INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. (2016)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.4649

RMetS
Royal Meteorological Society

Cointegration modelling for empirical South American seasonal temperature forecasts

Alemtsehai A. Turasie[a*] and Caio A. S. Coelho[b]

[a] *School of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, South Africa*
[b] *Centro de Previsao de Tempo e Estudos Climaticos (CPTEC), Instituto Nacional de Pesquisas Espaciais (INPE), Cachoeira Paulista, Brazil*

ABSTRACT: This study investigates an alternative modelling approach for empirical seasonal temperature forecasts over South America. Seasonal average temperatures are found to be non-stationary at most parts of South America over the 1949–2012 period. Simple persistence and lagged regression methods have considerable correlation skill in forecasting next season temperature using previous season temperature as predictor. However, the presence of trends in both predictor and predictand temperature variables can affect correlation skill. Models that can account for non-stationarity in these variables may do better in modelling and forecasting seasonal temperatures known to have trends. A novel method (cointegration), introduced here for empirical seasonal climate forecasting, is found to perform better than the traditional persistence and regression forecasts for places where the predictor and predictand temperatures have stochastic trends. Potential skill pairwise comparisons between temperature forecasts produced with cointegration and those produced using persistence and lagged regression have shown that the alternative cointegration method performs significantly better than the other two. One of the main reasons for the better performance of cointegration method is that the modelling procedure accounts for the existing non-stationarity in the process, and thus enables the estimated model to predict out of the range as efficiently as possible. Overall, this method appears to be ideal for modelling and predicting climate under the current global warming scenario. This is because most of the climatic variables including temperature in particular cannot be assumed to be stationary through time under such warming scenario.

KEY WORDS empirical prediction; seasonal temperature; cointegration; VAR; regression; persistence

Received 13 July 2015; Revised 21 October 2015; Accepted 9 December 2015

1. Introduction

Seasonal forecasting aims to estimate the change in the likelihood of a climatic event happening in the future. It attempts to predict deviations from climatology of the weather averaged generally over 2–4 months (Jan van Oldenborgh *et al.*, 2005). A number of methods have been proposed in literature for empirical seasonal forecasting (e.g. Barnston and He, 1996; Barnston and Smith, 1996; Landman and Mason, 2001; Colman and Davey, 2003). However, these methods generally do not pay due attention to the training data time series behaviour. Particularly, stationarity/non-stationarity of the training data is a feature strongly advised to be examined critically beforehand. This examination is required because some methods such as regression/correlation analysis may end up with misleading findings in terms of forecast skill for non-stationary data. This situation is referred to as 'spurious regression' by Granger and Newbold (1974).

Simple persistence and lagged regression methods are some of the commonly employed methods in empirical forecasts of seasonal climate (e.g. Coelho *et al.*, 2004, 2006). This study introduces a relatively novel cointegration modelling approach, which accounts for inherited non-stationarity in the data, for seasonal temperature forecast. A comparative analysis is presented in this article between forecasts produced with the cointegration method and two existing methods, namely lagged regression and persistence forecasts. These forecast methods are applied to seasonal temperature forecasts over South America, and are compared and verified over the 1949–2012 period. We consider case of forecasting March–April–May (MAM) seasonal temperature using previous season, November–December–January (NDJ), temperatures as predictor.

The article is organized as follows. In Section 2, we discuss the source, type and time series properties of the data used in this study. Section 3 presents some of the widely known empirical forecast methods (persistence and lagged regression) and their correlation skill. Section 4 discusses how trends in climatic variables can affect regression and correlation analyses results, which may lead to spurious regression/correlation problems. After showing the impact of stochastic trends on regression/correlation, Section 5 presents a commonly used statistical method for testing the presence of stochastic trends in seasonal average temperatures. Having examined the presence of stochastic trends

* Correspondence to: A. A. Turasie, School of Statistics and Actuarial Science, University of the Witwatersrand, Private Bag 3, Johannesburg, 2050, South Africa. E-mail: alemtsehai.turasie@wits.ac.za; alemtsehaia@yahoo.co.uk

in seasonal temperatures, we next introduce the cointegration model, which is designed to model relationships among two or more non-stationary time series. Section 6 discusses a bivariate cointegrating vector autoregressive (VAR) model with an application for two non-stationary seasonal temperature time series. Section 7 compares the skill of seasonal temperature forecasts produced with cointegration with those of persistence and lagged regression. Finally Section 8 summarizes the article.

