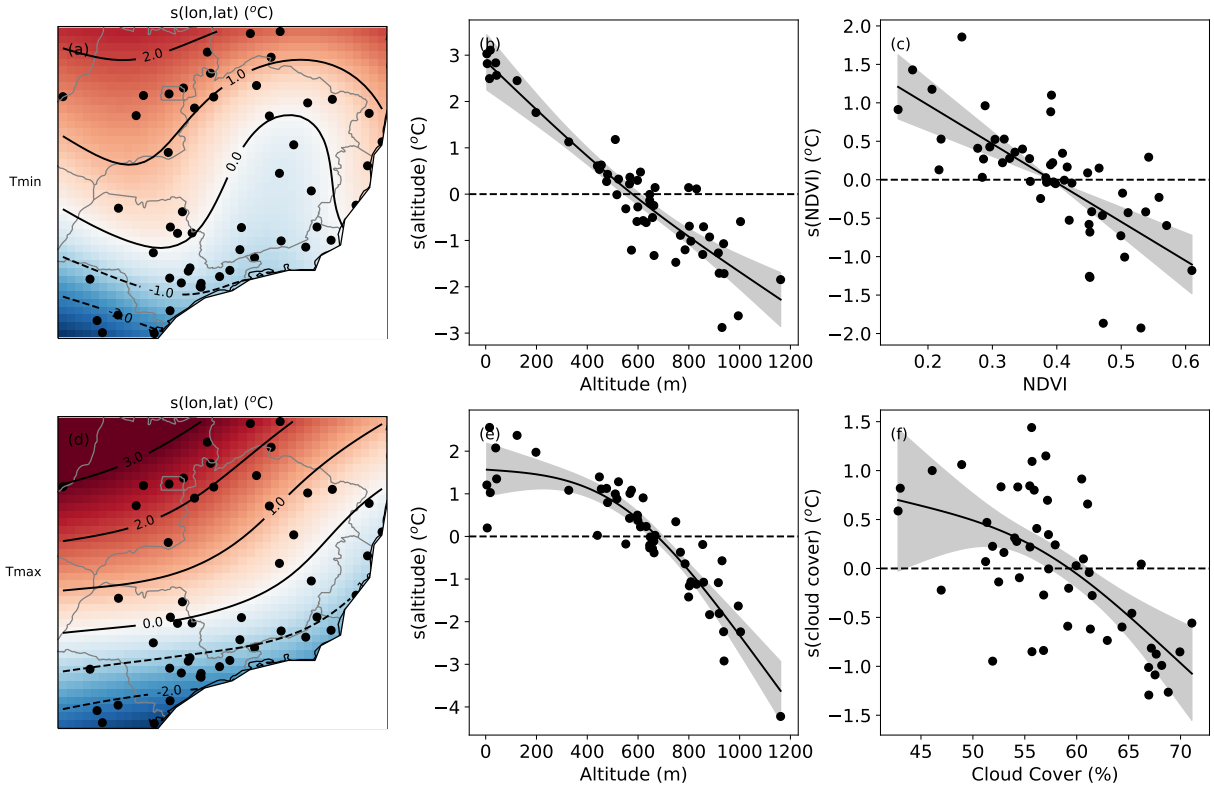

Figure 3.6: Contribution of GAM terms, in ^{o}C, for the annual mean of Tmin (a, b, c) and Tmax (d, e, f). In (a) and (d), is the function related to the geographical position s(lon, lat); in (b) and (e) is the altitude in meters above sea level (a.s.l.); (c) the NDVI and (f) the cloud cover in %. In (a) and (d) we show the position of each station used to fit the model. In (b), (c), (e), and (f): the points are the partial residual of the given function, and the fitted GAM response is displayed as a solid line with a 95 % confidence interval

With GAM we tried to identify the relationship between temperature and altitude, analogous to the TLR. We estimated a reduction of Tmin and Tmax with altitude under a range of regional response of approximately 6 ^{o}C (Figures 3.6b and 3.6e), similarly to SLR (Figure 3.5c), but with very distinct response patterns between Tmin and Tmax. For Tmin, the variability is well established and with a slightly non-linear response (edf = 1.55), with an approximate variation of -4.4 ^{o}C km^{-1}, roughly estimated using the difference between the highest and lowest altitudes. Differently, for Tmax we noticed that the sensitivity for changes in altitude is low in the first 500 m above sea level (a.s.l.), but it is more well established above 500 m. We estimated the TLR between 500 and 1200

m a.s.l., equal to -7.0 ^{o}C km^{-1} for Tmax and -4.0 ^{o}C km^{-1} for Tmin. This value of TLR for Tmax is steeper than Tmin and is reasonably comparable with the estimates of -7 ^{o}C km^{-1} from a study made in a rural area in a 10 km^{2} basin in complex terrain in Serra da Mantiqueira (Martin et al., 2019).

The cloud cover response acts by reducing temperature as cloud cover increases, due to the increased albedo cooling effect, and it is statistically significant in GAM only for Tmax (Figure 3.6f). Its contribution is lower than functions like s(lon, lat) and s(altitude), but it is still relevant with an amplitude of about 2 ^{o}C, with a cloud cover variation between 40 % and 75 %. The contribution for Tmax was not very clear between 40 and 55 % of cloud cover, but it is better established between 55 % to 75 %, with a temperature reduction of almost 2 ^{o}C for a 20 % increase in cloud cover (Figure 3.6f).

The s(NDVI) contribution is to reduce the temperature as NDVI increases, which is statistically significant in GAM only for Tmin (Figure 3.6c), with an approximately linear function. The amplitude is close to 2.5 ^{o}C between minimum and maximum NDVI values, comparable with the amplitude from cloud cover for Tmax, and is lower than the altitude response.

The NDVI range is relatively wide, from 0.15 to 0.61, where most weather stations are located in urban or suburban areas (Figure B.7), and the minority are in rural areas. As a rough estimate, we suppose a NDVI of 0.4 as a threshold in which below this value the vegetation cover is too low, with a predominance of built-up areas, and 0.65 as a threshold which above this value there is a high vegetation cover (Lambin and Ehrlich, 1996; Bhang, 2014), even though there are no exact values of NDVI that separate areas with low or high vegetation cover, especially when considering urban areas. In our sample data, we have 54 % of stations with NDVI < 0.4 and none of them with NDVI > 0.65. However, this does not imply that the pixels used to compute the average in a 300 m radius circle (with a spatial resolution of 30 m), do not contain areas with NDVI > 0.65, like in urban parks and rural areas, as we will discuss in the next section about the heterogeneity of land use.

The variability of s(NDVI) that is inversely proportional to NDVI has probably the same causes of the urban heat island, the increase in stored energy inside the urban canopy during the day and a slow release in the form of longwave radiation that is persistent throughout the night, heating the air temperature (Oke et al., 2017). The increase in Tmax can occur, primarily due to the increase in the Bowen ratio in the built-up areas,

but this is not a consensus due to losses from atmospheric turbulence (Oke et al., 2017). However, our analysis did not show a statistically significant response between Tmax and NDVI for both GAM and MLR, even though the SLR showed (Figure 3.5d). This is possibly due to the correlation between NDVI and latitude (Pearson correlation of -0.38, with p-value < 0.01) that affects the estimation of the parameters (Kutner et al., 2005).

3.3.3 Seasonality of GAM response

We further use GAM in a more detailed manner by fitting the summer and winter averages to see how the seasonality changes the estimated functions (Figure 3.7). For Tmin, the geographical position function, s(lon,lat), had a more significant contribution during winter, when horizontal gradients are more prominent, with a north-south difference of about 8 ^{o}C (Figure 3.7b), while in summer they are at most 3 ^{o}C (Figure 3.7a). For maximum temperature, the seasonality response was similar to Tmin, with meridional gradients above 8 ^{o}C in winter and lower than 3 ^{o}C in summer.

For the altitude function, we noticed little seasonal variation for Tmax (Figure 3.7g), where the lower sensitivity remains for altitudes that are lower than 500 m a.s.l., but with a steeper gradient above this altitude. For Tmin, we especially noticed that the dispersion around the fitted function was larger during winter than during summer, mostly in weather stations with an altitude lower than 500 m (Figure 3.7c). Kirchner et al. (2013) and Li et al. (2015) suggest that this is related to the increase of days with thermal inversions during winter, decreasing the lapse-rate. In fact, Martin et al. (2019) showed that in small basins of meso-γ scale in Southeastern Brazil, the lapse-rate for Tmin is positive in the first 200 m of altitude, being more intense during winter, where it is limited by the thermal belt of the nocturnal boundary layer, and it becomes negative for altitudes higher than 200 m. It seems like this effect is incorporated in some of the analyzed stations, which helps to explain the pattern in winter (Figure 3.7c), but being difficult to quantify it.

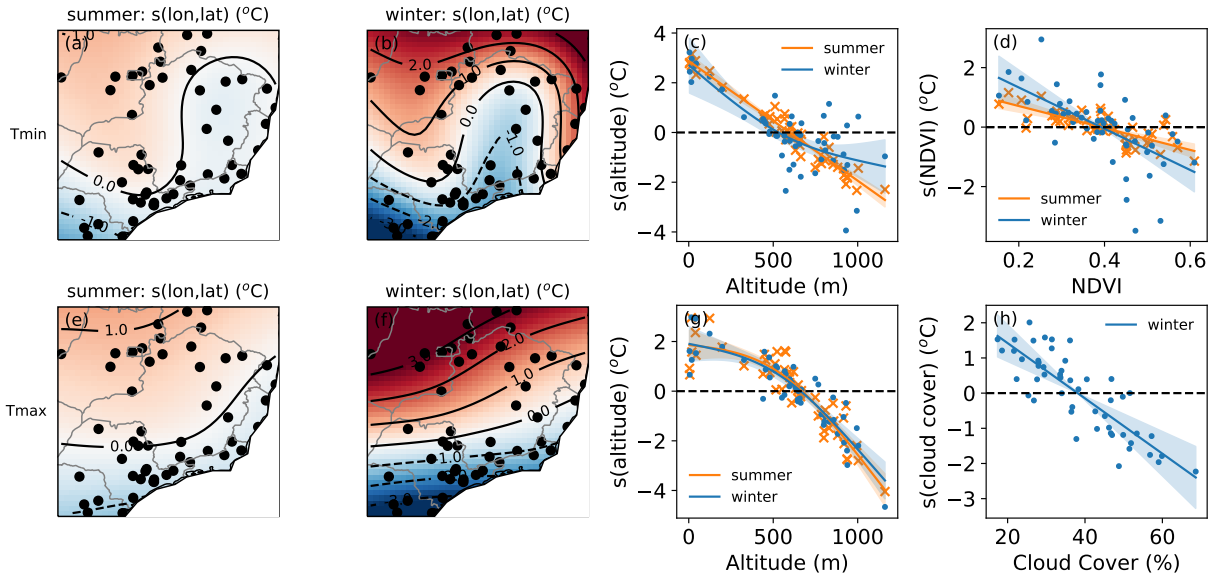

Figure 3.7: Contribution of GAM terms, in ^{o}C, for the seasonal mean of Tmin (a, b, c and d) and Tmax (e, f, g and h) for summer and winter. (a), (b), (c) and (f) shows the geographical position, s(lon, lat); altitude in (c) and (g); NDVI in (d); and cloud cover in (h). In (a), (b), (c) and (f) we show the position of each station used to fit the model. In (c), (d), (g) and (h),: the points are the partial residual of the given function, and the fitted GAM response is displayed as a solid line with a 95 % confidence interval.

The cloud cover functions were statistically significant only for winter (Figure 3.7h) and showed an amplitude of approximately 4 ^{o}C for cloud cover changes between 20 and 70 %, comparable with the response from altitude and (lat, lon). The NDVI function for Tmin was significant for both seasons, with higher amplitude in winter (3 ^{o}C) than in summer (1.5 ^{o}C), even though there is a higher dispersion in winter (Figure 3.7e). Southeastern Brazil is characterized by dry winters and consequent lower cloud cover (Reboita et al., 2010), also seen in Figure 3.7h, where values are mostly below 50 %. This characteristic can potentially increase local contributions from vegetation and urbanization, with clear sky nights that are favorable to promote heating in urban areas (low NDVI) when compared to rural areas (high NDVI) (Oke et al., 2017).

3.3.4 Impact of land use heterogeneity in urban areas

As discussed in the previous sections, we found a correlation between minimum temperature and NDVI defined in a 300 m radius circle (hereafter referenced as NDVI300) and for Tmax but only in the SLR case, with no statistical significance in GAM and MLR. However, other factors may influence the signal's amplitude, like wind speed which increases this spatial variability when wind speed is low (Section S4). There is also the possibility

of contribution from urban heat islands that may extend up to 60 km from the city center (Hicks et al., 2010), with great seasonal variability from parameters that are related to land cover and temperature of the order of 10^3 m (Suomi et al., 2012).

According to Stewart (2011), the spatial homogeneity of land cover up to the order of 10^3 m is important to evaluate the contributions from land use in urban temperature and urban heat island. Therefore, we reevaluated the spatial distribution of NDVI in an area of higher land cover spatial heterogeneity around the weather station. We take the example of the city of São Paulo, with a large urbanized area of 2,310 km^2, and two stations that are separated by a distance of approximately 4 km: iag01 (Figure 3.8a and 3.8b) located inside a park with more vegetation and higher NDVI, that is however surrounded by a densely urbanized area with lower NDVI; and ice03 (Figure 3.8a and 3.8c) station located in an airport with low NDVI in all surrounding area. On the other hand, we also identified two other stations, one at the north of SEB (inm07) (Figure 3.8a and 3.8d) and another one at the south (inm37) (Figure 3.8a and 3.8e) with different NDVI spatial distributions.

After finding differences in NDVI distribution, one of them in the same city, we reassessed the GAM for the prediction of average Tmin and Tmax, by considering the influence of NDVI by a smoother with multiple predictors using NDVI300 and NDVI3000 (NDVI defined in a 3,000 m radius circle), referenced as s(NDVI300, NDVI3000). This term considers the local scale land use (NDVI300) and the regional scale (NDVI3000), which involves transport at the scale of the urban boundary layer. We chose the radius of 3,000 m to capture the influence of the regional scale based on the leveling off of NDVI correlation with temperature, as noticed in Figure B.5.

We fitted the GAM with s(NDVI300, NDVI3000), which was statistically significant for Tmin in all temporal aggregations, but not Tmax, similarly to the previous models using s(NDVI300) only. However, by using the term with multiple parameters, there was a best overall fit, with an increase of the explained variance from 93.4 % to 96.1 %, and a reduction in the BIC, from 127.2 to 123.7 (Table B.4).

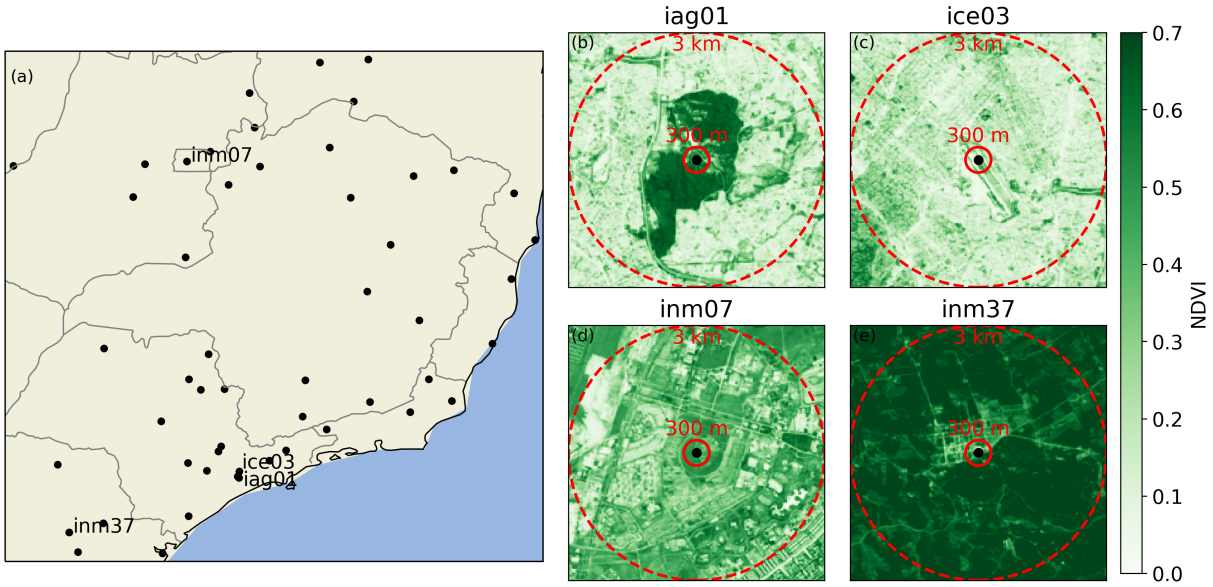

Figure 3.8: (a) Map of southeastern Brazil with all weather stations in this study, with a highlight for: iag01, ice03, inm08 e inm37; annual average NDVI (1985 to 2010) for: (b) iag01, (c) ice03, (d) inm08 e (e) inm37.

