Evaluation of time series causal detection methods on the influence of Pacific and Atlantic Ocean over Northeastern Brazil precipitation*

Juliano E. C. Cruz[1,2], Mary T. Kayano[3], Alan J. P. Calheiros[3], Sâmia R. Garcia[4], and Marcos G. Quiles[4]

[1] Applied Computing Postgraduate Program, National Institute for Space Research (INPE), São José dos Campos, Brazil
[2] Research and Technology Dept., EMBRAER S.A., São José dos Campos, Brazil
[3] National Institute for Space Research (INPE), São José dos Campos, Brazil
[4] Federal University of São Paulo (UNIFESP), São José dos Campos, Brazil

Abstract. The detection of causation in natural systems or phenomena has been a fundamental task of science for a long time. In recent decades, data-driven approaches have emerged to perform this task automatically. Some of them are specialized in time series. However, there is no clarity in literature what methods perform better in what scenarios. Thus this paper presents an evaluation of causality detection methods for time series using a well-known and extensively studied case study: the influence of El Niño-Southern Oscillation and Intertropical Convergence Zone on precipitation in Northeastern Brazil. We employed multiple approaches and two datasets to evaluate the methods, and found that the SELVAR and SLARAC methods delivered the best performance.

Keywords: causality · time series · ENSO · precipitation

1 Introduction

Understanding the causes behind an observed phenomena is among the great goals of science. There are basically two paths to investigate causes: by using observational data, either in raw format or through models, or employing interventionist experiments under well-controlled conditions [23]. In order to discover the cause behind symptoms or behaviors, medicine and social sciences predominantly employ randomized controlled experiments [19], but for most Earth science fields, it is done by employing computer simulation, which can be very expensive, time-consuming, and may be strongly based on assumptions and knowledge from specialists [23]. In the past few decades, there has been a significant rise in the availability of time series data, originating from both observational data and models. This trend, coupled with the rapid growth in computational

* This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

power, has resulted in the creation of new opportunities to leverage data-driven approaches.

Causal detection methods and applications have been a subject of research for quite some time [8]. However, with the introduction of a more comprehensive and consolidated causal framework [19,20], there has been a surge in the number of studies published, including those that focus on time series data [5,9,17]. Time series has an additional feature and challenge: time order indexation. Regarding climate time series, there are also several studies targeting different parts of the globe, i.e. Arctic [23], Europe [26], India [3], Atlantic Ocean [26], Pacific Ocean [24] etc. and using different types of meteorological variables. However, it is really hard to foresee which methods may work or not only based on time series aspects such as frequency, shape, noise etc. Most published papers only target specific climatic scenarios or systems and they do not extrapolate results for general usage. Therefore, the best way to evaluate the causal detection methods is to employ them in a case study of interest.

Although there are several studies done in the last decades that connect the Northeastern Brazil (NEB) precipitation variation to ENSO and to Tropical Atlantic phenomena [11,13,18], none of them employ causal detection methods. In that region, severe impacts on the precipitation behavior have been registered since the sixteenth century and due to its characteristics, it has a high seasonal climate predictability [18]. One explanation for the phenomenon is that extreme changes in sea surface temperature (SST) in the tropical Pacific, the El Niño-Southern Oscillation (ENSO), influence precipitation anomalies through changes in the Walker circulation [1]. But it accounts for only part of the rainfall variability. For example, from the 46 strong or moderate El Niño events between 1849 and 1992, only 21 were associated with droughts in north of Northeastern Brazil (NEB) [13,18]. The other ones can be explained by an anomalous northern position of the Intertropical Convergence Zone (ITCZ) over the Atlantic, caused by a warmer tropical North Atlantic [18].

The NEB precipitation anomalies have a well-understood behavior, therefore they are an excellent case study for evaluating the performance of data-driven causal detection methods. Thus, the objective of this study is to evaluate causal detection methods for time series on the relationship of NEB precipitation with ENSO and with Tropical Atlantic SST. Three experiments were performed. The first one aims to test whether the methods are capable of detecting the already known causality between the raw (but linear detrended) time series. The second one tests how the causal detection methods behave when filtering is applied on the time series. The third experiment checks how stable the best methods found in the prior experiments are when varying the time series length and their start and end over time.

2 Data

Seven types of monthly time series are used, where the time span ranges from 1950 to 2016. Three precipitation time series are from the region of interest,

Northern NEB. One precipitation time series (SRUNK) is from east of Southern Ural Mountains region. Two time series are related to the causative phenomena: Oceanic Niño Index (ONI) and the Tropical North Atlantic (TNAtl) SST. The last one is a synthetic random series, which does not affect or is caused by any of Earth system time series.

All the precipitation time series were extracted from GPCC (NOAA Global Precipitation Climatology Centre) repository, which has 0.5°x0.5° resolution. A polygon in the north part of Northeastern Brazil was used after two papers [10,13] also used the same area for precipitation anomaly prediction. The polygon is located between 3°-8°S and 36°-41°W. As already said, three time series from this region were used: the polygon precipitation average, the precipitation of the cells with the highest (4°-4.5°S and 40.5°-41°W) and the lowest (7.5°-8°S and 36°-36.5°W) correlation with SST of the TNAtl region and with ONI. The latter two time series were chosen, because it is quite normal in large areas to exist local variability, then the extremes (most correlated and less correlated) were picked in order to see how the detection methods would behave. The correlation with SST of the TNAtl region and with ONI were performed after a Morlet wavelet filtering considering the range from 2 to 7 year frequency [6,30].