2. Temperature time series assessment

In this study, we use the Global Historical Climatology Network (GHCN)/Climate Anomaly Monitoring System (CAMS) 2 m temperature analysis (0.5×0.5) data (Fan and Van den Dool, 2008). This is a high resolution (0.5×0.5 degrees in latitude and longitude) global land surface temperature data set covering the period 1948 to near present. Seasonal averages are extracted from this monthly data set by averaging over consecutive 3 months. We consider data for January 1948 to December 2012. Thus, we have time series of monthly and seasonal average surface temperatures of length 65 for each grid point.

Assessing the time series properties of the data is essential in determining which model to use for forecasting purposes. Most of the statistical models, and the least squares regression in particular, assume stationarity over time. Using non-stationary data to fit such models may lead to spurious results as it will be discussed later in Section 4. Thus, examining how temperatures behave through time is advisable before choosing specific methods for modelling and forecasting purposes. Non-stationarity may be caused by either deterministic or stochastic trends. In this section, we assess linear deterministic time trends in seasonal temperatures. Section 5 will later assess stochastic trends. By stationary time series, we refer to time series whose statistical properties such as mean and variance are constant through time and autocorrelations depend only on length of time lags.

For temperature T_t (T_{NDJ_t}: NDJ temperature and T_{MAM_t}: MAM temperature) at year t, a simple test of significance for deterministic linear time trend for the period of 1949–2012 is carried out on $T_t = \alpha + \beta t + \varepsilon_t$, where α and β are unknown constants with β measuring changes in seasonal temperatures per year and ε_t is the random-error term assumed to be independently and identically distributed normal variate with zero mean and constant variance.

Locations with significant time trend, at 5% level of significance, are shown in Figure 1. This figure shows that seasonal temperatures have significant linear time trends in most parts of the continent. Presence of trends can affect the reliability of the ordinary least square (OLS) regression estimates and the correlation coefficient between predictor and predictand variables in such regression models. For example, if T_{NDJ} is used as predictor for T_{MAM} and both T_{NDJ} and T_{MAM} have trends in the same direction, the resulting positive correlation between T_{NDJ} and T_{MAM} could be due to the presence of trends in both predictor and predictand variables (i.e. not necessarily due to the ability of T_{NDJ} to be a good predictor for T_{MAM}).

3. Some simple forecasts and their potential skill

3.1. Persistence forecast

During February, having just observed the previous NDJ temperature, one could think of a simple persistence forecast by assuming the previous NDJ temperature to persist into next MAM. In this simple forecasting model, the mean temperature of MAM at year t (T_{MAM_t}) is forecast to be the same as the observed mean temperature of NDJ at year $t-1$ $\left(T_{\mathrm{NDJ}_{t-1}}\right)$, i.e. $T_{\mathrm{MAM}_t} = T_{\mathrm{NDJ}_{t-1}}$.

This is one of the simplest, sometimes referred to as a lazy person's, forecast with lesser expense involved. However, due to the fact that the atmosphere (and especially the ocean) vary slowly, except in special and somewhat rare circumstances, this sort of forecast usually has higher skill than climatology, which uses the historical (mean) value for a particular observed period as the forecast for the future (Van den Dool, 2006).

Skill is here assessed using correlation. Correlation indicates potential skill, measuring association between the deterministic forecast (i.e. the mean of the forecast distribution) and observations. This skill measure is not affected by the existence of biases in forecast. Figure 2(a) shows correlation skill of persistence of previous NDJ temperatures into next MAM. This may imply that the previous season temperature value is a good proxy for the next season temperature value.