We did not notice a significant difference in the patterns associated with s(altitude) and s(lon, lat), but there are relevant applications to s(NDVI300, NDVI3000), as shown in Figure 3.9. The pattern of the function for the annual average Tmin shows negative values and colder weather stations (blue area in Figure 3.9a), which are dominated by high values of NDVI. In the opposite direction, the positive values represent warmer stations (red area in Figure 3.9a), that covers the entire range of NDVI. Still, in the annual average Tmin, in general, NDVI3000 was greater than NDVI300, as shown by the position of most stations distributed around the 1:1 line (Figure 3.9a). This implies that the areas closest to the station's position have relatively lower green cover than the surrounding areas (up to 3,000 m). There are few but important exceptions, like iag01 and inm08, that are relatively warm stations with high NDVI300, positioned in local areas with more vegetation than the surrounding areas (Figures 3.8b,d and 3.9a). With respect to the seasonality of s(NDVI300, NDVI3000), the amplitude of the contribution was greater in winter, with approximately $\pm$ 2 ^{o}C (Figure 3.9c), when compared to the summer, which was $\pm$ 1 ^{o}C (Figure 3.9b).

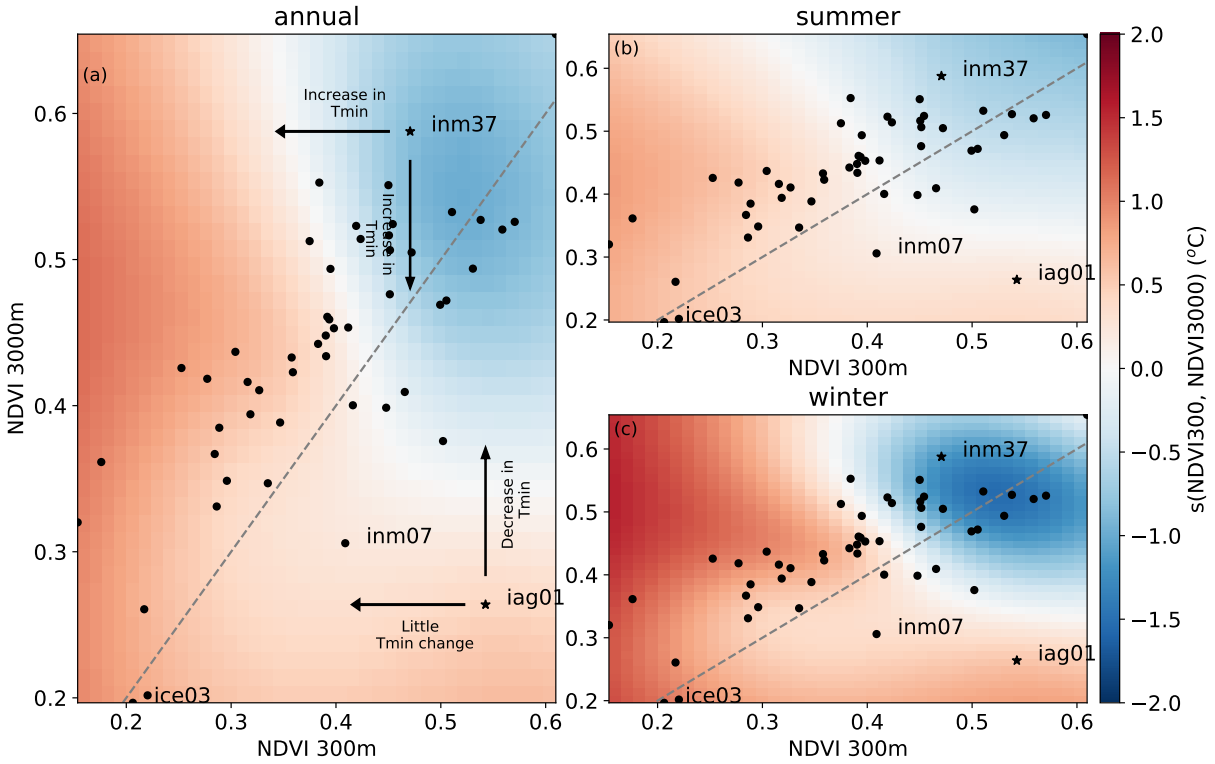

Figure 3.9: Contribution from the s(NDVI300, NDVI3000) function from GAM, in ^{o}C, for minimum temperature in the following time aggregations: (a) annual, (b) summer, and (c) winter. The black dots represent the NDVI in each station, with the names highlighting the position of the stations given in Figure 3.8. The arrows indicate changes in the contribution of s(NDVI300, NDVI3000) due to changes in NDVI at given direction (increase or decrease of NDVI). The dashed line is the 1:1 line, where NDVI300 = NDVI3000.

An application of the function s(NDVI300, NDVI3000) is to verify how a hypothetical change in NDVI would contribute to altering the temperature in a given weather station. For example, shifting the station iag01 in the x direction (as displayed by the arrows in Figure 3.9a) by reducing the NDVI300 (keeping NDVI3000 constant), there is little temperature change, however, shifting it in the y direction and increasing NDVI3000 (keeping NDVI300 constant) leads to a decrease in temperature by reaching the blue part of the plot. Another example with a different result is the station inm37 which has high values of both NDVI300 and NDVI3000. In this case, a decrease in NDVI300 (keeping NDVI3000 constant) with changes in the x direction, increases temperature when reaching the warmer side of the plot, as well as a change in the y direction with a reduction of NDVI3000 (keeping NDVI300 constant). We used a large spatial domain to fit the model, that expresses the combined effects of a large sampling, and that might not be the ideal approach to detect changes at individual stations. Despite that, our simplified method used both

the regional and local scale NDVI to suggest how the temperature relationship with NDVI depends on the heterogeneity of land-cover distribution across these scales.

3.4 Conclusions

With the purpose to assess the influence of local vegetation and regional scale geophysical controls on the spatial variability of near-surface temperature, we used 26 years of meteorological measurements from 52 conventional weather stations in Southeast Brazil. We used parsimonious statistical models of MLR and GAM, that helped to obtain the relationships of temperature with regional geophysical features (zonality, continentality, topography, and cloud cover) and local scale vegetation. The fitting of the average near-surface Tmax and Tmin showed the best overall performance with GAM, which used a single function to describe the combined effect of zonality and continentality, and for NDVI at local (300 m) and regional scale (3,000 m).

Our results were generally consistent with the knowledge established in the literature. However, some studies relied on larger areas (up to 10^4 m of radius) due to uncertainties in the information of station position, to attribute relationships of land-use and near-surface air temperature (Wang et al., 2017; Cao et al., 2019). As a novelty, our results fitted the dependency of geographical position and temperature with a non-linear method that represents the background climate conditions. Considering the additive property of the model we have on top of the spatial distribution, at least for Southeast Brazil, the land-use component, represented by NDVI, showed that, the variability of the near-surface air temperature is mostly correlated with local scale NDVI (300 m radius). Still, there is a combined effect of local vegetation with regional vegetation (3,000 m) that represents the complexity of heterogeneous surfaces.

In GAM, the independent variables that accounted for the variation of the annual average Tmin were geographical position and altitude, each with an amplitude of $\simeq 5$ ^{o}C, and the NDVI that contributed with an amplitude of $\simeq 3$ ^{o}C. Similarly, the variables that accounted for Tmax variation were geographical position and altitude, each with an amplitude of $\simeq 5$ ^{o}C, and cloud cover that contributed with an amplitude of $\simeq 3.5$ ^{o}C. The seasonality of the amplitude of each fitted function was relatively small across variables, except for the geographical position and altitude in the Tmin model, which was slightly

higher in winter compared to the annual mean.

Limitations of the proposed method are present, among them the fact that GAM is less generalizable than MLR, making it harder to extrapolate its results to other regions. Another limitation is the low density of stations available in the area that meet all required criteria for the analysis. The main geographical patterns are similar to those of Alvares et al. (2013) and Rodríguez-Lado et al. (2007), but a higher density is important for the characterization of the impact of land use heterogeneity. It would also help define more localized relationships with land use and topography, which would benefit from high-resolution data. For example, Local Climate Zones in the vicinity of the weather stations have information about architectural and urban morphology that are relevant to the surface energy budget (Stewart and Oke, 2012), as well as topographic aspect and slope (Sun and Zhang, 2016).

Finally, our results stress the need to clarify the causality of near-surface air temperature, at both the mean state and the temporal trends. Improving prediction of local temperature is key to adapting to global climate change and increasing urbanization. Especially in urban areas, it is recognized the need to achieve levels of climate resilience that assimilate current changes of the earth system. This issue can be driven partially through nature-based solutions (Mallick et al., 2021; McClymont et al., 2020), whereby public policies can design green spaces, using quantifying metrics that predict the possible effects distributed within in the urban space.

Chapter 4

Long-range temperature trends in Southeast Brazil weather stations, and urbanization impact

4.1 Introduction

The increase in global temperature of 0.85 ^{o}C [from 0.65 to 1.06 ^{o}C] between 1880 and 2012 is mainly attributed to the rise in anthropogenic emitted greenhouse gases (Gulev et al., 2021). However, there are significant regional differences in the recorded trends due to various causes like, for example, aerosols, internal variability, land cover, and topography (Doblas-Reyes et al., 2021). Climate simulations suggest a further increase in temperature and precipitation extremes, which are particularly impactful in more urbanized areas (Li et al., 2021). In Southeast Brazil (SEB), a region with more than 40 % of the population of Brazil (IBGE, 2018a), many different urban areas are affected by these changes, like the metropolitan region of São Paulo (MRSP), with more than 21 million inhabitants (SEADE, 2022). Examples of this vulnerability in MRSP are the hydrological drought in 2014 that compromised the human, industrial, and agricultural water supply on a regional scale (Coelho et al., 2015), and the 2001 drought (Cavalcanti and Kousky, 2004) that led to a shortage of energy supply.

Studies using long-range climate stations were made in SEB to detect and quantify trends in air temperature, which try to connect the results with the increase in greenhouse gases and local features like land cover changes and urbanization. From a regional perspective, Regoto et al. (2021) identified a statistically significant increase in both maximum and minimum temperature, Tmax and Tmin, respectively, with a larger trend for Tmax, and de Abreu et al. (2019) attributes a great part of this warming due to anthropogenic influence. In São Paulo, Sugahara et al. (2012) identified a significant trend of 0.26 ^{o}C

10 yr.$^{-1}$ for Tmax and 0.30 ^{o}C 10 yr.$^{-1}$ for Tmin between 1958 and 2004, which is larger than the regional average and the authors suggest an influence from urbanization. The increase in temperature is linked with changes in other climatic variables, like an increase in extreme precipitation events and a decrease in fog formation (Mühlig et al., 2020).

In other cities in the state of São Paulo, like Piracicaba and Campinas, Blain et al. (2009) reported a significant increase in minimum temperature for the entire series of Campinas and a no trend in Piracicaba between 1947 and 1976, with the authors suggesting influence from local radiative forcings like urbanization and internal variability as the main responsible for these differences. More recently, Alvares et al. (2022) identified an increase of 0.9 ^{o}C between 1917 and 2016 and a transition between climate types in Piracicaba, from a more subtropical climate to a more tropical one, with an increase in temperature in winter.

Most studies rely on the definition of a linear trend to calculate the changes in temperature, and other climate variables, in the last few decades. However, there is no reason to believe this is always the case, with the slope being highly dependent on the interval of the time series (Peng-Fei et al., 2015; Xu et al., 2021). Other methods, which are nonlinear, can be used to estimate the trend, like the Empirical Mode Decomposition (EMD), wavelets, and Generalized Additive Model (GAM) (Franzke, 2010; Hartmann et al., 2013; Simpson, 2018; Coelho et al., 2008), adding information about the time evolution of the trend. In the most recent Intergovernmental Panel on Climate Change (IPCC) report, for example, they computed global warming as a difference from 1850-1900 period instead of using linear regression, since the linear trend underestimates the current warming by at least 0.2 ^{o}C (Gulev et al., 2021).

Therefore, in this study, our objective is to quantify the trends in long-term climate stations from Southeast Brazil, and its variability in time for maximum and minimum temperature using a nonlinear method, more specifically GAM, and compare it with linear regression.

4.2 Materials and Methods

4.2.1 Weather stations

After an extensive automation process that took part in the last few decades, just a few conventional weather stations with a long record (> 50 years) were maintained (Alvares et al., 2022). Also, as shown by de Abreu et al. (2022) there is great spatial variability in temperature in Southeast Brazil from geographical position and altitude. Therefore, we focus our analysis on a small area in SEB for minimum and maximum temperatures for the stations in São Paulo. We used three stations in the city of São Paulo (iag, mrs, and cgn), one in Campinas and another one in Piracicaba (Table 4.1).

Table 4.1 - Geographical location of the analyzed stations, period in years, and the city where it is contained.

id	latitude (^{o})	longitude (^{o})	altitude (m)	city	period
iag	-23.651242	-46.622424	799	São Paulo	1933-2018
mrs	-23.496389	-46.62	802	São Paulo	1961-2018
cgn	-23.623106	-46.657749	785	São Paulo	1951-2018
cpn	-22.867442	-47.072914	667	Campinas	1901-2018
pcb	-22.708333	-47.633333	546	Piracicaba	1917-2018

The stations iag, mrs, and cgn are located in the city of São Paulo, where the dominant land cover type is urban infrastructure, but with some differences among the stations (Figure 4.1). iag is located in a vegetated area in the south of the city, while cgn is close to iag but with a predominance of urban infrastructure. mrs is located further north in São Paulo and in an urbanized area. In the case of cpn, the station is located in a vegetated region inside Instituto Agronômico de Campinas (IAC), with most of the city development to the south, similar to pcb station located in Piracicaba. The station cpn was moved in 1956, but according to Mello et al. (1994), the series can still be considered homogeneous, being used in other studies (Astolpho et al., 2004; Blain et al., 2009), so we decided to keep the series for the whole period that was made available.

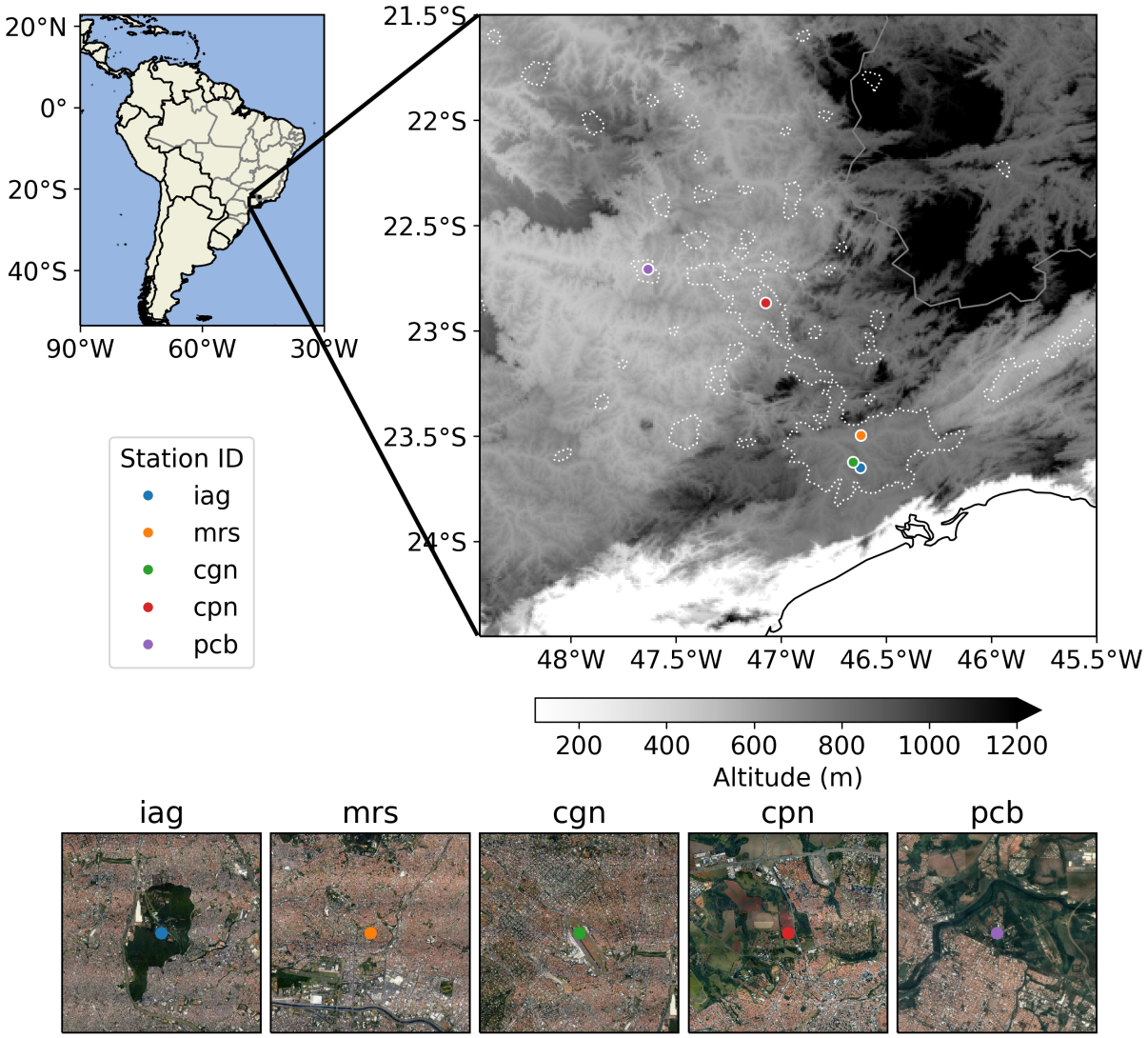

Figure 4.1: Geographical position of the weather stations used in this research. Shading represents the altitude in meters. The dotted white line delimits the urban area estimated from nighttime lights (`naturalearthdata.com`). In the bottom row is a 6 km × 6 km square surrounding each weather station with Google Earth image as the background.