SRUNK is an average from the polygon located between 52°-55°N and 60°-70°E, which is south of the Russian Ural District, east of the Ural Mountains, and in Northern Kazakhstan. It was chosen mainly because Lin and Qian [16] showed that ENSO has little or no influence over this area.

ONI and a region on the TNAtl are used as time series related to the causative phenomena. ONI, Oceanic Niño Index, is an index used to monitor ENSO, which can be used to know if the phenomenon is on El Niño or La Niña phase and its severity. It is calculated using a 3-month running mean on the SST of the region Niño 3.4 [7]. The polygon of Tropical North Atlantic SST is between 6°-22°N and 15°-60°W and was extracted from ERSST (Extended Reconstructed Sea Surface Temperature) dataset, which has 2°x2° resolution.

Prior studies [14,25,26] calculate time series anomalies and other studies [2,4,27] besides that, also apply filters in order to remove unwanted frequencies when trying to establish relationships among meteorological variables. Therefore, another set of time series was also employed in this study, which was created by applying a transformation on the prior dataset presented. Besides wide usage, the objective is also to check whether a less noisy time series can contribute to a better performance in causal detection. The process applied on precipitation data was sequentially the following: month-wise z-score calculation, linear trend removal, and finally, wavelet filtering.

3 Methods

There are two main groups of methods that were employed in this study. The ones that use hypothesis significance test (Granger, FullCI, PC, PCMCI, and PCMCI+) and the ones (QRBS, SLARAC, LASAR, SELVAR) that output

scalar results between 0 and positive infinity, which is used to infer the causal link likelihood [31].

3.1 Statistical hypothesis methods

Granger causality is a statistical concept of causality based on prediction, where its mathematical formulation is based on linear regression modeling of stochastic processes [8]. According to this method, for detecting causation of X on Y, Y_{t+1} has to be better predicted using X_{t-a} and Y_{t-b}, where $a, b \geq 0$, than with only Y_{t-b}. There is two principles that must exist in order to claim a Granger causality exists: the cause happens before the effect; and the cause series has unique information about the future behaviors of the effect series that would not be available otherwise [5,8].

FullCI, acronym for Full Conditional Independence, is one of the most direct methods known for causal link detection [21]. In its original formulation, the Granger causality between time series X and Y is based on the use of a linear or non-linear model which may include possible confounders for Y. A causal connection $X \rightarrow Y$, is evaluated by quantifying whether the inclusion of the history of variable X in the model significantly reduces the forecast error about Y. Thus, FullCI can be interpreted as a Granger version that uses specific time windows [21].

PC, is an algorithm that was originally formulated for general random variables without assuming temporal order and was named after its creators Peter and Clark [28]. In the structural discovery phase, this method generates an undirected graphical model in which connections are driven using a set of rules. The version for time series uses the temporal ordering information, which naturally creates an orientation rule for the links [21].

PCMCI is the junction of the PC method [28] with MCI, acronym for Momentary Conditional Independence [24]. PCMCI is a method that was proposed later than PC and presents an approach that solves some limitations of the latter algorithm [24]. The PC version used in PCMCI differs from the canonical one, because only thesubset with the highest yield is used, instead of testing all possible combinations. It employs some approaches in order to have fewer tests, which theoretically does not take away the capacity to remove spurious connections [21]. An extended version, PCMCI+, was later proposed in [22]. In addition to the features of prior version, it can also detect contemporaneous links.

Three conditional independence tests are used in this study are: Partial Correlation, GPDC (Gaussian Process Distance Correlation), and CMI (Conditional Mutual Information). The first test is linear and the other ones, non-linear. Except for Granger, all the other methods used Partial Correlation. GPDC and CMI were just used with PCMCI and PCMCI+. Partial correlation is a very fast algorithm. On the other hand, GPDC and CMI are very costly and they can be more than one hundred times slower then the former.

3.2 Scalar output methods

Although being able to handle linear and non-linear causal-effect problems, all the four methods in this group use internally linear approaches. Furthermore, they do not make any data normalization or hypothesis testing, as Weichwald *et al.* [31] claims it considerably decreases the accuracy.

QRBS is the abbreviation for Quantiles of Ridge regressed Bootstrap Samples [31]. The general idea is to regress present values on past ones and then, verify the coefficients in other to decide whether one variable cause (according to Grange [8]) another variable. The method employs ridge regression [12] of time-deltas $X_t - X_{t-1}$ on the preceding values X_{t-1}. All the samples used are generated with bootstrap, that is, random sampling with replacement.

SLARAC stands for Subsampled Linear Auto-Regression Absolute Coefficients and was proposed by Weichwald *et al.* [31]. It also uses the concept of regression applied to past values and inspection of the coefficients in order to determine whether one variable causes (according to Grange) another. SLARAC fits a vector autoregression model on bootstrap samples, each time choosing a random number of lags to include. It sums all coefficient values obtained for every lag and in the end, it selects the highest score as result.

LASAR stands for LASSO Auto-Regression and was also proposed by Weichwald *et al.* [31]. As QRBS and SLARAC, it also relies in regression analysis in order to infer causation [8] between two variables, and different from the other approaches as it uses LASSO [29]. LASAR also uses bootstrap samples.

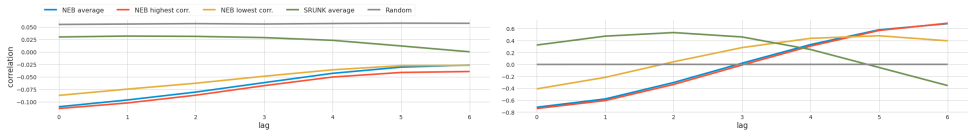
SELVAR or Selective Auto-Regressive model, detects causality with a hill-climbing procedure based on the leave-one-out residual sum of squares and at the final step, it scores the selected causal links with the absolute values of the regression coefficients [31].