The persistence of the previous NDJ mean temperature into next MAM mean temperature provides the central (mean) value of the forecast distribution. In order to provide an uncertainty estimate for this persistence forecast, the NDJ temperature cross-validated variances are estimated leaving 1 year out at a time. That is, forecast variance s_i^2 for i^{th} year is computed as $s_i^2 = \frac{\sum_{t=1,t\neq i}^{n}\left(T_{\mathrm{NDJ}_t} - \overline{T}_{\mathrm{NDJ}_i}\right)^2}{n-2}$ where $\overline{T}_{\mathrm{NDJ}_i} = \frac{\sum_{t=1,t\neq i}^{n} T_{\mathrm{NDJ}_t}}{n-1}$ for $i = 1, 2, \cdots, n$, $t = 1, 2, \cdots, n$ and $n = 65$ is number of years. These forecast variances will be used latter for constructing and assessing the skill of probability forecasts in Section 7.2.

3.2. Lagged regression forecast

An alternative simple linear regression model of next MAM temperature (T_{MAM_t}) on previous year's NDJ temperature ($T_{\mathrm{NDJ}_{t-1}}$)

$$T_{\mathrm{MAM}_t} = \alpha_t + \beta_t T_{\mathrm{NDJ}_{t-1}} + \varepsilon_t \tag{1}$$

is estimated using *leave-one-out* cross validation, and the respective set of retrospective forecasts (hindcasts) are made using the estimated model. Here α_t is mean MAM temperature for year t when previous NDJ

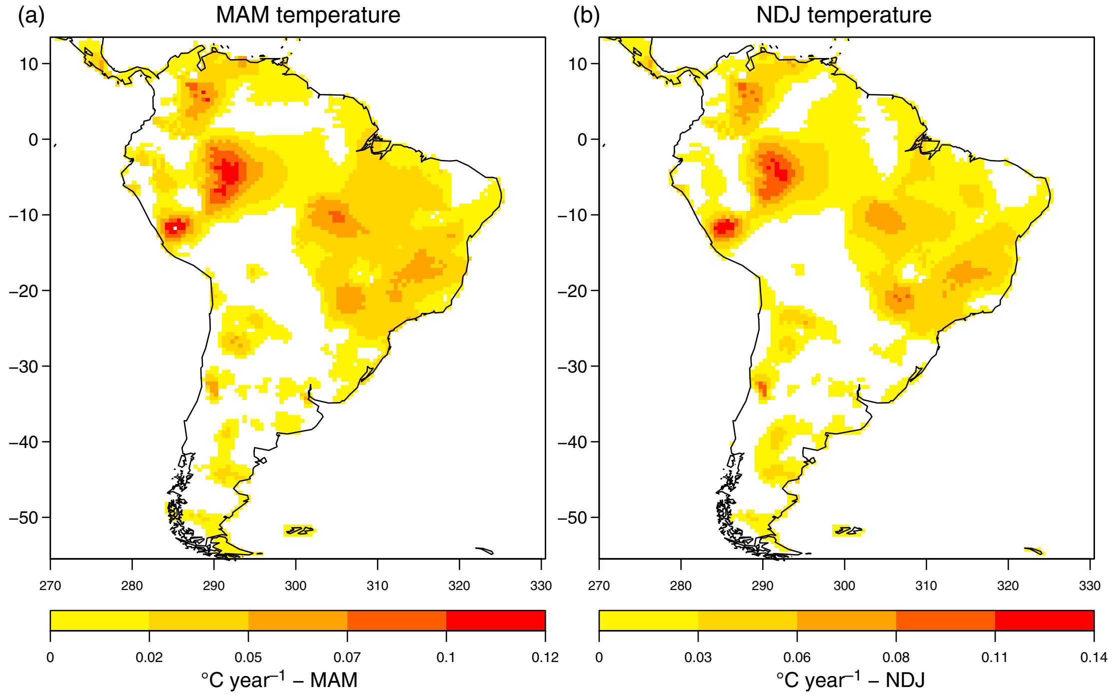

Figure 1. Coefficient of linear deterministic time trends (°C year^{-1}): (a) MAM for the 1949–2012 period and (b) NDJ for the 1949–2011 period. For the areas in white either the time-trend coefficients were not found to be statistically significant at the 5% level or are missing numerical data.