The metropolitan region of São Paulo has great economic importance for the country, with the city development starting in the 1930s and gaining momentum after 1950. The state reached a contribution to the national GDP of 50 % in the 1970s with an extensive horizontal urban area (Silva and Fonseca, 2013). From the population data (Table 4.2) in Campinas and Piracicaba, there is a significant population increase at the beginning of the series and in the 1960∼1980s, which is more significant in Campinas, with a 76 % increase from 1970 to 1980. This increase in population, especially in Campinas, is directly related to the decentralization of the industry in São Paulo in the 1970s, causing a larger migration of people to the interior of the state (Baeninger, 2001).

Table 4.2 - Total population for the cities of São Paulo, Campinas, and Piracicaba, with the percentage of increase from one year to the other in parenthesis (IBGE, 2012, 2001, 1992, 1980, 1971, 1962, 1954, 1950, 1926, 1905, 1892, 1874).

Year	São Paulo	Campinas	Piracicaba
1872	31,385	-	-
1890	64,934 (106%)	-	-
1900	239,820 (269%)	67,694	25,374
1920	579,033 (141%)	115,602 (70%)	67,732 (166%)
1940	1,326,261 (129%)	129,940 (12%)	76,416 (12%)
1950	2,198,096 (65%)	152,547 (17%)	87,835 (14%)
1960	3,825,351 (74%)	217,219 (42%)	115,403 (31%)
1970	5,978,977 (56%)	375,864 (73%)	152,505 (32%)
1980	8,587,665 (43%)	664,559 (76%)	214,295 (40%)
1991	9,626,894 (12%)	846,084 (27%)	283,540 (32%)
2000	10,405,867 (8%)	969,396 (14%)	329,158 (16%)
2010	11,253,503 (8%)	1,080,113 (11%)	364,571 (10%)

4.2.2 Statistical Model

To estimate how the trend varies in time we used the Generalized Additive Model (GAM) and compare it with linear regression (LR). Both of them can be represented as (Hastie et al., 2009):

$$Y = \alpha + f(\text{Year}) + \epsilon \tag{4.1}$$

Where Y is the annual minimum or maximum temperature, α is the intercept term, $f(\text{Year})$ is a function of the year, and ϵ is the residual error which has a normal distribution with zero mean and constant variance. In the case of linear regression, $f(\text{Year}) = \beta\text{Year}$, and the estimate of β, $\hat{\beta}$ is estimated using ordinary least squares (OLS). The GAM is a generalization of Generalized Linear Models (GLM) and, therefore of linear regression, where $f(\text{Year})$ is a smooth function fitted using penalized least squares that penalizes too wiggly functions (Wood, 2017). The penalization is determined using the Generalized Cross Validation (GCV) so that even linear functions can be fitted using GAM, and we can estimate if the relationship between temperature and time is linear or non-linear using the estimated degrees of freedom (edf). We use the `mgcv` package for R (Wood, 2017) for fit the GAM.

We compare the linear regression with GAM based on the coefficient of determination (R^2) and Bayesian Information Criteria (BIC). We fit the series of both minimum and maximum temperature for the five selected stations and for the global average temperature. To compute the global annual timeseries we use the Berkeley Earth Data (Rohde and Hausfather, 2020), a gridded product which has data since 1850 and a grid spacing of $1^o \times 1^o$ in latitude and longitude. We use a weighted average of the area of each grid box, similar to Morice et al. (2012). Both the stations and global average are represented in the form of annual mean temperature anomalies, computed based on the 1981-2010 climatology.

4.3 Results and Discussion

We first applied LR and GAM to minimum and maximum temperature for the global average (Figure 4.2). For both Tmin and Tmax the timeseries (Figure 4.2a and b) have an increase from the beginning of the series up to 1940, when it is mostly constant until the 1980s, when the temperature starts to rise again in an almost linear way, with a warming rate of approximately 0.26 ^{o}C 10 yr^{-1} for Tmin and 0.30 ^{o}C 10 yr^{-1} for Tmax according to the GAM fitted curve. Little is known about the causes of the slowdown between 1940 and 1980, with influences from internal variability or increase in aerosols emission during this period being the most probable causes (Trenberth, 2015; Xu et al., 2022).

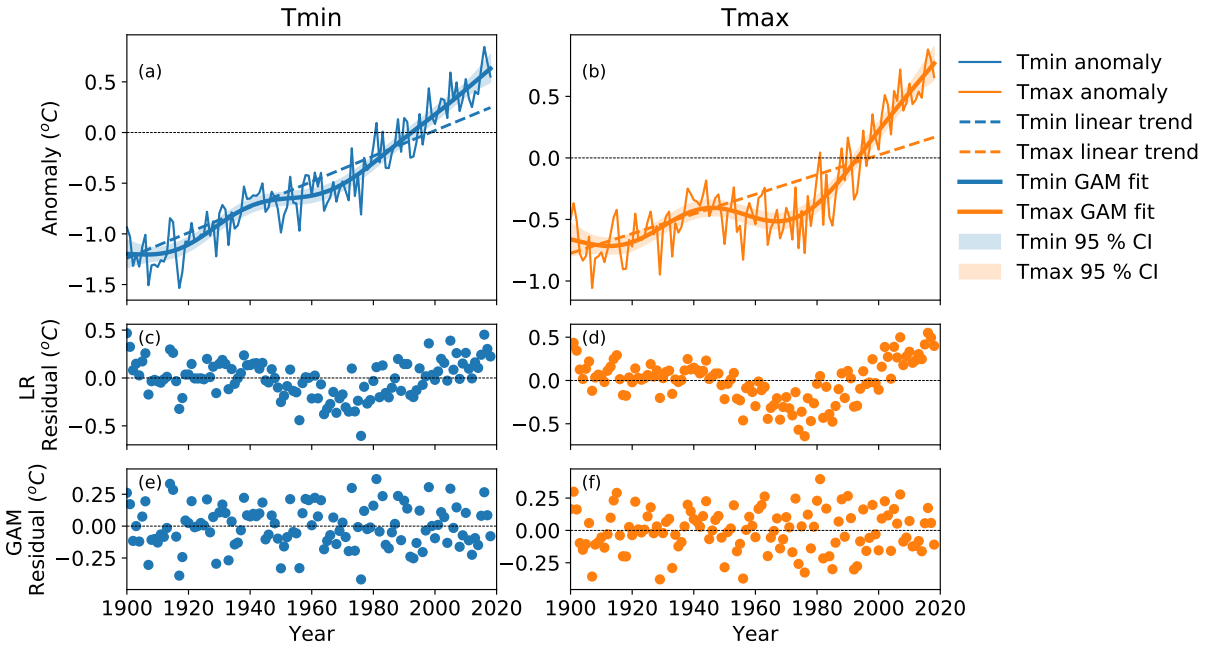

Figure 4.2: Annual mean temperature anomaly in ^{o}C for minimum temperature (Tmin) and maximum temperature (Tmax) for global average temperature with the fitted GAM and LR estimates (a and b); residuals of the linear fit (c and d), residuals of the GAM fit (e and f).

The linear regression fits a curve with a constant warming rate of 0.15 ^{o}C 10 yr^{-1} for Tmin and 0.10 ^{o}C 10 yr^{-1}, which is very different from the GAM estimate. The coefficient of determination R^2 is higher for GAM (Table 4.3), especially for Tmax with 86.4 % compared to 65.6 % from LR, with a lower BIC. The residuals from LR (Figure 4.2c and d) also show a more defined pattern of the linear fit overestimating temperature at the 1980s and underestimating at the end of the series, while GAM residuals (Figure 4.2e and f) are more randomly distributed around the zero, which is more consistent with the assumption of normal residues.

Table 4.3 - Coefficient of determination (R^2) in percentage, and Bayesian Information Criteria (BIC) for both linear regression (LR) and Generalized Additive Model (GAM) for each of the fitted timeseries (iag, mrs, cgn, cpn, pcb, and the global mean). The estimated degrees of freedom (edf) is also available.

	Variable	Global	iag	mrs	cgn	cpn	pcb
BIC (LR)		-33.25	112.11	76.10	99.68	122.99	215.04
BIC (GAM)		-56.27	110.82	76.10	98.29	121.49	114.77
R^2 (LR)	Tmin	87.4	78.3	62.0	71.6	81.3	34.0
R^2 (GAM)		91.3	79.7	62.0	74.0	84.1	56.4
edf		5.46	2.00	1.00	2.07	4.75	5.22
BIC (LR)		29.74	156.89	105.97	140.91	258.33	205.50
BIC (GAM)		-59.50	156.89	107.64	142.52	237.64	207.78
R^2 (LR)	Tmax	65.6	50.4	54.4	39.5	50.1	28.7
R^2 (GAM)		86.4	50.4	57.9	46.0	64.7	32.6
edf		5.43	1.00	2.58	3.22	5.22	2.73

We then fit the weather station's timeseries with LR and GAM and present the results in Figure 4.3. For minimum temperature, the trends in the city of São Paulo (iag, mrs, and cgn) fitted with GAM are almost the same as LR, especially in mrs. In iag and cgn, however, there are changes from the linear fit in the boundaries of the series, with GAM showing lower values at the start and end of the series compared to the LR. This means that the rate of change in Tmin is decreasing with time. When we compare the anomalies of the weather stations with the global mean (Figure 4.2a) we have a significant difference at the beginning, with the stations showing larger negative anomalies, which means that the stations in the city of São Paulo have a significantly higher average temperature in the reference period (1981-2010) than in the past when compared to the global average. For iag, for example, there is a difference of almost 1 ^{o}C in the 1930s. The R^2 (Table 4.3) is similar between the two methods, but higher for GAM, varying between 62 % in mrs and 79.7 % in iag, with a lower BIC.

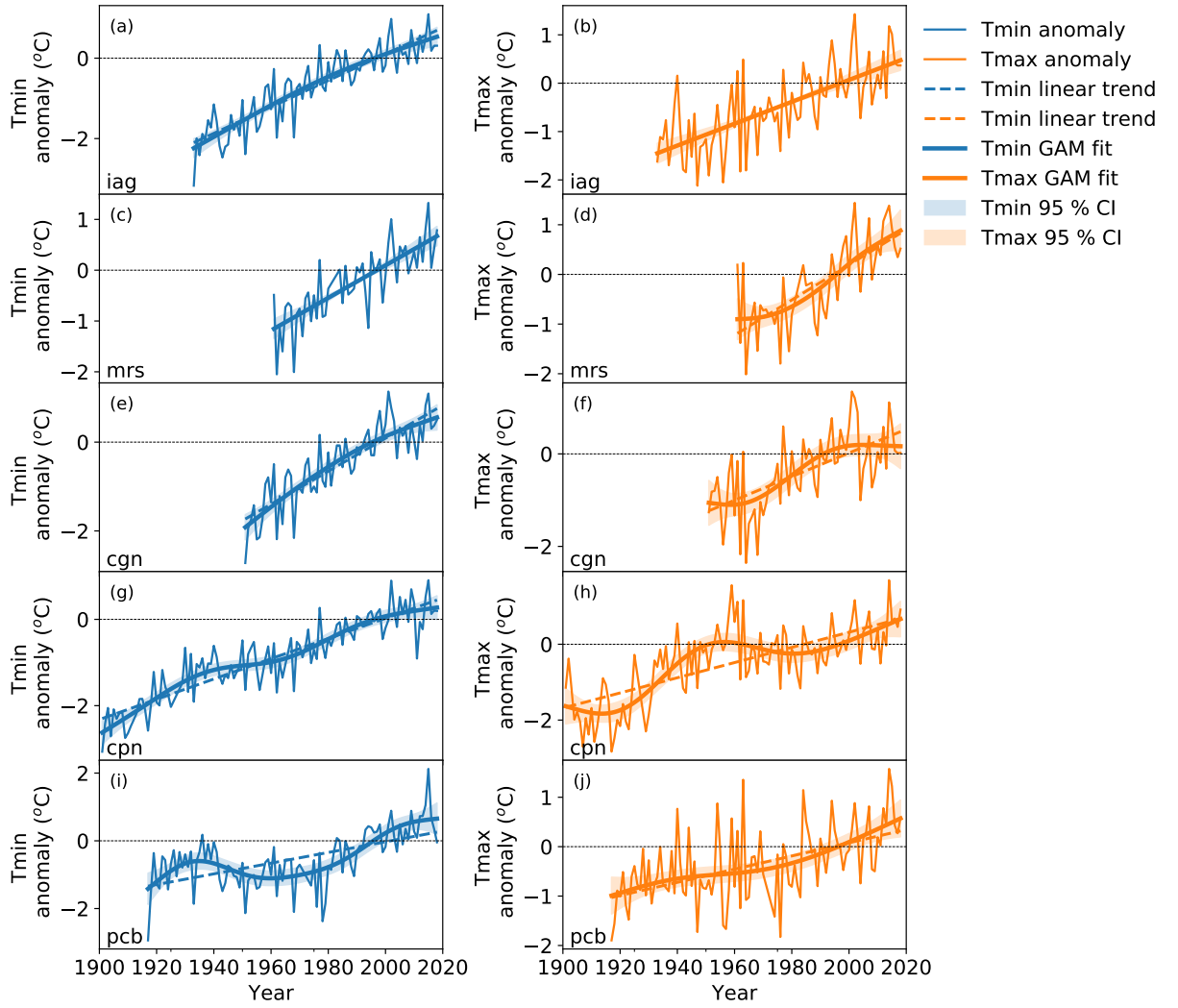

Figure 4.3: Annual temperature anomaly in ^{o}C for minimum temperature (Tmin) and maximum temperature (Tmax) for stations iag (a and b); mrs (c and d), cgn (e and f), cpn (g and h), pcb (i and j). The dashed line is the linear trend calculated with the ordinary least squares method, colored solid line represents the GAM fitted trend. The shaded colors represents the confidence interval (CI) of 95 % of the GAM fitted curve.

In the case of cpn and pcb we have longer timeseries, and the difference between the GAM and linear regression is more evident. Both series Tmin show an increase up until 1940s, when there is a slower warming rate up until 1970s in cpn and 1980s, which is when the the differences between LR and GAM are more evident, especially for pcb. After that, an increase of almost 1 ^{o}C is recorded for both series which starts to decline after 2000. The anomalies in pcb, and cpn weather stations have a more similar pattern to the global temperature than the stations in São Paulo, with an exception for the end of the series that shows more constant temperatures.

There are larger differences among the stations in the city of São Paulo for maximum

temperature. For example, both iag and mrs show an increasing trend with GAM, which is close to the LR, especially for iag, while cgn suggests that Tmax starts to level off after 1980. The linear trend, however, cannot capture this kind of difference, underestimating the warming rate from the 1970s to 1990s, and overestimating after that. The difference between the measured timeseries and the global average for iag is similar to Tmin, with significant warming at the beginning of the series.

In the stations cpn and pcb, located in the state's interior, there are also differences in the trends for Tmax between the GAM fitted curve and LR. In cpn there is an overestimation of the temperature anomalies by LR until the 1930s, and an underestimation between 1940 and 1980, when Tmax is more constant. After that, the temperature increases again by almost 1 ^{o}C in 40 years. For pcb there is a similar pattern, but the difference from LR is lower than cpn. The coefficient of determination has a significant improvement for cpn, from 50.1 % using LR to 64.7 % with GAM (Table 4.3), a lower BIC and an edf of 5.22, similar to the global average. However, for pcb there is a lower R^2 and higher BIC using GAM.

We then calculate the instantaneous trend for every decade, given by the derivative of the GAM fitted curve at the given point (Figure 4.4). For minimum temperature, the warming rate is more than double the regional average for the São Paulo stations (iag, mrs, and cgn) at the beginning of the series, with a maximum value of 0.50 ^{o}C 10 yr.$^{-1}$ in 1960 for cgn. The warming rate decreases with time in cgn and iag, closer to the global average of 0.26 ^{o}C 10 yr.$^{-1}$ in 2010. For cpn we have a different pattern, where the warming rate is especially higher than the global average in 1970 and 1980 reaching up to 0.34 ^oC 10 yr.$^{-1}$ in 1980. In pcb, Tmin has a significant increase between 1980 and 2000, with a maximum of 0.56 ^{o}C 10 yr.$^{-1}$.