4 Results

First, charts with lag correlations are shown in order to have a baseline for the rest of the paper. Then, the causal detection methods itself are evaluated in Sec. 4.2. The performance ranking regarding the previous experiments is subsequently presented in Sec. 4.3. Finally, a temporal stability experiment is done, which aims to find out if the best methods identified in Sec. 4.3 have the same behavior in smaller time windows as it does in full-length time series.

4.1 Correlation

The lag correlations among the precipitation time series and ONI and TNAtl SST were calculated for both time series set, raw and filtered. The results for raw time series are shown in Fig. 1 . As expected, the correlation of ONI with SRUNK precipitation average is low. The Northern NEB precipitations also have low correlation. Regarding the Atlantic, Fig. 1(b) , the correlation have higher values in magnitude compared to chart (a). SRUNK precipitation has a

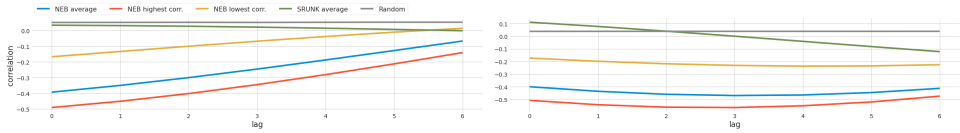


(a) ONI (b) TNAtl SST

Fig. 1: Precipitation time series correlation with Pacific and Atlantic. One time lag unit represents one month.

significant correlation, reaching its peak at lag 2. The SRUNK precipitation kept with low correlation, a fact already expected [16].

The results for the filtered times series set can be seen in Fig. 2. The wavelet filtering improved drastically the correlation values for ONI when compared to raw data correlation results. On the other hand, Fig. 2(b) did not keep the behavior saw in raw data chart, and it has lower peak values. Finally, all the random time series correlations for all lags stayed between 0.06 and 0, and had imperceptible tiny oscillations.



(a) ONI (b) TNAtl SST

Fig. 2: Filtered time series correlation with Pacific and Atlantic. One time lag unit represents one month.

4.2 Raw and filtered dataset

Most studies used as reference in this paper use p-value as 0.05 [21,22,24], which is also adopted here.

The famous quote says "correlation does not imply causation", which means in this context that causal detection methods must be capable of detecting causality having or not a significant correlation for true causal links and not detecting causality between time series despite there is significant correlation. Thus, the perfect result for either datasets would be the one that only ONI and TNAtl SST causes the NEB precipitations (average, highest correlation, and lowest correlation time series) and obviously, no other link is detected. That is exactly what is shown in Fig. 3.

Even though five of six real links were detected, the results of Granger using the raw dataset, Fig. 4(a), had plenty of spurious causal links with most of

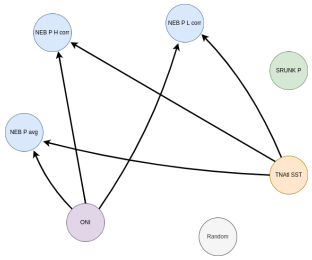
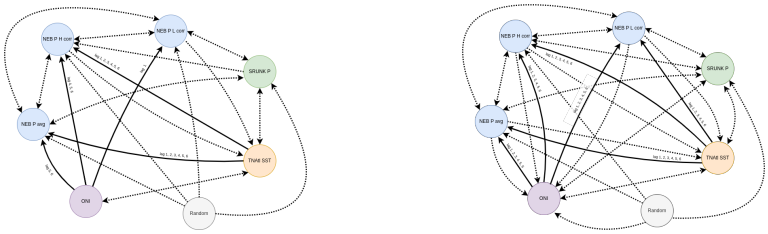


Fig. 3: The perfect result, where only the time series (ONI and TNAtl SST) related to the causative phenomena have a causality link with NEB precipitation time series.

them being bidirectional. Except for the random time series, all the rest is being caused at least by another node. Most of the links had detection for all time lags, which most of them have p-value lower than 0.01. When the filtered dataset were used, Fig. 4(b), all the expected links were detect, but more spurious links were detected and the lag amount detected also raised. Like the raw dataset, most of the p-value are under 0.01.



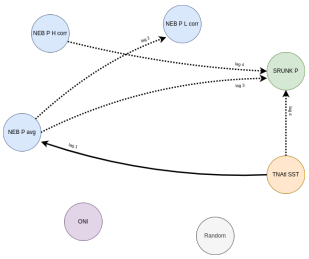
(a) raw dataset (b) filtered dataset

Fig. 4: Granger results. Solid lines are real relationships and dashed lines spurious ones. Due to the large amount of links, lag information for spurious links were suppressed, but they are shown in Appendix.

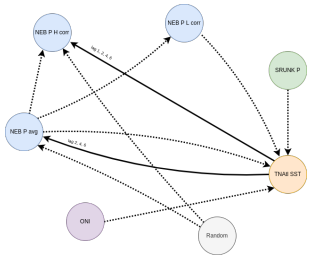
The results for PC and FullCI, which employed partial correlation, can be seen in Fig. 5. Using the raw dataset, the PC algorithm detected the causation from the TNAtl SST in the SRUNK and Northern NEB average precipitation, in time lag 4 and 1 respectively. The other detections are among precipitation time series. No link was detected when the filtered dataset was used, which explains why it was suppressed from Fig. 5.