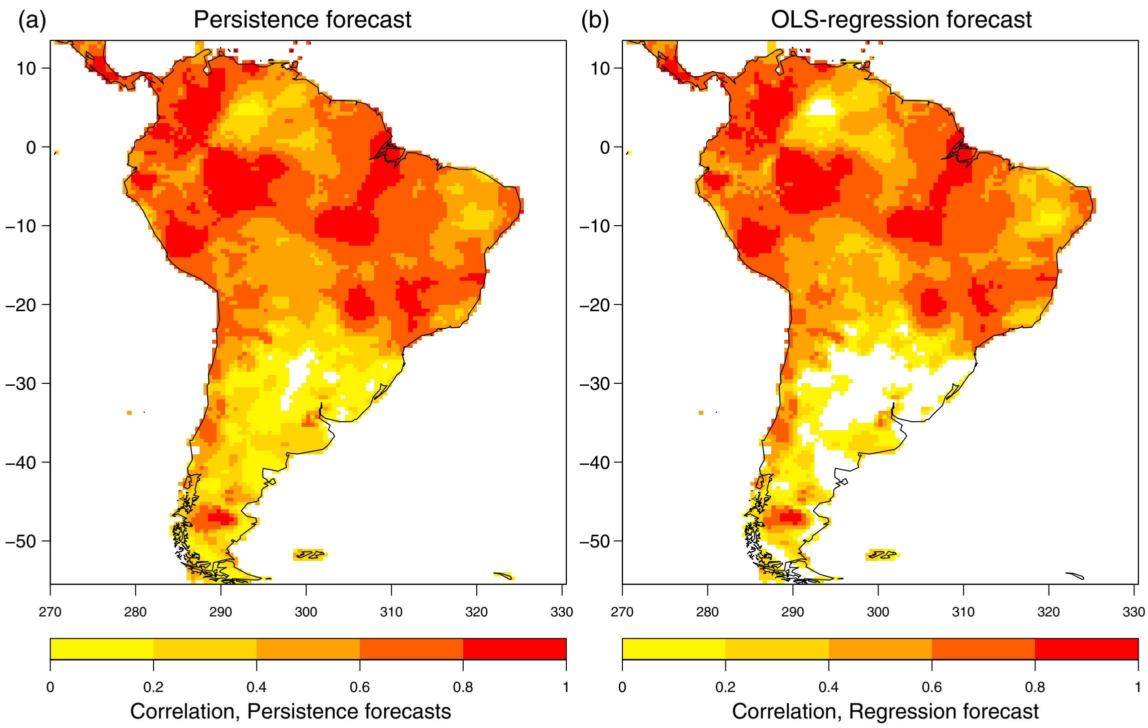

Figure 2. Maps of correlation skill for the 1949–2012 period: (a) persistence and (b) OLS regression forecasts for MAM mean temperature using previous NDJ temperature as predictor. For the areas in white either the correlation coefficients were not found to be statistically different from zero at the 5% level or they were not computable due to missing numerical data.

temperature is 0 °C, β_t is a change in MAM temperature for year t for every 1 °C change in previous NDJ temperature and ε_t is a random-error term assumed to be distributed as $\varepsilon_t \overset{ii}{\sim} N\left(0,\sigma^2\right)$. The two time-dependent parameters α_t and β_t, which are functions of $T_{\mathrm{NDJ}_{t-1}}$, are estimated by the OLS method using all previously observed NDJ temperature values except the ones corresponding to the forecast year (i.e. using one year out cross validation). Correlation skill of the resulting retrospective forecasts is shown in Figure 2(b).