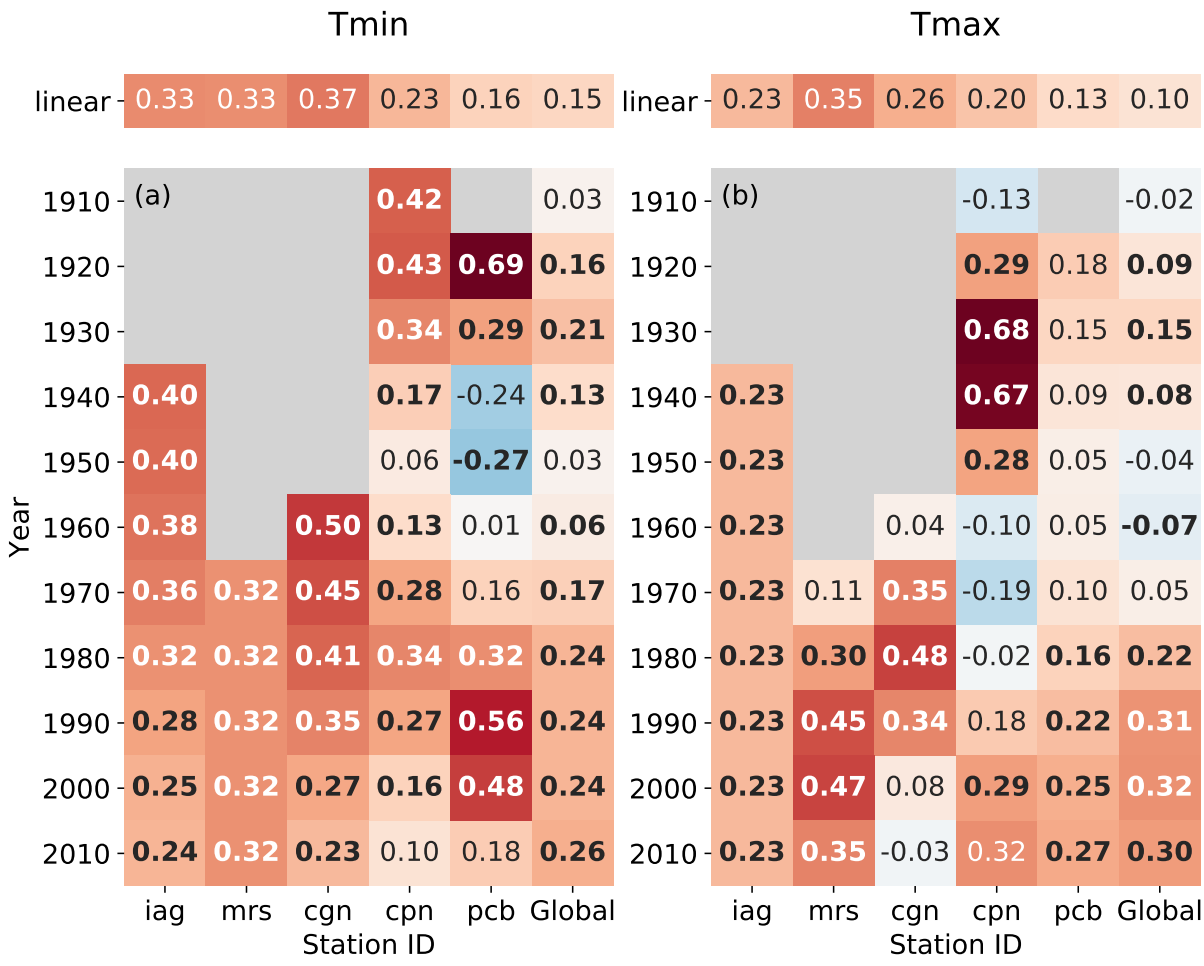

Figure 4.4: Instantaneous trend, given by the derivative of the GAM fitted curve at the given point in time, for each station and Southeast Brazil global average for minimum (a) and maximum temperature (b). The estimated value we calculated using GAM fitted curve, which is also compared with the linear fit. The bold letter shows where the derivative is statistically significantly different from zero, with a confidence interval (CI) of 95 %.

de Abreu et al. (2022) shows that the vegetation in Southeast Brazil explains an important part of the spatial variability of the average temperature, especially for Tmin, with a high degree of complexity depending on the spatial heterogeneity of land cover. The minimum temperature is mainly affected by urbanization with an increased stored heat during the day that is slowly released to the air during the night, increasing Tmin (Oke et al., 2017). Also, as discussed in previous studies (Sugahara et al., 2012; de Lima and Rueda, 2018), there is a likely impact of urbanization in São Paulo trends for both temperature and precipitation. This could be the effect that we see in iag, mrs and cgn, since there is significant urbanization in the decades prior to 1980 and a considerable warming rate that is not observed in the global average temperature and in cpn and pcb. After

that, the population increases in São Paulo at a much slower rate (Table 4.2), land cover is mostly the same, with urban infrastructure dominating the area (Figure 4.5), and the instantaneous trend for Tmin is closer to the global average. According to Jones et al. (2008) the differences in trends between rural and urban sites tend to decrease over time after the urban area is established.

In cpn and pcb, there was much slower warming between 1940 and 1980, comparable to the global average. In the following decades, the increase in temperature at a higher rate of change than the global average could also suggest the influence of urbanization since the urban expansion was most significant at a later time in Campinas and Piracicaba than in São Paulo (Table 4.2). Blain et al. (2009) even suggests that most of the increase in Tmin in Campinas might be due to localized effects. After the 2000s, there was also a decrease in temperature trend in those stations. Many studies call the period between 1998-2012, when trends decreased, the "global warming hiatus" (Trenberth, 2015; Medhaug et al., 2017), that could have several possible causes such as radiative forcing or internal variability, most specifically the negative phase of the Pacific Decadal Oscillation (PDO). A Pearson correlation between the PDO and Tmin series, using the annual mean, of 0.49 and 0.38 is statistically significant between cpn and pcb weather stations, respectively, suggesting that they might be related.

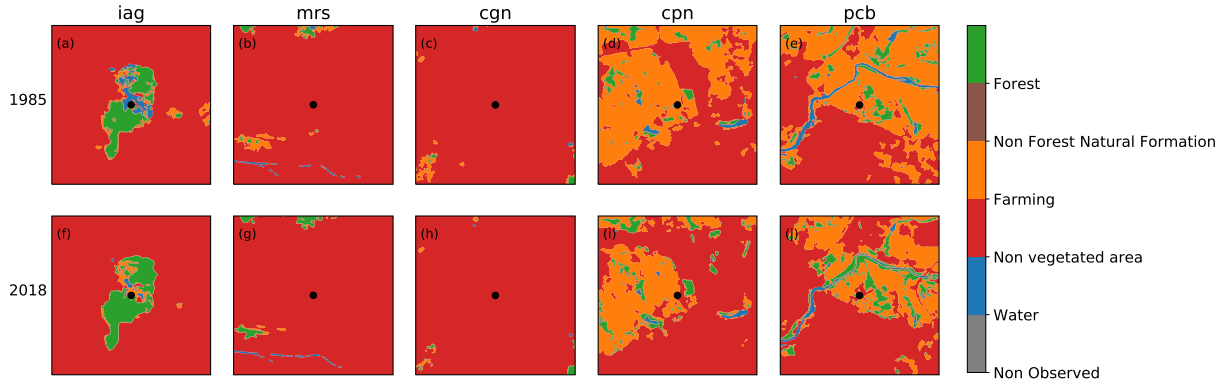

Figure 4.5: land cover from MapBiomas version 5.0 classification (Souza et al., 2020) in a 6 km × 6 km box around each weather station coordinates, for the years of 1985 (a-e) and 2018 (f-j).

For maximum temperature, the global average temperature trend shows a significant increase after 1980 with the GAM method, reaching 0.30 ^{o}C 10 yr.$^{-1}$ in 2010, while the linear trend is much lower, 0.10 ^{o}C 10 yr.$^{-1}$. We see this same pattern of increasing warming rate also locally in cpn, and pcb. For cgn, however, there is a decrease in the temperature

trend over time, with -0.03 ^{o}C 10 yr.$^{-1}$ in 2010, which is not statistically different from zero. In iag the temperature trend fitted by the GAM is linear with the same warming rate, while for mrs there is an increase in warming up until 2000 with 0.47 ^{o} 10 yr.$^{-1}$ and decreases to 0.35^o 10 yr.$^{-1}$. The effect of urbanization in Tmax is more complex than for Tmin, with no significant warming or even cooling in some areas (Oke et al., 2017; de Abreu et al., 2022; Kalnay and Cai, 2003). In São Paulo, for example, Freitas et al. (2007) shows an interaction between the urban heat island effect and sea breeze circulation, and Umezaki et al. (2020) shows a higher spatial heterogeneity of the urban heat island effect during the afternoon than during the night. During the day, cloud cover, and therefore precipitation, plays an important role in temperature spatial variability (de Abreu et al., 2022), with a decrease of temperature in areas with a larger cloud cover, due to the increased albedo cooling effect. From a regional perspective, Zilli et al. (2019) shows a poleward shift of the South Atlantic Convergence Zone (SACZ) in recent decades that affects the cloud cover patterns, and Regoto et al. (2021) shows that the stations in São Paulo are in a transition area where stations show negative precipitation trends to the north and positive trends to the south.

4.4 Conclusions

Many studies use a prior definition of a linear trend to calculate the increase in temperature in the last few decades, which is not always the best fit. Therefore, we used the Generalized Additive Model (GAM) to evaluate the temporal temperature variability in long-term weather stations in São Paulo state. We used five stations, three located in the city of São Paulo (iag, mrs, and cgn). The population in 2010 was over 11 million inhabitants, and the urbanization process was most expressive in the 1950s. We also used one station in Campinas (cpn), and another one in Piracicaba (pcb), where the urbanization process was most expressive after the 1970s.

We used the non-linear estimate to compare with the linear fit in the selected stations, where for Tmin, the stations in São Paulo had an almost linear trend, with a decrease in instantaneous trend over time, from 0.40 ^{o}C 10 yr.$^{-1}$ in 1940 to 0.24 ^{o}C 10 yr.$^{-1}$ in 2010, for iag station, close to the global average (0.26 ^{o}C 10 yr.$^{-1}$ in 2010). Even though no objective method was used to attribute the changes to land cover only, the observed

changes from regional trends agree with the literature on the impact of urbanization on the increase of minimum temperature (Oke et al., 2017; Kalnay and Cai, 2003; de Abreu et al., 2022). In Campinas and Piracicaba the variability is more closely related to the global average, with a much slower warming during 1940 and 1980 than in São Paulo and a subsequent increase in temperature.

The departure from the linear trend is even more apparent for maximum temperature in a few stations like mrs, cgn, cpn, and pcb. Stations like pcb, mrs and iag have a similar pattern of increasing maximum temperature trend over time, which is different from cgn and cpn where temperature anomalies are almost constant after a particular year. Those differences might have other causes, like cloud cover and precipitation, since the effect of urbanization in Tmax is not as evident as for Tmin.

We should consider other possible causes for the observed differences among the weather stations, like other local features, in which more extensive work should be done. However, in the current study, we addressed the limitations of using a linear trend to compute the trend in the selected stations and how a non-linear estimate can reveal more details about temperature changes, like urbanization impacts.

Chapter 5

Conclusions and future work

5.1 Conclusions

Given the climate projections of an increase in temperature and precipitation extremes at the global scale, the current work aimed to bridge different aspects of temperature variability and attribution in Southeastern Brazil, from local to regional scale analysis. At the regional scale, we used the concept of Detection and Attribution (D&A) and the novel statistical model from Ribes et al. (2017) to attribute the observed trend in average temperature of 1.1 ^{o}C in 50 years. The results showed that we can not explain the recent warming only with internal variability and natural forcing, and anthropogenic greenhouse gases being the main contributor. We should also point that model error is the main source of uncertainty in the estimation of the parameters, with more than half of the error coming from it.

Even tough greenhouse gases are the major source of the observed warming in Southeastern Brazil, we noted pronounced differences in the trends of individual weather stations, which led us to analyze the main regional geographical controls of the average temperature. By using a non-linear model, the Generalized Additive Model (GAM) and 26 years of data from 52 weather stations, we showed that geographical position and altitude accounted for the largest spatial variability of $\simeq$ 5.0 ^{o}C for both minimum and maximum temperature, Tmin and Tmax, respectively. For Tmin however, NDVI has a statistically significant contribution of approximately 3 ^{o}C, while for Tmax cloud cover is more important. Also, we suggest there is a heterogeneity in the Normalized Difference Vegetation Index (NDVI) response, that needs to account for regional and local NDVI.

Since land cover is an important sources of local variability in temperature for Southeast

Brazil, and as reported by other studies (Oke et al., 2017; Kalnay and Cai, 2003; Kagawa-Viviani and Giambelluca, 2020) we decided to look at local long range temperature trends in the state of São Paulo, in the cities of São Paulo, Campinas, and Piracicaba. Using the GAM we give more detail to the temporal variability of the temperature anomalies than by using a linear fit. For Tmin a rapid increase in the magnitude of the multidecadal trend is observed for São Paulo stations at beginning of the series, reaching up to 0.40 ^{o}C 10 yr.$^{-1}$ in 1940 and slowly decreasing in recent decades. In Piracicaba and Campinas, Tmin trends are more consistent with the global average, with a period of much slower warming during 1940 and 1980 than in São Paulo, and a subsequent increase in temperature. This is consistent with the urbanization process that started first in São Paulo and a migration further inland in the 1970s due to industry decentralization policies (Baeninger, 2001). For maximum temperature the effect is not so evident, and other sources of variability might be of interest like cloud cover and precipitation.

In a more broad perspective, in this study we were able to show that greenhouse gas concentration is the main driver of regional temperature trends in Southeast Brazil. These changes are similar to the global temperature changes, which suggests that impacts of anthropogenic global warming is becoming more relevant at smaller scales. However, other sources of temperature variation should be accounted for, like land cover and vegetation, which has complex spatial distribution but are relevant at the local scale average temperature, and trend, and are key to adapt urban areas to the effects of human induced climate change.

5.2 Future work

The results found in this research suggest important future work that could be done:

- The detection and attribution of regional trends in other variables like precipitation, and indices related to extreme events. This work could also be expanded to other areas of Brazil, and South America;
- Expand the attribution analysis with data from the Coupled Model Intercomparison Project Phase 6 (CMIP6);
- Incorporate information about the Local Climate Zones (LCZ) in the analysis of

temperature dependency with land use;

- High resolution long-range dynamic simulations in metropolitan areas of Southeast Brazil to evaluate the impact of different land use types in temperature and precipitation patterns;

Bibliography

Abreu R. C., Cunningham C. Rudorff C. M., Rudorff N., Abatan A. A., Dong B., Lott F. C., Tett S. F. B., Sparrow S. N., Contribution of anthropogenic climate change to April-May 2017 heavy precipitation over the Uruguay River basin, Bull. Amer. Meteor. Soc., 2018

Alexandersson H., Moberg A., Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends, International Journal of Climatology: A Journal of the Royal Meteorological Society, 1997, vol. 17, p. 25

Allen M. R., Stott P., Estimating signal amplitudes in optimal fingerprinting, Part I: Theory, Climate Dynamics, 2003, vol. 21, p. 477

Allen M. R., Tett S. F., Checking for model consistency in optimal fingerprinting, Climate Dynamics, 1999, vol. 15, p. 419

Alvares C. A., Sentelhas P. C., Dias H. B., Southeastern Brazil inland tropicalization: Köppen system applied for detecting climate change throughout 100 years of meteorological observed data, Theoretical and Applied Climatology, 2022, vol. 149, p. 1431

Alvares C. A., Stape J. L., Sentelhas P. C., de Moraes Gonçalves J. L., Modeling monthly mean air temperature for Brazil, Theoretical and applied climatology, 2013, vol. 113, p. 407

Ambrizzi T., Ferraz S. E., An objective criterion for determining the South Atlantic Convergence Zone, Frontiers in Environmental Science, 2015, vol. 3, p. 23

Astolpho F., Camargo M. B. P. d., Bardin L., Probabilidades mensais e anuais de ocorrência

de temperaturas mínimas do ar adversas à agricultura na região de Campinas (SP), de 1891 a 2000, Bragantia, 2004, vol. 63, p. 141

Baeninger R., Região Metropolitana de Campinas: expansão e consolidação do urbano paulista, Migração e ambiente nas aglomerações urbanas, 2001, vol. 1, p. 319

Beckers J.-M., Rixen M., EOF calculations and data filling from incomplete oceanographic datasets, Journal of Atmospheric and oceanic technology, 2003, vol. 20, p. 1839

Bhang K. J., Anomalous variations of NDVI for a nonvegetated urban industrial area of Gumi, Korea, Am J Remote Sensing, 2014, vol. 2, p. 44

Bindoff N., Stott P., AchutaRao K., Allen M., Gillett N., Gutzler D., Hansingo K., Hegerl G., Hu Y., Jain S., Mokhov I., Overland J., Perlwitz J., Sebbari R., Zhang X., , 2013 Detection and Attribution of Climate Change: from Global to Regional. Cambridge University Press Cambridge, United Kingdom and New York, NY, USA pp 867–952

Blain G. C., Picoli M. C. A., Lulu J., et al., Análises estatísticas das tendências de elevação nas séries anuais de temperatura mínima do ar no Estado de São Paulo, Bragantia, 2009, vol. 68, p. 807