The FullCI results for raw time series, Fig. 5(b), show that besides the detection about the TNAtl and the NEB precipitations, the other ones are spurious: i.e. precipitation causing SST variation and especially the random time series

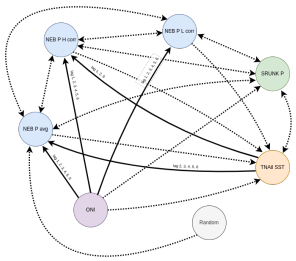
causing the NEB precipitations. Using the filtered dataset, Fig. 5(c), the amount of true links increased (there are five out of six), however the spurious ones and the detected lag amount also increased considerably. Most of links are bidirectional.



(a) PC (raw dataset)



(b) FullCI (raw dataset)



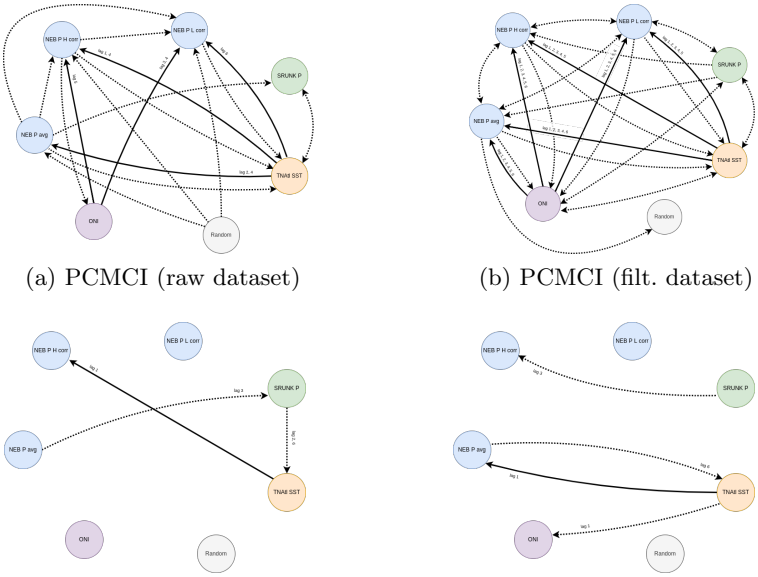
(c) FullCI (filtered dataset)

Fig. 5: Results of methods employing linear conditional independence test (Partial Correlation). PC result with filtered dataset was suppressed due to no link detection. Solid lines are real relationships and dashed lines spurious ones. Additional information can be found in Appendix.

The results for PCMCI and PCMCI+ employing linear conditional independence test are shown in Fig. 6. PCMCI succeed in detecting most of the expected links but the price paid was to also detect a lot of spurious links. When using filtered dataset, PCMCI detected all true links, but had much more spurious links and detected lags than the raw dataset results and had much more bidirectional links and much lower p-values. On the other hand, PCMCI+ had very few spurious links, but also very few true links. All links had p-value lower than 0.01.

When the non-linear conditional tests were employed, Fig. 7 and Fig. 8, no significant improvement was seen compared to Partial Correlation. Except for Fig. 7(c), which has four spurious links and PCMCI+ results had no links.

PCMCI with GPDC and CMI keeps detecting the correct links and had a slightly increase in the amount of spurious links, which most of them are bidirectional. Compared to filtered dataset, Fig. 7(b), PCMCI with GPDC using



(a) PCMCI (raw dataset) (b) PCMCI (filt. dataset)

(c) PCMCI+ (raw dataset) (d) PCMCI+ (filt. dataset)

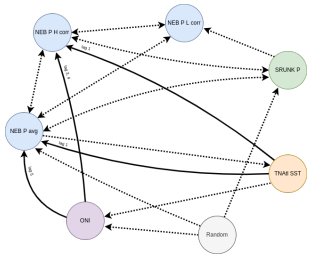
Fig. 6: Results of methods employing linear conditional independence test (Partial Correlation). Solid lines are real relationships and dashed lines spurious ones. Additional information can be found in Appendix.

raw dataset has lower link and lag amount, where the spurious links are among precipitation nodes themselves or caused by the random time series, precipitations and TNAtl SST. When the p-value threshold is decreased to 0.01, shown in Appendix, the spurious links among precipitation nodes are still present, and so the precipitation causing TNAtl SST.

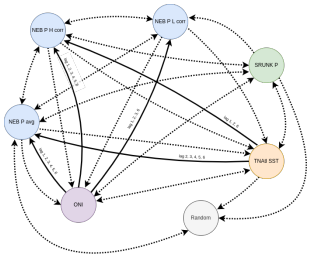
Employing CMI, as shown in Fig. 8, PCMCI also detected a lot of spurious links, i.e. precipitation causing ONI and TNAtl SST, and the random time series causing and being caused by precipitation nodes. Decreasing the p-value to lower than 0.01 does not help either to improve the accuracy.

One general conclusion regarding the method results that employ statistical test is that the method accuracy do not get much better when the threshold is decreased to 0.01 or lower. Most of link detection already have a very low p-value. When the filtered dataset was used, it considerably increased the occurrence of causal connections.