Camargo M. B. P. d., The impact of climatic variability and climate change on arabic coffee crop in Brazil, Bragantia, 2010, vol. 69, p. 239

Camilloni I., Barros V., On the urban heat island effect dependence on temperature trends, Climatic Change, 1997, vol. 37, p. 665

Cao W., Huang L., Liu L., Zhai J., Wu D., Overestimating impacts of urbanization on regional temperatures in developing megacity: Beijing as an example, Advances in Meteorology, 2019, vol. 2019

Cavalcanti I. F., Kousky V. E., Drought in Brazil during Summer and Fall 2001 and associated atmospheric circulation features, Revista Climanálise Ano, 2004, vol. 2

Ceppi P., Scherrer S. C., Fischer A. M., Appenzeller C., Revisiting Swiss temperature trends 1959–2008, International Journal of Climatology, 2012, vol. 32, p. 203

Coelho C., Ferro C., Stephenson D., Steinskog D., Methods for exploring spatial and temporal variability of extreme events in climate data, Journal of Climate, 2008, vol. 21, p. 2072

Coelho C. A., de Oliveira C. P., Ambrizzi T., Reboita M. S., Carpenedo C. B., Campos J. L. P. S., Tomaziello A. C. N., Pampuch L. A., de Souza Custódio M., Dutra L. M. M., et al., The 2014 southeast Brazil austral summer drought: regional scale mechanisms and teleconnections, Climate Dynamics, 2015, pp 1–16

Dalagnol R., Gramcianinov C. B., Crespo N. M., Luiz R., Chiquetto J. B., Marques M. T., Neto G. D., de Abreu R. C., Li S., Lott F. C., et al., Extreme rainfall and its impacts in the Brazilian Minas Gerais state in January 2020: Can we blame climate change?, Climate Resilience and Sustainability, 2022, vol. 1, p. e15

de Abreu R., Tett S., Schurer A., Rocha H., Attribution of Detected Temperature Trends in Southeast Brazil, Geophysical Research Letters, 2019, vol. 46, p. 8407

de Abreu R. C., Hallak R., da Rocha H. R., Effects of Local Vegetation and Regional Controls in Near-Surface Air Temperature for Southeastern Brazil, Atmosphere, 2022, vol. 13, p. 1758

de Lima G. N., Rueda V. O. M., The urban growth of the metropolitan area of Sao Paulo and its impact on the climate, Weather and climate extremes, 2018, vol. 21, p. 17

Doblas-Reyes F. J., Sorensson A. A., Almazroui M., Dosio A., Gutowski W. J., Haarsma R., Hamdi R., Hewitson B., Kwon W.-T., Lamptey B. L., Maraun D., Stephenson T. S., Takayabu I., Terray L., Turner A., Zuo Z., , 2021 in Masson-Delmotte V., Zhai P., Pirani A., Connors S. L., Pean C., Berger S., Caud N., Chen Y., Goldfarb L., Gomis M. I., Huang M., Leitzell K., Lonnoy E., Matthews J. B. R., Maycock T. K., Waterfield T., Yelekci O., Yu R., Zhou B., eds, , Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press

El Kenawy A., López-Moreno J. I., Stepanek P., Vicente-Serrano S. M., An assessment of the role of homogenization protocol in the performance of daily temperature series and

trends: application to northeastern Spain, International Journal of Climatology, 2013, vol. 33, p. 87

Farr T. G., Rosen P. A., Caro E., Crippen R., Duren R., Hensley S., Kobrick M., Paller M., Rodriguez E., Roth L., et al., The shuttle radar topography mission, Reviews of geophysics, 2007, vol. 45

Foss M., Chou S. C., Seluchi M. E., Interaction of cold fronts with the Brazilian Plateau: a climatological analysis, International Journal of Climatology, 2017, vol. 37, p. 3644

Franzke C., Long-range dependence and climate noise characteristics of Antarctic temperature data, Journal of Climate, 2010, vol. 23, p. 6074

Freitas E. D., Rozoff C. M., Cotton W. R., Dias P. L. S., Interactions of an urban heat island and sea-breeze circulations during winter over the metropolitan area of São Paulo, Brazil, Boundary-Layer Meteorology, 2007, vol. 122, p. 43

Gulev S., Thorne P., Ahn J., Dentener F., Domingues C., Gong S. G. D., Kaufman D., Nnamchi H., Rivera J. Q. J., Sathyendranath S., Smith S., Trewin B., von Schuckmann K., Vose R., Allan R., Collins B., Turner A., Hawkins E., , 2021 in , Climate Change 2021: The Physical Science Basis: Working Group I contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press

Hartmann D., KleinÂ Tank A., Rusticucci M., Alexander L., Broönnimann S., Charabi Y., Dentener F., Dlugokencky E., Easterling D., Kaplan A., Soden B., Thorne P., Wild M., Zhai P., , 2013 Observations: Atmosphere and Surface. Cambridge University Press Cambridge, United Kingdom and New York, NY, USA p. 159â254

Hastie T., Tibshirani R., Friedman J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009

Hegerl G. C., voN SToRcH H., Hasselmann K., Santer B. D., Cubasch U., Jones P. D., Detecting greenhouse-gas-induced climate change with an optimal fingerprint method, Journal of Climate, 1996, vol. 9, p. 2281

Henn B., Raleigh M. S., Fisher A., Lundquist J. D., A comparison of methods for filling gaps in hourly near-surface air temperature data, Journal of Hydrometeorology, 2013, vol. 14, p. 929

Hicks B. B., Callahan W. J., Hoekzema M. A., On the heat islands of Washington, DC, and New York City, NY, Boundary-layer meteorology, 2010, vol. 135, p. 291

Hoesly R. M., Smith S. J., Feng L., Klimont Z., Janssens-Maenhout G., Pitkanen T., Seibert J. J., Vu L., Andres R. J., Bolt R. M., et al., Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS), Geoscientific Model Development (Online), 2018, vol. 11

Hunt J. D., Stilpen D., de Freitas M. A. V., A review of the causes, impacts and solutions for electricity supply crises in Brazil, Renewable and Sustainable Energy Reviews, 2018, vol. 88, p. 208

Hurrell J. W., Holland M. M., Gent P. R., Ghan S., Kay J. E., Kushner P., Lamarque J.-F., Large W. G., Lawrence D., Lindsay K., et al., The community earth system model: a framework for collaborative research, Bulletin of the American Meteorological Society, 2013, vol. 94, p. 1339

IBGE, 1874 Recenseamento do Brazil em 1872 <https://biblioteca.ibge.gov.br/biblioteca-catalogo?id=225477&view=detalhes> Acessado: 2022-11

IBGE, 1892 Recenseamento geral da Republica dos Estados Unidos do Brazil em 31 de dezembro de 1890, Comarca de Palmas, Estado do Parana = Recensement General de la Republique des Etats Unis du Bresil au 31 decembre de 1890, Comarca de Palmas, Etat du Parana / Minis) <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=216866> Acessado: 2022-11

IBGE, 1905 Synopse do recenseamento de 31 de dezembro de 1900 = Precis du recensement 31 decembro 1900 / Directoria Geral de Estatistica) <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?id=225474&view=detalhes> Acessado: 2022-11

IBGE, 1926 Recenseamento de 1920 (4^o Censo geral da população e 1^o da agricultura e das industrias) <https://biblioteca.ibge.gov.br/visualizacao/livros/liv6478.pdf> Acessado: 2022-11

IBGE, 1950 Recenseamento geral do Brasil 1940 : censo demográfico : censos

econômicos <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?id=765&view=detalhes> Acessado: 2022-11

IBGE, 1954 Censo demográfico : 1950 <https://biblioteca.ibge.gov.br/?view=detalhes&id=767> Acessado: 2022-11

IBGE, 1962 Censo demográfico : 1960 <https://biblioteca.ibge.gov.br/index.php/bibliotecacatalogo?id=768&view=detalhes> Acessado: 2022-11

IBGE, 1971 Censo demográfico : 1970 <https://biblioteca.ibge.gov.br/biblioteca-catalogo.html?id=769&view=detalhes> Acessado: 2022-11

IBGE, 1980 Censo demográfico : 1980 : dados gerais, migração, instrução, fecundidade, mortalidade <https://biblioteca.ibge.gov.br/index.php/bibliotecacatalogo?view=detalhes&id=772> Acessado: 2022-11

IBGE, 1992 Censo demográfico 1991 : resultados preliminares / IBGE <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=283450> Acessado: 2022-11

IBGE, 2001 Sinopse preliminar do censo demográfico : 2000 <https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7308> Acessado: 2022-11

IBGE, 2012 Censo demográfico de 2010: Educação e Deslocamento - Resultados da Amostra <https://censo2010.ibge.gov.br/resultados.html> Acessado: 2022-06

IBGE, 2018a Etimativas da População residente no Brasil e unidades da federação com data de referência em 1Âº de julho de 2018 Retrieved from: ftp://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2018/estimativa_dou_2018_20181019.pdf

IBGE, 2018b Produto Interno Bruto pela ótica da renda, Brasil, Grandes Regiões e as Unidades da Federação, pelas óticas da renda e da produção - 2010-2016 Retrived from: ftp://ftp.ibge.gov.br/Contas_Regionais/2016/xls/PIB_Otica_da_Renda.xlsx

Jones P., Lister D., Li Q., Urbanization effects in large-scale temperature records, with an emphasis on China, Journal of Geophysical Research: Atmospheres, 2008, vol. 113

Jones P., Lister D., Osborn T., Harpham C., Salmon M., Morice C., Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010, Journal of Geophysical Research: Atmospheres, 2012, vol. 117

Kagawa-Viviani A., Giambelluca T., Spatial patterns and trends in surface air temperatures and implied changes in atmospheric moisture across the Hawaiian Islands, 1905–2017, Journal of Geophysical Research: Atmospheres, 2020, vol. 125, p. e2019JD031571

Kalnay E., Cai M., Impact of urbanization and land-use change on climate, Nature, 2003, vol. 423, p. 528

Karoly D. J., Stott P. A., Anthropogenic warming of Central England temperature, Atmospheric Science Letters, 2006, vol. 7, p. 81

Kay J., Deser C., Phillips A., Mai A., Hannay C., Strand G., Arblaster J., Bates S., Danabasoglu G., Edwards J., et al., The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, Bulletin of the American Meteorological Society, 2015, vol. 96, p. 1333

Kirchner M., Faus-Kessler T., Jakobi G., Leuchner M., Ries L., Scheel H.-E., Suppan P., Altitudinal temperature lapse rates in an Alpine valley: trends and the influence of season and weather patterns, International journal of climatology, 2013, vol. 33, p. 539

Kutner M. H., Nachtsheim C. J., Neter J., Li W., et al., Applied linear statistical models. vol. 5, McGraw-Hill Irwin Boston, 2005

Lambin E. F., Ehrlich D., The surface temperature-vegetation index space for land cover and land-cover change analysis, International journal of remote sensing, 1996, vol. 17, p. 463

Ledoit O., Wolf M., A well-conditioned estimator for large-dimensional covariance matrices, Journal of multivariate analysis, 2004, vol. 88, p. 365

Lee J.-Y., Marotzke J., Bala G., Cao L., Corti S., Dunne J., Engelbrecht F., Fischer E., Fyfe J., Jones C., Maycock A., Mutemi J., Ndiaye O., Panickal S., Zhou T., , 2021 Future

Global Climate: Scenario-Based Projections and Near-Term Information. Cambridge University Press Cambridge, United Kingdom and New York, NY, USA p. 553â672

Li C., Zwiers F., Zhang X., Li G., Sun Y., Wehner M., Changes in annual extremes of daily temperature and precipitation in CMIP6 models, Journal of Climate, 2021, vol. 34, p. 3441

Li Y., Zeng Z., Zhao L., Piao S., Spatial patterns of climatological temperature lapse rate in mainland China: A multi–time scale investigation, Journal of Geophysical Research: Atmospheres, 2015, vol. 120, p. 2661

Lott F. C., Christidis N., Ciavarella A., Stott P. A., The effect of human land use change in the Hadley Centre attribution system, Atmospheric Science Letters, 2020

McClymont K., Cunha D. G. F., Maidment C., Ashagre B., Vasconcelos A. F., de Macedo M. B., Dos Santos M. F. N., Júnior M. N. G., Mendiondo E. M., Barbassa A. P., et al., Towards urban resilience through Sustainable Drainage Systems: A multi-objective optimisation problem, Journal of Environmental Management, 2020, vol. 275, p. 111173

Mallick S. K., Das P., Maity B., Rudra S., Pramanik M., Pradhan B., Sahana M., Understanding future urban growth, urban resilience and sustainable development of small cities using prediction-adaptation-resilience (PAR) approach, Sustainable Cities and Society, 2021, vol. 74, p. 103196

Manoli G., Fatichi S., Schläpfer M., Yu K., Crowther T. W., Meili N., Burlando P., Katul G. G., Bou-Zeid E., Magnitude of urban heat islands largely explained by climate and population, Nature, 2019, vol. 573, p. 55

Marengo J. A., Mudanças climáticas globais e regionais: Avaliação do clima atual do Brasil e projeções de cenários climáticos do futuro, Revista Brasileira de Meteorologia, 2001, vol. 16, p. 01

Martin T. C., da Rocha H. R., Joly C. A., Freitas H. C., Wanderley R. L., da Silva J. M., Fine-scale climate variability in a complex terrain basin using a high-resolution weather station network in southeastern Brazil, International Journal of Climatology, 2019, vol. 39, p. 218

Medhaug I., Stolpe M. B., Fischer E. M., Knutti R., Reconciling controversies about the 'global warming hiatus', Nature, 2017, vol. 545, p. 41

Meek D., Hatfield J., Data quality checking for single station meteorological databases, Agricultural and Forest Meteorology, 1994, vol. 69, p. 85

Mello M. d. A., Pedro Junior M., Ortolani A., Alfonsi R., Chuva e temperatura: cem anos de observações em Campinas. vol. 154, IAC Campinas, Brazil, 1994

Menne M. J., Williams Jr C. N., Homogenization of temperature series via pairwise comparisons, Journal of Climate, 2009, vol. 22, p. 1700

Mitchell J., Karol D., Hegerl G., Zwiers F., Allen M., Marengo J., Barros V., Berliner M., Boer G., Crowley T., Folland C., Free M., Gillett N., Groisman P., Haigh J., Hasselmann K. Jones P., Kandlikar M., Kharin V., Kheshgi H., Knutson T., MacCracken M., Mann M., North G., Risbey J., Robock A., Santer B., Schnur R., Schönwiese C., Sexton D., Stott P., Tett S., Vinnikov K., Wigley T., , 2001 Detection of Climate Change and Attribution of Causes. Cambridge University Press Cambridge, United Kingdom and New York, NY, USA pp 695–778

Morice C. P., Kennedy J. J., Rayner N. A., Jones P. D., Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, Journal of Geophysical Research: Atmospheres, 2012, vol. 117

Mühlig A. C., Klemm O., Gonçalves F. L. T., Fog, Temperature and Air Quality Over the Metropolitan Area of São Paulo: a Trend Analysis from 1998 to 2018, Water, Air, & Soil Pollution, 2020, vol. 231, p. 1

Nobre C., Young A. F., Saldiva P. H. N., Orsini J. A. M., Nobre A. D., Ogura A. T., Thomaz O., Párraga G. O. O., da Silva G. C. M., Valverde M., et al., Vulnerability of Brazilian megacities to climate change: the São Paulo Metropolitan Region (RMSP), CLIMATE CHANGE IN BRAZIL, 2011, p. 197

Oke T. R., Mills G., Voogt J., Urban climates. Cambridge University Press, 2017

Oliveira A. P., Bornstein R. D., Soares J., Annual and diurnal wind patterns in the city of São Paulo, Water, Air, & Soil Pollution: Focus, 2003, vol. 3, p. 3

Oliveira A. P. d., Silva Dias P. L. d., Aspectos observacionais da brisa marítima em São Paulo, II CBM: Anais 1980-2006, 1982

Otto F. E., Haustein K., Uhe P., Coelho C. A., Aravequia J. A., Almeida W., King A., Coughlan de Perez E., Wada Y., Jan van Oldenborgh G., et al., Factors other than climate change, main drivers of 2014/15 water shortage in southeast Brazil, Bulletin of the American Meteorological Society, 2015, vol. 96, p. S35

Peng-Fei L., Xiao-Li F., Juan-Juan L., Historical trends in surface air temperature estimated by ensemble empirical mode decomposition and least squares linear fitting, Atmospheric and Oceanic Science Letters, 2015, vol. 8, p. 10

Pereira V. R., Blain G. C., Avila A. M. H. d., Pires R. C. d. M., Pinto H. S., Impacts of climate change on drought: changes to drier conditions at the beginning of the crop growing season in southern Brazil, Bragantia, 2017, vol. 77, p. 201

Pettit A., A non-parametric approach to the change-point problem, Applied statistics, 1979, vol. 28, p. 126

Reboita M. S., Gan M. A., Rocha R. P. d., Ambrizzi T., Regimes de precipitação na América do Sul: uma revisão bibliográfica, Revista brasileira de meteorologia, 2010, vol. 25, p. 185

Regoto P., Dereczynski C., Chou S. C., Bazzanela A. C., Observed changes in air temperature and precipitation extremes over Brazil, International Journal of Climatology, 2021, vol. 41, p. 5125

Ribes A., Azaïs J.-M., Planton S., Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate, Climate Dynamics, 2009, vol. 33, p. 707

Ribes A., Planton S., Terray L., Application of regularised optimal fingerprinting to attribution. Part I: Method, properties and idealised analysis, Climate dynamics, 2013, vol. 41, p. 2817

Ribes A., Zwiers F. W., Azais J.-M., Naveau P., A new statistical approach to climate change detection and attribution, Climate Dynamics, 2017, vol. 48, p. 367

Rodríguez-Lado L., Sparovek G., Vidal-Torrado P., Dourado-Neto D., Macías-Vázquez F., Modelling air temperature for the state of São Paulo, Brazil, Scientia Agricola, 2007, vol. 64, p. 460

Rohde R. A., Hausfather Z., The Berkeley Earth land/ocean temperature record, Earth System Science Data, 2020, vol. 12, p. 3469

SEADE, 2022 Informações dos Municípios Paulistas <`http://www.imp.seade.gov.br/`> Acessado: 2017-11

Shafer M. A., Fiebrich C. A., Arndt D. S., Fredrickson S. E., Hughes T. W., Quality assurance procedures in the Oklahoma Mesonetwork, Journal of Atmospheric and Oceanic Technology, 2000, vol. 17, p. 474

Silva G., Fonseca M. d. L., São Paulo, city-region: constitution and development dynamics of the São Paulo macrometropolis, International Journal of Urban Sustainable Development, 2013, vol. 5, p. 65

Silva Dias M. A., Dias J., Carvalho L. M., Freitas E. D., Silva Dias P. L., Changes in extreme daily rainfall for São Paulo, Brazil, Climatic Change, 2013, vol. 116, p. 705

Silva Dias M. A., Vidale P. L., Blanco C. M., Case study and numerical simulation of the summer regional circulation in São Paulo, Brazil, Boundary-Layer Meteorology, 1995, vol. 74, p. 371

Simpson G. L., Modelling palaeoecological time series using generalised additive models, Frontiers in Ecology and Evolution, 2018, vol. 6, p. 149

Souza C. M., Z Shimbo J., Rosa M. R., Parente L. L., A Alencar A., Rudorff B. F., Hasenack H., Matsumoto M., G Ferreira L., Souza-Filho P. W., et al., Reconstructing three decades of land use and land cover changes in Brazilian biomes with landsat archive and earth engine, Remote Sensing, 2020, vol. 12, p. 2735

Stewart I. D., A systematic review and scientific critique of methodology in modern urban heat island literature, International Journal of Climatology, 2011, vol. 31, p. 200

Stewart I. D., Oke T. R., Local climate zones for urban temperature studies, Bulletin of the American Meteorological Society, 2012, vol. 93, p. 1879

Stott P. A., Tett S. F. B., Jones G. S., Allen M. R., Mitchell J. F. B., Jenkins G. J., External Control of 20th Century Temperature by Natural and Anthropogenic Forcings, Science, 2000, vol. 290, p. 2133

Sugahara S., Da Rocha R. P., Ynoue R. Y., Da Silveira R. B., Homogeneity assessment of a station climate series (1933–2005) in the Metropolitan Area of São Paulo: instruments change and urbanization effects, Theoretical and applied climatology, 2012, vol. 107, p. 361

Sun R., Zhang B., Topographic effects on spatial pattern of surface air temperature in complex mountain environment, Environmental Earth Sciences, 2016, vol. 75, p. 621

Sun X., Cook K. H., Vizy E. K., The South Atlantic subtropical high: climatology and interannual variability, Journal of Climate, 2017, vol. 30, p. 3279

Sun Y., Zhang X., Ren G., Zwiers F. W., Hu T., Contribution of urbanization to warming in China, Nature Climate Change, 2016, vol. 6, p. 706

Suomi J., Hjort J., Käyhkö J., Effects of scale on modelling the urban heat island in Turku, SW Finland, Climate research, 2012, vol. 55, p. 105

Trenberth K. E., Has there been a hiatus?, Science, 2015, vol. 349, p. 691

Umezaki A. S., Ribeiro F. N. D., de Oliveira A. P., Soares J., de Miranda R. M., Numerical characterization of spatial and temporal evolution of summer urban heat island intensity in São Paulo, Brazil, Urban Climate, 2020, vol. 32, p. 100615

Verstraeten G., Poesen J., Demarée G., Salles C., Long-term (105 years) variability in rain erosivity as derived from 10-min rainfall depth data for Ukkel (Brussels, Belgium): Implications for assessing soil erosion rates, Journal of Geophysical Research: Atmospheres, 2006, vol. 111

Vincent L. A., Peterson T., Barros V., Marino M., Rusticucci M., Carrasco G., Ramirez E., Alves L., Ambrizzi T., Berlato M., et al., Observed trends in indices of daily temperature extremes in South America 1960–2000, Journal of climate, 2005, vol. 18, p. 5011

Wallace J. M., Hobbs P. V., Atmospheric science: an introductory survey. vol. 92, Elsevier, 2006

Wan H., Zhang X., Zwiers F., Human influence on Canadian temperatures, Climate Dynamics, 2019, vol. 52, p. 479

Wanderley R. L., M. Domingues L., A. Joly C., R. da Rocha H., Relationship between land surface temperature and fraction of anthropized area in the Atlantic forest region, Brazil, PloS one, 2019, vol. 14, p. e0225443

Wang J., Tett S., Yan Z., Correcting urban bias in large-scale temperature records in China, 1980–2009, Geophysical Research Letters, 2017, vol. 44, p. 401

Wang J., Yan Z.-W., Urbanization-related warming in local temperature records: a review, Atmospheric and Oceanic Science Letters, 2016, vol. 9, p. 129

Wang Y., Sun Y., Hu T., Qin D., Song L., Attribution of temperature changes in Western China, International Journal of Climatology, 2018, vol. 38, p. 742

Wang Z., Jiang Y., Wan H., Yan J., Zhang X., Detection and attribution of changes in extreme temperatures at regional scale, Journal of Climate, 2017, vol. 30, p. 7035

Wijngaard J., Klein Tank A., Können G., Homogeneity of 20th century European daily temperature and precipitation series, International Journal of Climatology: A Journal of the Royal Meteorological Society, 2003, vol. 23, p. 679

Wilson A. M., Jetz W., Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions, PLoS biology, 2016, vol. 14, p. e1002415

WMO, 1967 Technical report A note on climatological normals. World Meteorological Organization

WMO, 2018 Technical report Guide to Instruments and Methods of Observation: Volume I - Measurement of Meteorological Variables. World Meteorological Organization

Wood S. N., Generalized additive models: an introduction with R. CRC press, 2017

Xavier A. C., King C. W., Scanlon B. R., Daily gridded meteorological variables in Brazil (1980–2013), International Journal of Climatology, 2015

Xu Y., Li J., Fu H., The role of sea surface temperature variability in changes to global surface air temperature related to two periods of warming slowdown since 1940, Climate Dynamics, 2022, pp 1–19

Xu Z., Ji F., Liu B., Feng T., Gao Y., He Y., Chang F., Long-term evolution of global sea surface temperature trend, International Journal of Climatology, 2021, vol. 41, p. 4494

Zhao L., Lee X., Smith R. B., Oleson K., Strong contributions of local background climate to urban heat islands, Nature, 2014, vol. 511, p. 216

Zilli M. T., Carvalho L., Lintner B. R., The poleward shift of South Atlantic Convergence Zone in recent decades, Climate Dynamics, 2019, vol. 52, p. 2545

Appendix

Appendix A

Complementary information of Chapter 1

A.1 CMIP5 models

Table A.1 - CMIP5 models used for the attribution study. ALL is the simulations with both anthropogenic and natural forcings, NAT is the simulation with only natural forcings, and GHG is the simulation with only greenhouse gases. The experiment for that used the 1955-2014 time period used the RCP8.5 scenario to extended the ALL run.

Model	ALL	NAT	GHG	RCP8.5
ACCESS1-0	1	-	-	1
ACCESS1-3	3	-	-	1
BCC-CSM1-1	3	1	1	1
BCC-CSM1-1-M	3	-	-	1
BNU-ESM	1	1	1	1
CCSM4	8	4	3	6
CESM-LE	34	-	-	34
CESM1-CAM5	3	3	3	3
CESM1-FASTCHEM	3	-	-	-
CESM1-WACCM	4	-	-	3
CMCC-CESM	1	-	-	1
CMCC-CM	1	-	-	1
CMCC-CMS	1	-	-	1
CNRM-CM5	10	6	6	5
CNRM-CM5-2	1	-	-	-
CSIRO-Mk3-6-0	-	5	5	-

CanESM2	-	5	5	-
FGOALS-g2	-	-	1	-
FIO-ESM	3	-	-	3
GFDL-CM2p1	10	-	-	-
GFDL-CM3	5	3	-	1
GFDL-ESM2G	3	-	-	1
GFDL-ESM2M	1	-	-	-
GISS-E2-H	18	10	-	-
GISS-E2-H-CC	1	-	-	1
GISS-E2-R	25	10	5	-
GISS-E2-R-CC	1	-	-	1
HadGEM2-CC	1	-	-	-
HadGEM2-ES	4	4	-	4
INMCM4	1	-	-	1
IPSL-CM5A-LR	6	3	-	4
IPSL-CM5A-MR	3	3	3	-
IPSL-CM5B-LR	1	-	-	1
MPI-ESM-LR	3	-	-	-
MPI-ESM-MR	3	-	-	1
MPI-ESM-P	2	-	-	-
NorESM1-M	3	-	1	1
NorESM1-ME	1	-	-	-

A.2 CRUTEM4 decadal anomalies

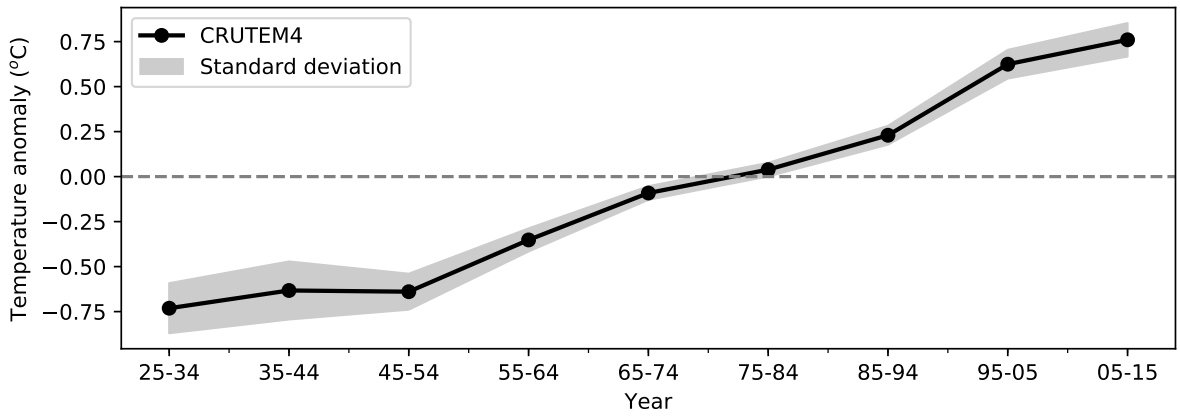

Figure A.1: Decadal anomalies for CRUTEM4 dataset. Also is displayed the ±1 standard deviation as shaded calculated using 100 ensemble members of the land only component of HadCRUT4 and the uncorrelated errors from the same dataset. The labels in x axis shows the decade that the anomaly was calculated with respecto the 1925 to 2014 climatology.

Appendix B

Complementary information of Chapter 2

B.1 Weather stations

Table B.1 - Geographical position and altitude of the stations used in this study. The stations that start with "iag"are from the Instituto de Astronomia, Geofísica e Ciências Atmosféricas/Universidade de São Paulo (IAG); "iac"from the Instituto Agronômico de Campinas/Secretaria de Agricultura e Abastecimento de São Paulo (IAC); "inm"from the Instituto Nacional de Meteorologia (INMET); and "ice"from the Instituto de Controle do Espaço Aéreo/Ministério da Aeronáutica (ICEA).

id	latitude (^{o})	longitude (^{o})	altitude (m)	id	latitude (^{o})	longitude (^{o})	altitude (m)
inm01	-13.332407	-44.617374	551	inm27	-21.461019	-47.579512	620
inm02	-13.251097	-43.405365	447	inm28	-21.226130	-44.979666	916
inm03	-14.089070	-46.366530	830	inm29	-21.769990	-43.364328	936
inm04	-14.949727	-46.235795	854	inm30	-21.204389	-41.905670	123
inm05	-15.902500	-52.245278	327	inm31	-21.742500	-41.332778	15
inm06	-15.854722	-48.966111	766	inm32	-22.022222	-42.364444	516
inm07	-15.789722	-47.925833	1161	inm33	-22.126273	-45.043327	930
inm08	-15.549167	-47.338889	938	inm34	-22.451111	-44.444722	439
inm09	-15.915229	-46.107120	523	inm35	-23.325556	-51.141667	566
inm10	-15.448054	-44.366321	480	inm36	-23.496389	-46.620000	785
inm11	-16.009560	-41.281027	647	inm37	-25.010777	-50.853734	808
inm12	-14.297500	-43.771389	455	inm38	-24.786944	-49.999167	994
inm13	-16.673056	-49.263889	748	inm39	-25.502846	-50.637609	881
inm14	-16.366268	-46.889321	595	inm40	-25.536111	-48.528333	4
inm15	-16.686333	-43.843759	645	inm41	-23.480000	-47.426667	597
inm16	-16.154862	-42.284921	476	iac01	-22.867442	-47.072914	667
inm17	-16.580810	-39.783182	197	iac02	-22.252333	-48.565944	599
inm18	-17.859776	-42.852647	919	iac03	-21.446076	-46.986794	662
inm19	-17.739444	-39.258611	6	iac04	-24.610217	-47.883792	43
inm20	-18.170278	-47.958056	857	iac05	-22.968783	-45.452533	568
inm21	-18.713975	-39.848749	39	iac06	-21.207042	-47.871414	632
inm22	-19.020355	-43.433948	663	iac07	-23.285510	-47.898402	574
inm23	-19.735765	-42.137222	609	iag01	-23.651242	-46.622424	799
inm24	-20.439722	-49.983611	510	ice01	-22.999617	-47.143638	657
inm25	-20.584314	-47.382427	1003	ice02	-23.221937	-45.868239	646
inm26	-20.316038	-40.317225	18	ice03	-23.623106	-46.657749	802

B.2 Timeseries homogenization

Many tests used to detect breakpoints in timeseries are available in the literature, and each of them may perform better than the others depending on specific situations, making the use of multiple tests an approach that is favorable to ensure that there are a lower

number of false alarms, i. e., the detection of breakpoints that are artificial (Wijngaard et al., 2003). Hence, we use two independent tests to detect the change points, the Standard Normal Homogeneity Test (SNHT; Alexandersson and Moberg (1997)) and Pettitt test (Pettit, 1979; Verstraeten et al., 2006). If both tests find a breakpoint in a two-year time window of the other one we apply the correction as in Alexandersson and Moberg (1997), if only one of the tests identify a given change point we discard it and no correction is applied.

The selected tests are applied in a difference series Q from the candidate station Y_k and a reference series R_k, which is a key point in determining the locations of the breakpoints. We follow the method described in Alexandersson and Moberg (Alexandersson and Moberg, 1997) where for a particular time i, R_k is the weighted average of the X_j sites based on the squared correlation ρ_j between the site and the candidate series Y_k:

$$R_{ki} = \frac{\sum_{j \neq k} \rho_j^2 [X_{ji} - \overline{X}_j + \overline{Y}]}{\sum_{j \neq k} \rho_j^2} \tag{B.1}$$

For each candidate station, the reference series is calculated using all stations with a correlation above 0.7, with an altitude difference lower than 500 m and a distance of less than 300 km from the candidate series, similar to El Kenawy et al. (2013). Figure B.1 shows a summary of the characteristics between the candidate stations and the ones used to compute their reference series. The median correlation is above 0.9 with an interquartile range that is greater than 0.85, indicating a high correlation between stations, with a median distance between stations close to 200 km because of the somewhat low station density. The altitude difference has more than 50 % of the distribution between 60 and 300 m, and we see a highly skewed distribution for the number of stations used for each candidate series, with more stations using fewer nearby stations than a higher number.