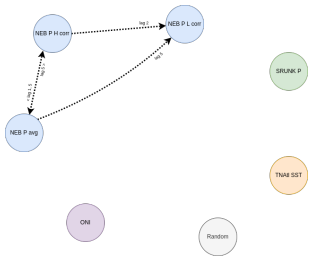
QRBS, SELVAR, SLARAC and LASAR have a scalar output and there is no preestablished threshold for these methods, then there are infinity thresholds that can be placed among data points. Thus, the approach employed is: if there is linear separability between classes (causal and non-causal), then it is established a threshold for each maximum lag value that could separate all the causal links from the non-causal links, with the lowest false positive rate possible. Fig. 9 show



(a) PCMCI (raw dataset)

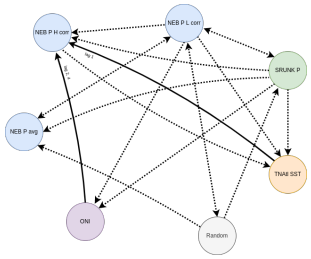


(b) PCMCI (filtered dataset)

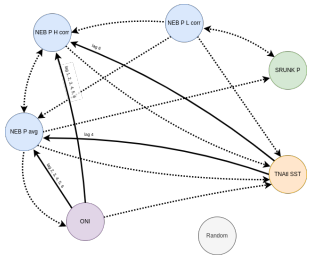


(c) PCMCI+ (raw dataset)

Fig. 7: Results of methods employing non-linear conditional independence tests GPDC. PCMCI+ result with filtered dataset was suppressed due to no link detection. Solid lines are real relationships and dashed lines spurious ones. Additional information can be found in Appendix.



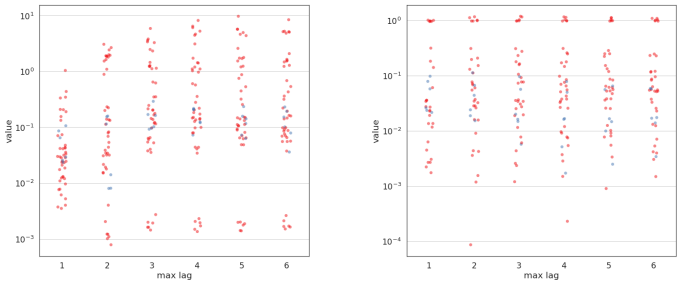
(a) PCMCI (raw dataset)



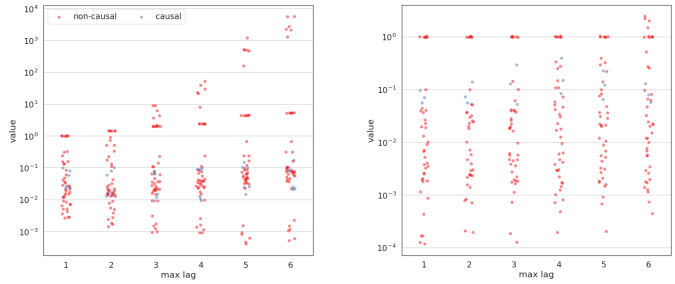
(b) PCMCI (filtered dataset)

Fig. 8: Results of methods employing non-linear conditional independence test CMI. PCMCI+ results were suppressed due to no link detection. Solid lines are real relationships and dashed lines spurious ones. Additional information can be found in Appendix.

the filtered dataset results and them clearly do not show any linear separability between classes, not allowing then, to establish a threshold.



(a) QRBS (b) SELVAR



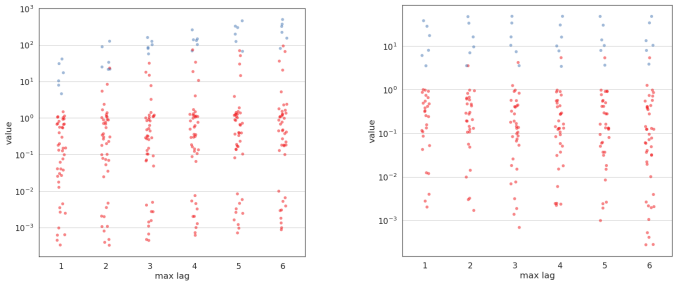
(c) SLARAC (d) LASAR

Fig. 9: Strip plots of the scalar output methods for filtered data, where y-axis is related to the causal detection and x-axis is the maximum time lag used for each execution, where one lag unit is one month.

Strip plots of scalar method outputs using the raw dataset are shown in Fig. 10, where most results had a linear separation between classes. SELVAR, Fig. 10(b), is the method that had the best separation between classes. QRBS and SLARAC also succeeded to separate classes but the decision boundary is not the same for all time lags. Finally, LASAR was the only method that did not succeed separating the classes properly.

Except for LASAR, the causal link outputs (blue dots) followed a sequence where the Northern NEB precipitation time series with the highest correlation had always the higher value, then the average precipitation one and finally, the NEB time series with the lowest correlation. Another pattern also happened in SELVAR, which the causal link outputs for TNAtl SST were always higher than the ones for ONI.In the four strip plots in Fig. 10, there is a red dot which sometimes appear near to an isolated blue dot on the boundary of the classes. It is correspondent to SRUNK average precipitation, which sometimes refers to the link with ONI, other times from TNAtl.

In general, they had much better performance and accuracy than the previous methods using both datasets. QRBS, SELVAR, and SLARAC correctly detected the causal effect of ONI and TNAtl SST on the Northern NEB precipitations.



(a) QRBS (b) SELVAR

(c) SLARAC (d) LASAR

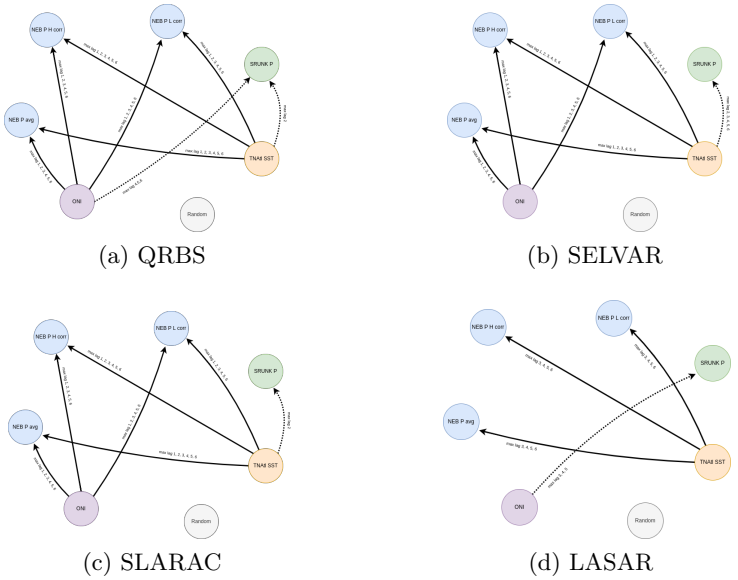
Fig. 10: Strip plots of the scalar output methods for raw dataset, where y-axis is related to the causal detection and x-axis is the maximum time lag used for each execution, where one lag unit is one month.