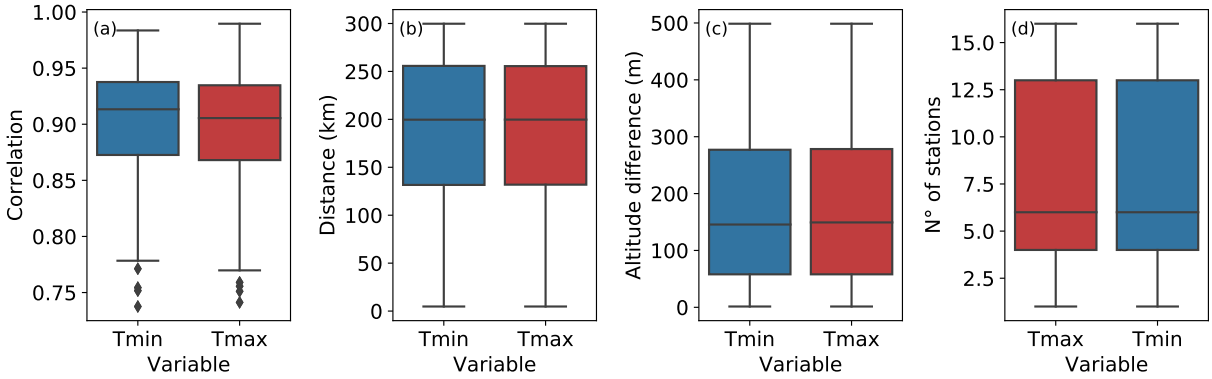

Figure B.1: Boxplot that summarizes the relationship between the candidate series and the surrounding stations used to calculate the reference series R_k for Tmin and Tmax: (a) Correlation; (b) Distance between stations; (c) Altitude difference; (d) the number of stations used to compute R_k.

To detect an inhomogeneity, the test goes as follows: (1) We calculate the reference series R_k so we can obtain the difference series Q; (2) The SNHT and Pettitt tests are applied so we can identify the change points. Since both tests only identify a single change point, we check whether or not the station has multiple change points by dividing the series into smaller subsections: one before the change point and another one after it. The test is then reapplied for each subsection and this process is repeated iteratively until no more breakpoints are found or the series length is shorter than 12 months; (4) We compare the breakpoints found in the SNHT and Pettit test and select only the ones that are in a two-year time window of the other; (5) The correction is applied to the candidate series based on its difference with the reference series.

An important source of inhomogeneity is caused by changes in the local environment like vegetation growing in the station's surroundings or gradual warming caused by urbanization which are key features that we would like to preserve for our analysis. However, the tests described here only deal with step changes and may homogenize a series around a breakpoint that is part of a larger trend. Therefore, to deal with this limitation, we classify the change points based on five models as in Menne and Williams Jr (2009) (Table B.2):

Table B.2 - Models used to classify the breakpoints timeseries. p is the number of parameters used to fit the model, ϵ_i is a random noise term, and μ and β are the parameters estimated to fit the model (Menne and Williams Jr, 2009).

Model	Description	Parameters (p)
M1	$Q_i = \mu + \epsilon_i$	1
M2	$Q_i = \mu + \beta i + \epsilon_i$	2
M3	$Q_i = \begin{cases} \mu_1 + \epsilon_i \; i \leq a \\ \mu_2 + \epsilon_i \; i > a \end{cases}$	3
M4	$Q_i = \begin{cases} \mu_1 + \beta i + \epsilon_i \; i \leq a \\ \mu_2 + \beta i + \epsilon_i \; i > a \end{cases}$	4
M5	$Q_i = \begin{cases} \mu_1 + \beta_1 i + \epsilon_i \; i \leq a \\ \mu_2 + \beta_2 i + \epsilon_i \; i > a \end{cases}$	5

To select the most suitable model we use the one with the lowest Bayesian Information Criteria (BIC) that aims to weight the sum of squared errors by the number of parameters added. With this classification, we only homogenize the breakpoints classified as either M3, M4, or M5. M1 model represents a constant value, while M2 is supposed to show breakpoints inside a longer trend, which could be caused by changes in the local environment, for example.

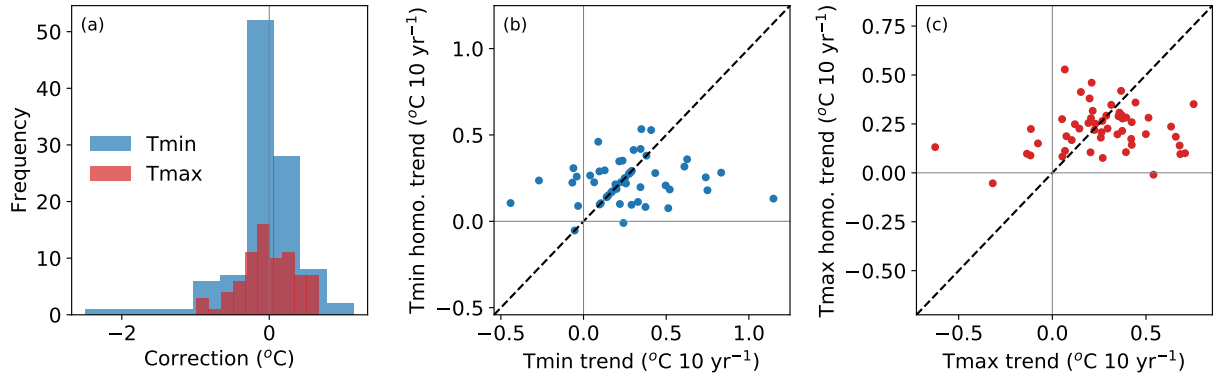

Figure B.2: (a) Histogram of the correction applied for all stations; (b) Scatter plot of the trend for minimum temperature with the raw data versus the homogenized data; (c) same as (b) but for maximum temperature.

A total of 36 and 35 stations were homogenized for minimum and maximum temperature, respectively. The correction applied in each breakpoint is shown in Figure B.2a, indicating that most corrections are lower than ± 1 °C, with a few of them reaching values as low as -2 °C for minimum temperature. The impact of the homogenization procedure

on the trend tends to smooth the difference between stations for both Tmin and Tmax (Figure B.2b and c).

B.3 Generalized Additive Model (GAM)

In this section we give some basic information about how generalized additive models (GAM) are derived from a penalized least squares perspective and some useful definitions that are intended to help the unfamiliar reader with the topic. We only consider the case where the residual is normally distributed, which is the case that we go through in the main text, and because the derivation is easier to follow in the authors opinion. Therefore, even tough we define the section as "Generalized Additive Models"we are actually dealing with additive models which are the special case of GAMs when the residual is normally distributed. The text is based on Wood (2017) and Hastie et al. (2009) which are highly recommended sources if the reader wants a more thorough explanation.

Let's consider that we can represent a random variable Y as a sum of smooth functions $f(X)$ plus a random noise term $\epsilon \sim N(0, \sigma^2)$, as represented by:

$$Y = f(X) + \epsilon = \sum_{k=0}^{q} \beta_k b_k(X) + \epsilon \tag{B.2}$$

Where β_k is the scaling parameter for the smooth function, b_k is a basis function and q is the number of basis functions. For example, for a cubic spline, we have that:

$$\begin{cases} b_k = X^k, & k = 0, ..., 3 \\ b_{(3+l)} = (X - \zeta_l)_+^3, & l = 1, ..., L \end{cases} \tag{B.3}$$

Where ζ_l is the l-th knot and $()_+$ is positive when the difference inside the parenthesis is greater than zero and zero otherwise, and in this case $q = L + 4$. Considering that we have $i = 1, \ldots, n$ samples, we can represent the spline function $f(X)$ from B.2 in matrix notation as:

$$\mathbf{f} = \mathbf{X}\boldsymbol{\beta} \tag{B.4}$$

In which the i-th row of $\mathbf{X}$ is represented by $\mathbf{X}_i = [b_0(x_i), b_1(x_i), \ldots, b_q(x_i)]$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_q]^T$. An example is presented in Figure B.3 where maximum temperature is

fitted as a function of altitude in two different ways: the first using third degree B-Splines (Figure B.3a) and the second using a second degree polynomial (Figure B.3b). In both cases the function $f(X)$ will be given as the sum of the basis function multiplied by their respective scaling parameter.

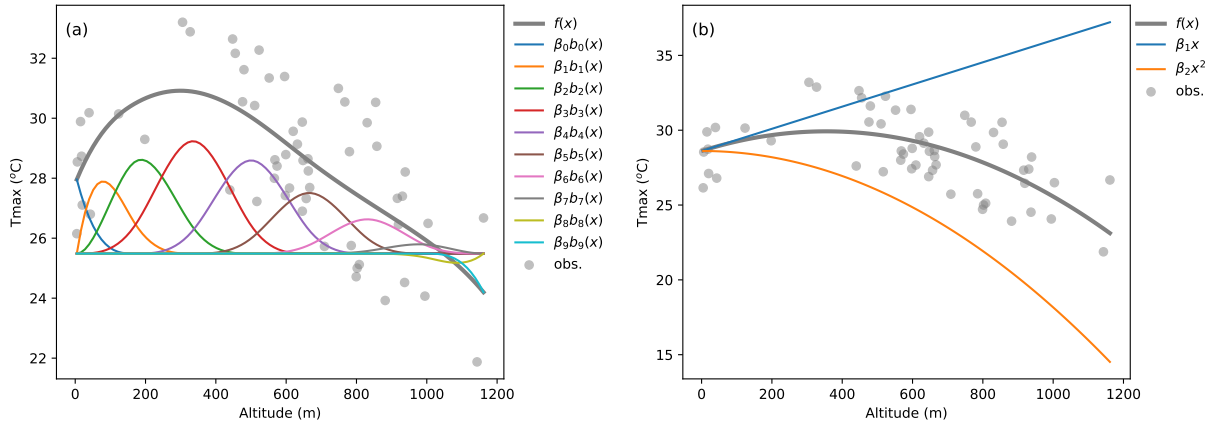

Figure B.3: Fitted maximum temperature as a function of altitude using: (a) Third degree B-Splines; (b) Second degree polynomial.

In the case of additive models, we consider a variable $\mathbf{y}$, which could be temperature, for example, that can be represented as the sum of a number p of independent variables, like altitude, latitude and longitude, that are represented as smooth function defined by Eq. B.2. Therefore:

$$y_i = \sum_{j=1}^{p} f(x_{j,i}) + \epsilon_i \tag{B.5}$$

Where $\epsilon_i \sim N(0, \sigma^2)$. In matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{B.6}$$

In which $\mathbf{X} = [\mathbf{X_1} : \ldots : \mathbf{X_p}]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta_1}^T : \ldots : \boldsymbol{\beta_p}^T]^T$. Instead of using ordinary least squares to obtain an estimator for $\boldsymbol{\beta}$, we use the penalized least squares, in order to penalize functions that are too wiggly, giving preference for simpler functions. This means that we have to minimize the following equation:

$$RSS(\boldsymbol{\beta}, \lambda_1, ..., \lambda_p) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} \tag{B.7}$$

Which is the residual sum of squares, given by the first term in the right hand side of

Eq. B.7, plus a penalization term, given by the second term in the right hand side of Eq. B.7. The term $\mathbf{S}$ is given by:

$$\mathbf{S} = \begin{bmatrix} \lambda_1\mathbf{S_1} & 0 & \dots & 0 \\ 0 & \lambda_2\mathbf{S_2} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \lambda_p\mathbf{S_p} \end{bmatrix} \tag{B.8}$$

And the entry from row i and column k from $\mathbf{S}_j$ is $\mathbf{S_j}_{,ik} = \int b_{ki}(x_k)''b_{jk}(x_j)''dx$. The parameter λ_j controls the degree of penalization for wiggly functions. As $\lambda_j \to \infty$, penalization increase and $\mathbf{f_j}$ tends to a linear function, while if $\lambda_j = 0$ the penalization is minimal and the resulting function will be an interpolator in all points. Computing the derivative of Eq. B.7 with respect to $\boldsymbol{\beta}$ and equating to zero:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^T\mathbf{y} \tag{B.9}$$

From this result we can define the smooth matrix $\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^T$. Here we introduce the concept of estimated degrees of freedom as edf $= \text{tr}(\mathbf{A})$, in an analogous form of the multiple linear regression case (MLR). For example, in MLR with p independent variables, we have that $\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ since there is no penalization, and therefore the number of degrees of freedom is:

$$\text{edf} = \text{tr}(\mathbf{A}) = \text{tr}(\mathbf{I}_p) = p \tag{B.10}$$

Where $\mathbf{I}_p$ is the identity matrix of size $p \times p$. In the additive model, since there is penalization in $\hat{\boldsymbol{\beta}}$ the edf will give a more appropriate measure of the degrees of freedom from the model, even if more parameters are used than the estimated edf. We can even estimate the edf of each β_i. If $\mathbf{P} = (\mathbf{X}^T\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^T$ we have that edf $= \text{tr}(\mathbf{XP})$. Now, if $\mathbf{P^0}_i$ is the matrix $\mathbf{P}$ with all rows equal to zero, with exception of the i-th row, the effective degrees of freedom for β_i is given by $\text{edf}_{\beta_i} = \text{tr}(\mathbf{XP^0}_i)$, and therefore the edf of each $\mathbf{f}_j$ will be given by the sum of the edfs of each $\beta_i \in \boldsymbol{\beta}_j$. If the edf is close to unit this means that the penalization is greater and a linear function could represent satisfactorily the dependency between independent and dependent variable, while higher values of edf indicate that non linear relationships are more appropriate to represent the given term.

From Eq. B.9 it is possible to find the expected value and variance of the estimator $\hat{\boldsymbol{\beta}}$, which are given, respectively, by:

$$\mathbf{E}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \tag{B.11}$$

$$\mathbf{V}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{S})^{-1}\sigma^2 \tag{B.12}$$

It is easy to see that $\mathbf{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}]$ is not an unbiased estimator of $\boldsymbol{\beta}_{\boldsymbol{j}}$, however, if $\boldsymbol{\beta}_{\boldsymbol{j}} = \mathbf{0}$ we also have that $\mathbf{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}] = \mathbf{0}$, so we can test the null hypothesis that $\boldsymbol{\beta}_{\boldsymbol{j}} = \mathbf{0}$. In that case we have that $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} \sim N(\mathbf{0}, \mathbf{V}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}])$ and that:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}{}^T\mathbf{V}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}]^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} \overset{\mathrm{H}_0}{\sim} \chi^2_d \tag{B.13}$$

If $\mathbf{V}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}]$ is of full rank, with $d = \dim(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}})$. However, because of the penalization term, usually $\mathbf{V}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}]$ is not full rank, so we use $\mathbf{V}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}]^{r-}$ which is the rank r pseudo-inverse of the covariance matrix, where $r = \mathrm{rank}(\mathbf{V}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}])$. We have also to estimate the variance σ^2, which is assumed to be unknown, using the sum of the squared error:

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{A}\mathbf{y})^T(\mathbf{y} - \mathbf{A}\mathbf{y})}{n - \mathrm{tr}(\mathbf{A})} \tag{B.14}$$

Finally, we combine B.13 and B.14 so that under $\mathrm{H}_0 : \boldsymbol{\beta}_{\boldsymbol{j}} = \mathbf{0}$:

$$\frac{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}{}^T\mathbf{V}[\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}]^{r-}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}/r}{\hat{\sigma}^2/(n - \mathrm{edf})} \overset{\mathrm{H}_0}{\sim} F_{r,\mathrm{edf}} \tag{B.15}$$

Which can be used to calculate the associated p-values for each one of the p independent variables and determine if they are statistically significant in the model.

B.4 Number of NDVI images

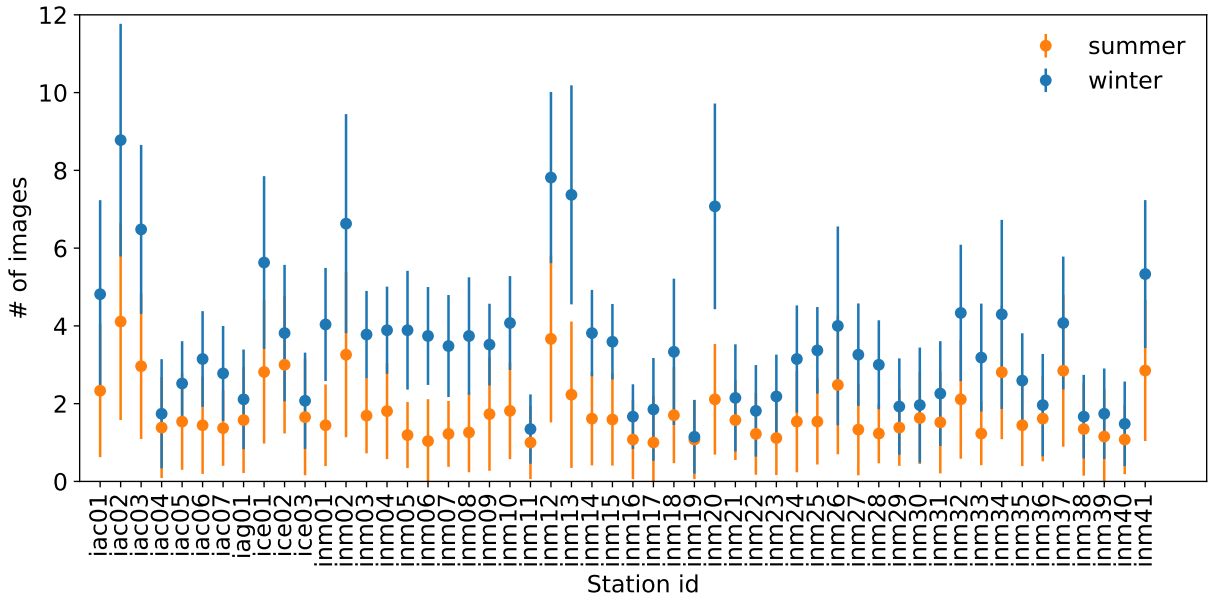

Figure B.4: Average number of images by season (summer and winter), for each available station. Vertical bars represent the ±1 standard deviation.