However, QRSB and LASAR detected the causal link from ONI to SRUNK precipitation.

The significant correlation values between SRUNK precipitation and TNAtl SST, shown in Fig. 1(b), was also detected by QRBS, SELVAR, and SLARAC. That fact is obviously not conclusive at all, but it may indicate a possible connection between them [15]. Another possibility is that there are other variables (confounders) that were not taken into account in this study that influence both variables creating a false impression of causal link.

4.3 Detection Comparison

Tab. 1 consolidates the result and allows a better comparison. The metrics for comparison are the True Positive rate, which is the percentage of correct links detected, and the False Positive rate which is the amount of spurious links detected divided by the total amount of possible spurious links. Even though some methods detect unitary time lags, the comparison will be done by the percentage of correct or incorrect links, because there is a group of methods that uses range of lags instead of unitary ones, thus making impossible a lag-wise comparison.



(a) QRBS (b) SELVAR

(c) SLARAC (d) LASAR

Fig. 11: Scalar method results for raw data. Solid lines are real relationships and dashed lines spurious ones.

The best methods are SELVAR and SLARAC with raw dataset, which detected all the expected links and just one spurious link each. QRBS also detected all expected links but detected two spurious ones, which one of them is the relationship $ONI \rightarrow SRUNK$. The most recent method from the statistical group, PCMCI+, had a very poor result for either datasets and conditional independence test types. In general, most of results with non-linear independence tests did not even detect a link.

In average, the results obtained using the filtered dataset increased the spurious link amount and lags detected. What was supposed to improve performance due to the theoretical removal of unnecessary information, produced the opposite effect. Analyzing each method individually, they performed better when employing partial correlation than non-linear tests.

4.4 Temporal Stability

This section presents some experiments that check the temporal stability of the best methods found in Tab. 1. The first one is regard growing time window, starting with 12 months and with a 6 month step, and has the objective to check the minimum window length and the method sensitivity over time. The second one employs temporal sliding windows, with 20 and 40 year length and also with a 6 month step, and has the objective to check if there are significant oscillations in the causal output with a fixed window size over time. The analysis

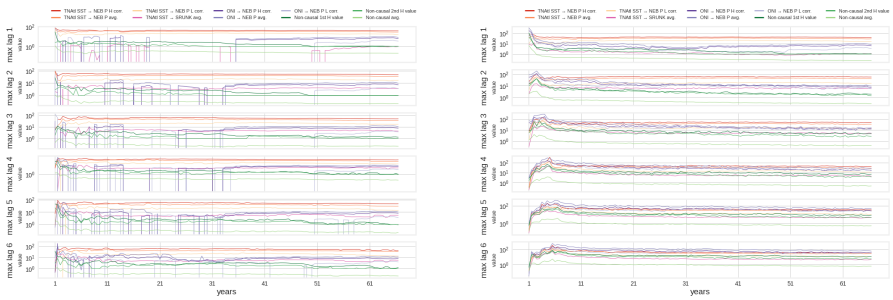
Table 1: Method ranking according to True Positive and False Positive rate.

Method	Dataset	TP rate	FP rate
SELVAR	raw	100%	2.8%
SLARAC	raw	100%	2.8%
QRBS	raw	100%	5.6%
PCMCI + ParCorr	filtered	100%	63.9%
Granger	filtered	100%	66.7%
PCMCI + GPDC	filtered	100%	77.8%
PCMCI + ParCorr	raw	83.3%	33.3%
Granger	raw	83.3%	52.8%
FullCI + ParCorr	filtered	83.3%	55.6%
PCMCI + GPDC	raw	66.7%	44.4%
LASAR	raw	50%	2.8%
PCMCI + CMI	filtered	50%	36.1%
FullCI + ParCorr	raw	33.3%	22.2%
PCMCI + CMI	raw	33.3%	44.4%
PCMCIPlus + ParCorr	raw	16.7%	8.3%
PCMCIPlus + ParCorr	filtered	16.7%	8.3%
PC + ParCorr	raw	16.7%	11.1%
PC + ParCorr	filtered	0%	0%
PCMCIPlus + CMI	raw	0%	0%
PCMCIPlus + CMI	filtered	0%	0%
PCMCIPlus + GPDC	filtered	0%	0%
PCMCIPlus + GPDC	raw	0%	11.1%

is done for TNAtl and ONI causing NEB precipitations, TNAtl causing SRUNK average precipitation, and the main (the highest and the second highest) non-causal links, as well the average for the rest of non-causal links. Due to the best results, the dataset used was the raw one.