B.5 Wind speed

The wind speed is an important factor to consider when quantifying the absolute contribution of each of the estimated functions, especially for land use (Oke et al., 2017; Kagawa-Viviani and Giambelluca, 2020), as suggested by Figure B.5. We noticed that the correlation is higher for days when wind speed is lower, below the 25 % percentile (P25), when compared to samples for days when wind speed is higher, above the 75 % percentile (P75). Therefore, we filtered daily Tmin and Tmax for days when wind speed was below P25 and above P75, and calculated temperature averages for the different temporal aggregations considered in this study: annual, summer, and winter. Daily wind speed was obtained for the gridded data of Xavier et al. (2015), in which we interpolated the data to each station coordinate.

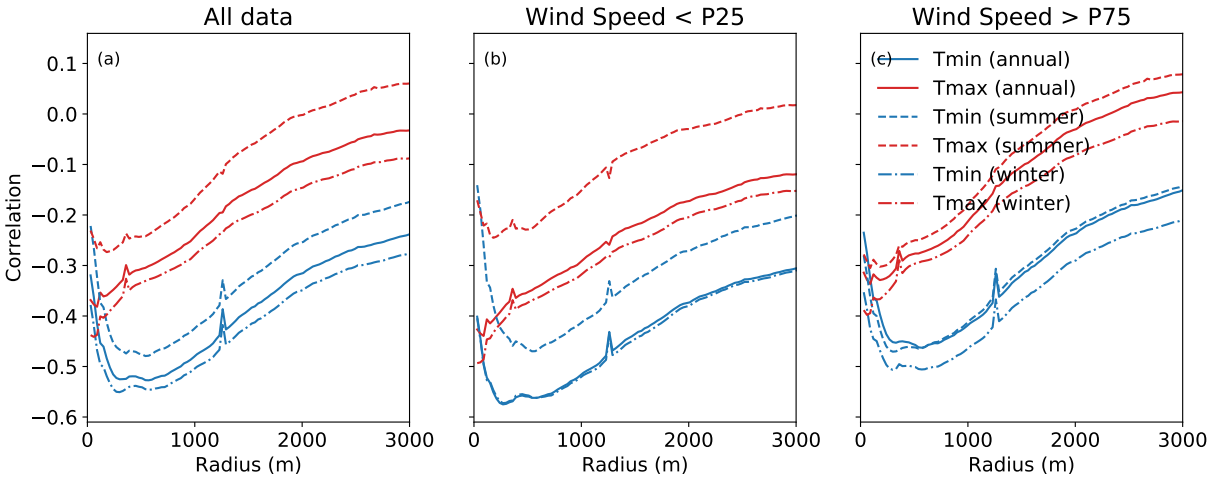

Figure B.5: Pearson correlation between the minimum and maximum temperature, Tmin and Tmax, respectively, and the NDVI in the annual, summer, and winter aggregations, based on the average between 1985 and 2010. The NDVI was calculated as the average of the pixels surrounding the station coordinates, considering a radius that varies according to the x-axis. The average temperature was calculated considering: (a) all available data; (b) only days when the wind speed was lower than the 25 % percentile; (c) only days when the wind speed was greater than the 75 % percentile.

We fitted the GAM individually and we present the amplitude of each function in Figure B.6, which was calculated as the difference between the maximum and minimum values of each function. Only statistically significant functions with a p-value below 5 % were considered. We noticed that the function of geographical position s(lon,lat) (Figures B.6a,d) has a greater amplitude in winter for Tmin when wind speed is below P25, while during summer the greater amplitude occurs when wind speed is above P75, suggesting a more intense thermal gradient. For cloud cover (Figure B.6f) the amplitude was also higher in winter for P75, while in summer it was not statistically significant.

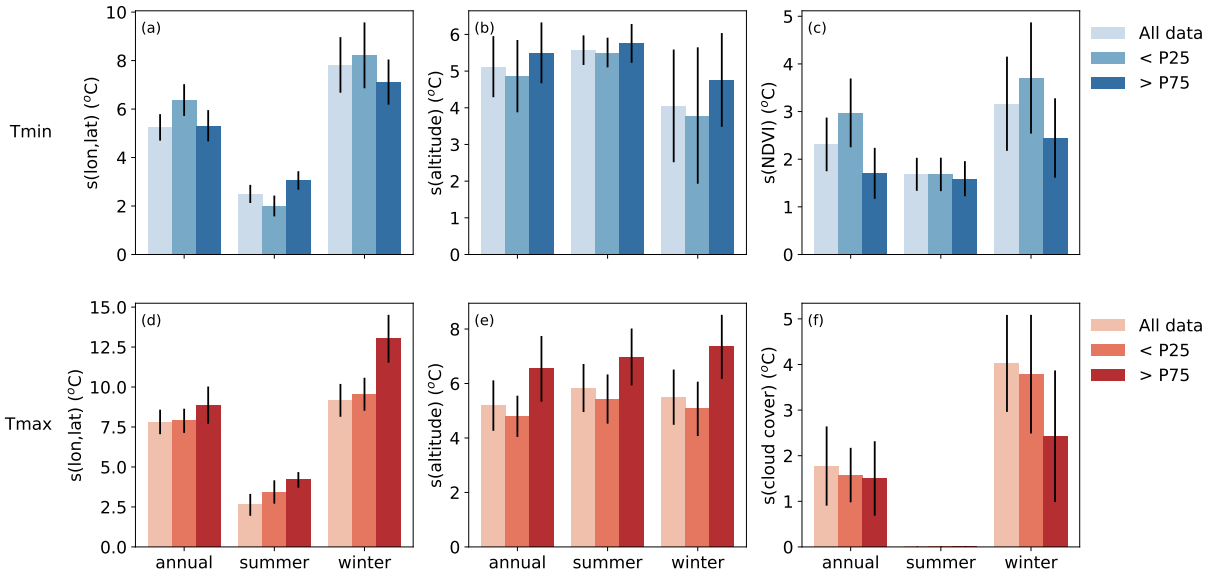

Figure B.6: Amplitude (difference between maximum and minimum values) of the contribution of each individual function of the GAM for three different datasets: 1) all data, 2) only days when wind speed was below the 25 % percentile (¡ P25) and 3) only days when the wind speed was above the 75 % percentile (¿ P75). The model was fitted individually for Tmin and Tmax in each of the given seasons (annual, summer, and winter) and for each of the three different datasets. Figures (a)-(c) represent the results for minimum temperature and figures (d)-(g) are for maximum temperature. The vertical lines represent the 95 % confidence interval.

The function s(altitude) (Figures B.6b,e) showed greater variability in summer for Tmin, with no significant difference between the different ranges of wind speed, while for Tmax even though there was no major variation between seasons the amplitude was greater for the case when wind speed is above P75. Still, for NDVI (Figure B.6c) there is an increased contribution for cases when wind speed is low for minimum temperature, being 50 % higher than P75 in winter and approximately 75 % in the annual case. During summer, the differences are lower and not very significant. For Tmax s(NDVI) was not statistically significant in any of the cases.

B.6 Urban Infrastructure timeseries

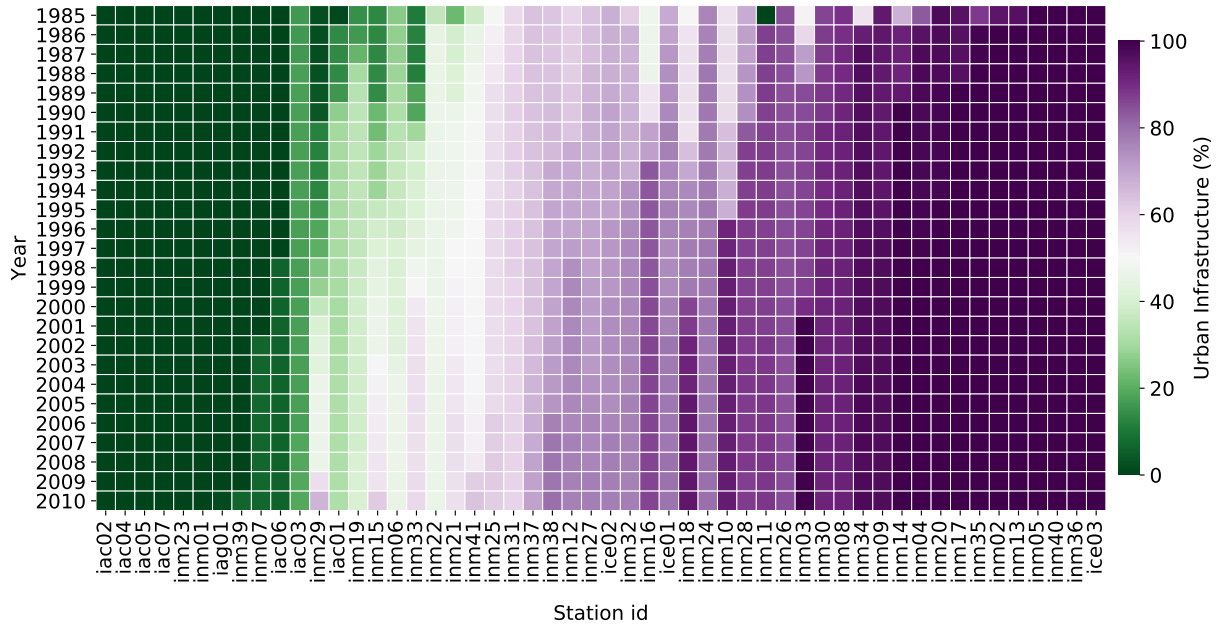

Figure B.7: Average annual urban fraction between 1985 and 2010 for the selected stations. The urban fraction was calculated as an average of all pixels inside a circle with a 300 m radius around each station based on MapBiomas version 5.0 classification (Souza et al., 2020).

B.7 GAM complementary results

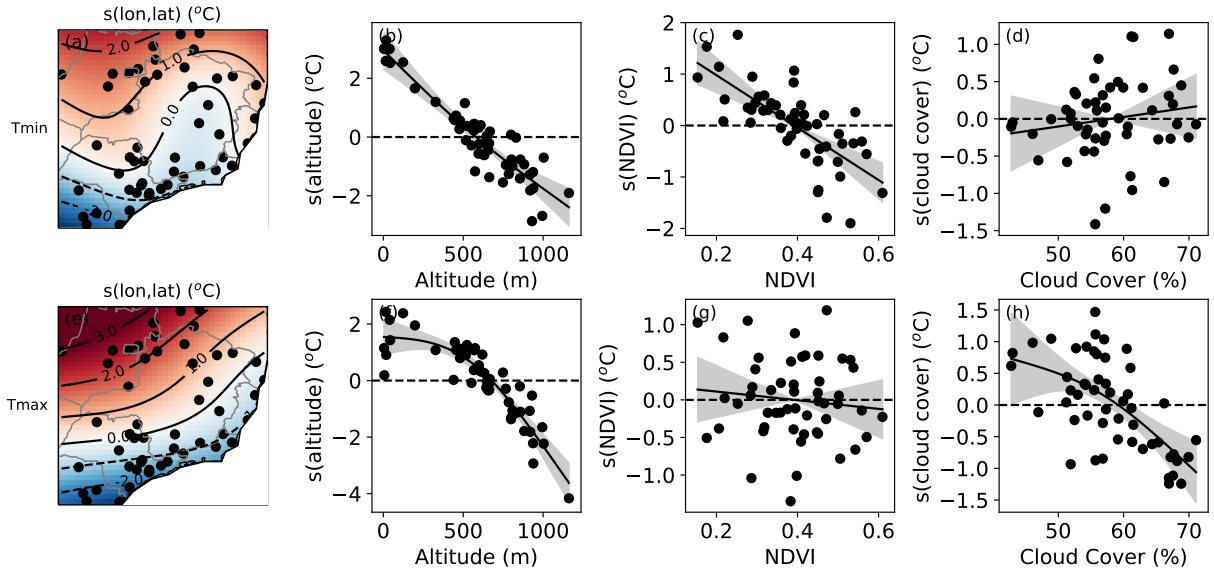

Figure B.8: Contribution of each function in the GAM, in °C, using all available independent variables as in Equation 2, even the ones that were not statistically significant, for the annual mean of Tmin (a, b, c, d) and Tmax (e, f, g, h). In (a) and (e), is the function related to the geographical position s(lon, lat); in (b) and (f) is the altitude in meters above sea level; (c) and (g) the NDVI; (f) and (h) the cloud cover. In (a) and (d) we show the position of each station used to fit the model. In (b), (c), (d), (f), (g) and (h): the points are the partial residual of the given function and the fitted GAM response is displayed as a solid line with a 95 % confidence interval.

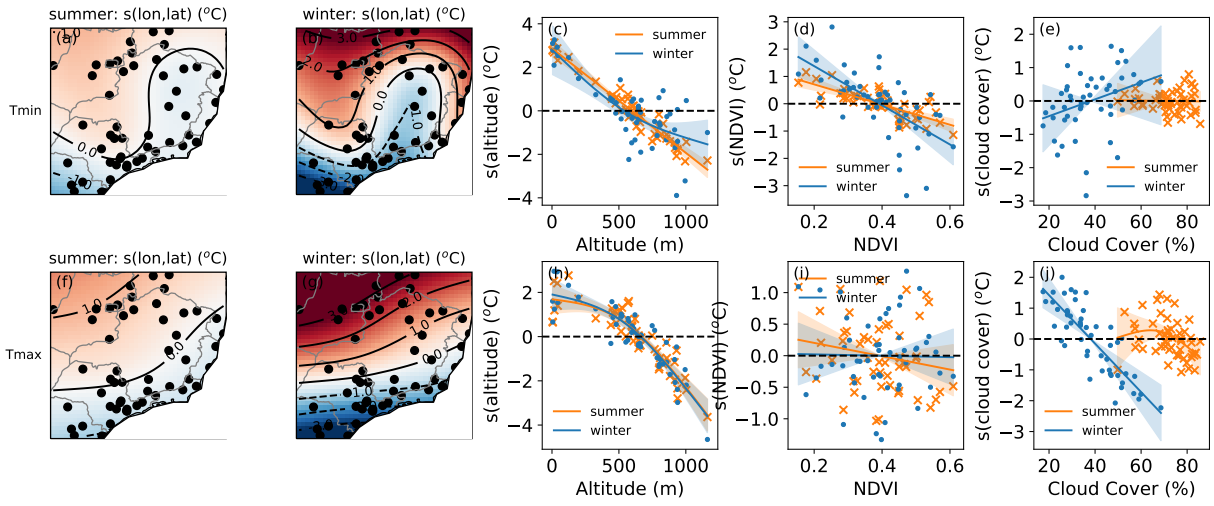

Figure B.9: Contribution of each function in the GAM, in °C, using all available independent variables as in Equation 2, even the ones that were not statistically significant, for the seasonal mean of Tmin (a, b, c, d and e) and Tmax (f, g, h, i and j) for summer and winter. (a), (b), (f) and (g) shows the geographical position, s(lon, lat); altitude in (c) and (h); NDVI in (d) and (i); and cloud cover in (e) and (j). In (a), (b), (f) and (g) we show the position of each station used to fit the model. In (c), (d), (e), (h), (i) and (j),: the points are the partial residual of the given function and the fitted GAM response is displayed as a solid line with a 95 % confidence interval.

Table B.3 - Results from the Generalized Additive Model (GAM) for maximum and minimum temperature (Tmax and Tmin, respectively) for summer and winter. We show the estimated degrees of freedom (edf), the coefficient of determination (R^2), and the Bayesian Information Criteria (BIC). Only the terms that were statistically significant with a p-value < 0.01 are displayed.

	Summer			Winter	
	Tmax (edf)	Tmin (edf)		Tmax (edf)	Tmin (edf)
Intercept	29.8	19.8	Intercept	26.6	13.9
s(lon,lat)	4.53	7.40	s(lon,lat)	5.60	7.92
s(altitude)	1.93	1.00	s(altitude)	1.95	1.69
s(NDVI)	-	1.00	s(NDVI)	-	1.00
s(cloud cover)	-	-	s(cloud cover)	1.00	-
R^2	85.9 %	95.7 %	R^2	95.8 %	87.7 %
BIC	125.1	74.7	BIC	137.4	186.4

Table B.4 - Results for the Generalized Additive Model (GAM) for minimum temperature in annual, summer and winter aggregations, using the s(NDVI300, NDVI3000) function for fitting. We also show the results for the model with s(NDVI300) as comparison. We show the estimated degrees of freedom (edf), the coefficient of determination (R^2), and the Bayesian Information Criteria (BIC). Only the terms that were statistically significant with a p-value < 0.01 are displayed.

	Annual	Summer	Winter
s(lon,lat)	8.50	7.88	8.62
s(altitude)	1.00	1.00	1.00
s(ndvi 300m, ndvi 3000m)	6.79	6.01	6.77
s(cloud cover)	-	-	-
R^2	96.1 %	97.3 %	92.1 %
BIC	123.7	72.5	186.4
R^2 s(NDVI300) only	93.4 %	95.7 %	87.7 %
BIC s(NDVI300) only	127.2	74.7	186.4