The growing window experiment, Fig. 12, shows that SELVAR outputs had several discontinuities while SLARAC did not. The time series in SELVAR that had discontinuities are: *ONI* $\rightarrow$ *NEB P H corr.*, *ONI* $\rightarrow$ *NEB P avg.*, *ONI* $\rightarrow$ *NEB P L corr.*, *TNAtl SST* $\rightarrow$ *SRUNK avg.*, and *TNAtl SST* $\rightarrow$ *NEB P L corr.* Another point to note in SELVAR is that the highest (non-causal 1st H value) and the second highest (non-causal 2nd H value) value for non-causal links in max lag 1 are almost always higher than the *TNAtl SST* $\rightarrow$ *SRUNK avg.* link, while on the other max lags that behavior did not happen after 12-year length. The behavior seen in SLARAC outputs are totally different. After 10-year length almost all max lags got stabilized. The only one that have some oscillations was in max lag 1, where ONI links do not go so straight. Analyzing all max lags, it is possible to see that ONI outputs become higher than TNAtl ones and as the max lag get higher, the non-causal average value tends to stay more time near to 1.

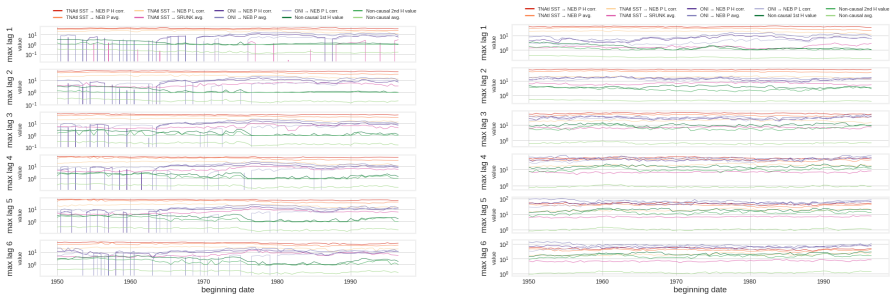
The second experiment uses sliding windows, which lengths – 20 and 40 years – were chosen, because it is when SLARAC and SELVAR start to get stabilized. SELVAR had plenty of discontinuities before the length of 35 years. As shown in Fig. 13(a), with a sliding window of 20-year length it still has a lot of discontinuities, but most of them vanishes after 1980. Except for *TNAtl SST* $\rightarrow$ *SRUNK avg.* and its discontinuities in max lag with 1, after 1980 the



(a) SELVAR (b) SLARAC

Fig. 12: Results using raw data and growing window with 12 months of initial length and 6 months of incremental step.

separability of causal and non-causal has the same aspect what was seen with full-length time series, Fig. 12(a).



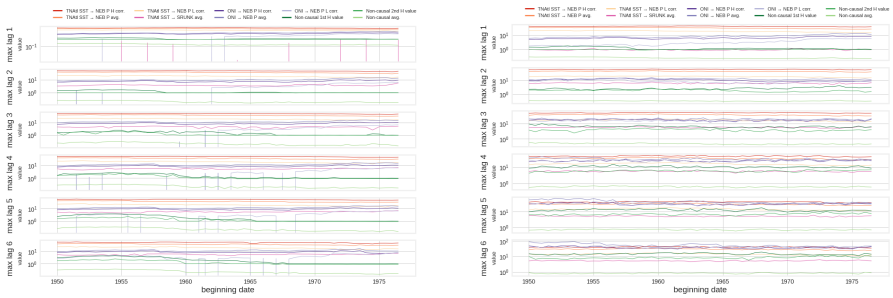
(a) SELVAR (b) SLARAC

Fig. 13: Results using raw data and sliding window with 20 year length.

The SLARAC results for the 20-year sliding window show almost the same behavior of the full-length time series: ONI links get higher than TNAtl links over the max lags, all the outputs are sort of stable except for max lag 1, and *TNAtl SST* → *SRUNK avg.* is higher than the other non-causal time-series in max lag 2. The differences are: *TNAtl SST* → *SRUNK avg.* get detached from the non-causal class after 1987 in max lag 1, and most of times in max lag 4, 5, and 6 *TNAtl SST* → *NEB P L corr.* stays inside or very near to the non-causal class.

When a 40-year length sliding window was used, Fig. 14, the outputs in SELVAR got much less discontinuities when comparing to the 20-year window, a fact already expected. The only time series that still experienced discontinuities

were *TNAtl SST → SRUNK avg.* in max lag 1 and *TNAtl SST → NEB P L corr.* in all max lags. On the other hand, SLARAC had pretty much the same behavior of the prior sliding window experiment.



(a) SELVAR (b) SLARAC

Fig. 14: Results using raw data and sliding window with 40 year length.

A point of attention regarding the prior charts – Fig. 12, 13, and 14 – is that *TNAtl SST → SRUNK avg.*, *TNAtl SST → NEB P L corr.*, and *ONI → NEB P L corr.*, had several issues in the growing window or sliding window experiments, what may suggest that their causal link may not be so strong as the other ones or not that consistent over time.

5 Conclusion

Several studies have employed causal detection methods in climate time series, but none of them in the connection of ENSO and ITCZ to the Northeastern Brazil precipitation. Moreover, it is not clear on the literature what scenarios each method performs better. Then, the goal of this study was to evaluate the performance of time series causal detection methods on the aforementioned phenomena. Nine methods were used, but only two had a satisfactory performance: SLARAC and SELVAR. Moreover, the employment of filtered time series also degraded the detection performance of the methods.

There are two facts that must be highlighted about the top three methods – SELVAR, SLARAC, and QRBS. First, besides the good performance, they were even able to detect causation properly even when the correlation was very low. That is the case of the link from ONI to NEB precipitation time series. Second, they detected causation of TNAtl SST on SRUNK precipitation which they have a significant correlation for the first lags. It is definetly not conclusive, however, it may suggest the existence of a real connection or it may be simply just the result of a confounder variable that was not considered in this study.

The temporal stability experiments did not show any other significant oscillation, besides the already expect discontinuities in SELVAR. Nevertheless, the

discontinuities of *TNAtl SST* $\rightarrow$ *SRUNK avg.* and *ONI* $\rightarrow$ *NEB P L corr.* in SELVAR and the fact *TNAtl SST* $\rightarrow$ *NEB P L corr.* most of the time (SLARAC with max lag 4, 5, and 6) stayed within or quite near to the non-causal cluster, may suggest that their causal link strength may not be as strong as the other ones.

The employment of SLARAC and SELVAR had good results indeed, but there are some caveats to be considered. It is not possible to know the exact time lag of a causal connection detection. The second point is that the window length is a very important parameter, which the longest, normally the better and the more stable. Another critical parameter is the threshold for causal and non-causal separability. In scenarios where there is no ground truth for finding the best threshold, some experiments should be firstly conducted in order to define the value range of causal and non-causal samples.

References

1. Ambrizzi, T., de Souza, E.B., Pulwarty, R.S.: The hadley and walker regional circulations and associated enso impacts on south american seasonal rainfall. In: The Hadley circulation: present, past and future, pp. 203–235. Springer (2004)
2. Canedo-Rosso, C., Uvo, C.B., Berndtsson, R.: Precipitation variability and its relation to climate anomalies in the bolivian altiplano. International Journal of Climatology **39**(4), 2096–2107 (2019)
3. Di Capua, G., Kretschmer, M., Donner, R.V., van den Hurk, B., Vellore, R., Krishnan, R., Coumou, D.: Tropical and mid-latitude teleconnections interacting with the indian summer monsoon rainfall: A theory-guided causal effect network approach. Earth System Dynamics **11**, 17–34 (2020)
4. Du, X., Hendy, I., Hinnov, L., Brown, E., Zhu, J., Poulsen, C.J.: High-resolution interannual precipitation reconstruction of southern california: Implications for holocene enso evolution. Earth and Planetary Science Letters **554**, 116670 (2021)
5. Eichler, M.: Causal inference in time series analysis. Wiley Online Library (2012)
6. Garcia, S.R., Kayano, M.T.: Some evidence on the relationship between the south american monsoon and the atlantic itcz. Theoretical and Applied Climatology **99**(1), 29–38 (2010)
7. Glantz, M.H., Ramirez, I.J.: Reviewing the oceanic niño index (oni) to enhance societal readiness for el niño's impacts. International Journal of Disaster Risk Science **11**, 394–403 (2020)
8. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society pp. 424–438 (1969)
9. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: Problems and methods. arXiv preprint arXiv:1809.09337 (2018)
10. Hastenrath, S.: Prediction of northeast brazil rainfall anomalies. Journal of Climate **3**(8), 893–904 (1990)
11. Hastenrath, S.: Circulation and teleconnection mechanisms of northeast brazil droughts. Progress in Oceanography **70**(2-4), 407–415 (2005)
12. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

13. Kane, R.: Prediction of droughts in north-east brazil: Role of enso and use of periodicities. International Journal of Climatology **17**(6), 655–665 (1997)
14. Kretschmer, M., Coumou, D., Donges, J.F., Runge, J.: Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. Journal of Climate **29**(11), 4069–4081 (2016)
15. Lim, Y.K.: The east atlantic/west russia (ea/wr) teleconnection in the north atlantic: climate impact and relation to rossby wave propagation. Climate Dynamics **44**(11-12), 3211–3222 (2014)
16. Lin, J., Qian, T.: A new picture of the global impacts of el nino-southern oscillation. Scientific Reports **9**(1), 1–7 (2019)
17. Malinsky, D., Danks, D.: Causal discovery algorithms: A practical guide. Philosophy Compass **13**(1), e12470 (2018)
18. Marengo, J.A., Torres, R.R., Alves, L.M.: Drought in northeast brazil—past, present, and future. Theoretical and Applied Climatology **129**(3-4), 1189–1200 (2017)
19. Pearl, J.: Causality: models, reasoning and inference, vol. 29. Springer (2000)
20. Pearl, J., et al.: Causal inference in statistics: An overview. Statistics surveys **3**, 96–146 (2009)
21. Runge, J.: Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos: An Interdisciplinary Journal of Nonlinear Science **28**(7), 310 (2018)
22. Runge, J.: Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In: Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence. vol. 124, pp. 1388–1397. PLMR (2020)
23. Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M.D., Muñoz-Marí, J., et al.: Inferring causation from time series in earth system sciences. Nature communications **10**(1), 2553 (2019)
24. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances **5**(11), 4996 (2019)
25. Runge, J., Petoukhov, V., Donges, J.F., Hlinka, J., Jajcay, N., Vejmelka, M., Hartman, D., Marwan, N., Paluš, M., Kurths, J.: Identifying causal gateways and mediators in complex spatio-temporal systems. Nature communications **6**, 8502 (2015)
26. Runge, J., Petoukhov, V., Kurths, J.: Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. Journal of climate **27**(2), 720–739 (2014)
27. Shaman, J.: The seasonal effects of enso on european precipitation: Observational analysis. Journal of Climate **27**(17), 6423–6438 (2014)
28. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Social science computer review **9**(1), 62–72 (1991)
29. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
30. Torrence, C., Compo, G.P.: A practical guide to wavelet analysis. Bulletin of the American Meteorological society **79**(1), 61–78 (1998)
31. Weichwald, S., Jakobsen, M.E., Mogensen, P.B., Petersen, L., Thams, N., Varando, G.: Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In: NeurIPS 2019 Competition and Demonstration Track. pp. 27–36. PMLR (2020)