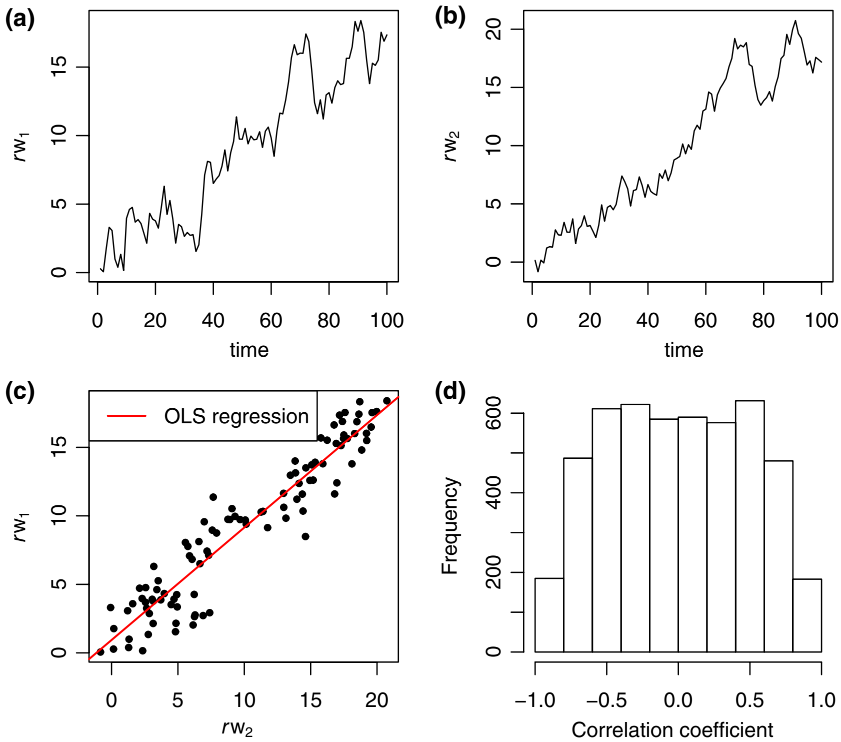

Figure 3. Graphical display of independently simulated random-walk processes and their linear relationship: (a) time series of rw_1 (b) time series of rw_2 (c) scatter plot of rw_1 versus rw_2 and the fitted regression line and (d) histogram showing frequency distribution of correlation coefficients for pairs of independently simulated random walks.

The i^{th} individual forecast variance, s_i^2, in this case is computed as $s_i^2 = \hat{\sigma}_i^2 \left(1 + \frac{1}{n-1} + \frac{\left(T_{\mathrm{NDJ}_i} - \overline{T}_{\mathrm{NDJ}_i}\right)^2}{\sum_{t=1, t\neq i}^{n} \left(T_{\mathrm{NDJ}_t} - \overline{T}_{\mathrm{NDJ}_i}\right)^2} \right)$, where $\overline{T}_{\mathrm{NDJ}_i} = \frac{\sum_{t=1,t\neq i}^{n} T_{\mathrm{NDJ}_t}}{n-1}$ and $\hat{\sigma}_i^2$ is the estimated residual variance for that period, which is estimated as $\hat{\sigma}_i^2 = \frac{\sum_{t=1,t\neq i}^{n} \left(T_{\mathrm{MAM}_t} - \hat{T}_{\mathrm{MAM}_t}\right)^2}{n-3}$, where $\hat{T}_{\mathrm{MAM}_t}$ is t^{th} period fitted value of T_{MAM} using estimates based on all but the i^{th} observation for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, n = 65$.

It is evident from Figures 1 and 2 that the correlation skill is stronger at locations with stronger time trend. However, we could not attribute the stronger correlation skills to stronger time trends at this stage. Existence of long-term linear relationship between the two seasonal temperatures should be studied before such inference (see Section 6).

One can detrend the series (i.e. remove the time trend) to address this type of non-stationarity. However, this simple test of significance for deterministic time trend may be highly influenced by presence of stochastic trend. Thus, before detrending a series, one should study if there is a stochastic trend therein using more reliable specialized techniques (see Section 5.1).

4. Spurious regression/correlation problems

Classical estimation methods are valid for stationary series. Many observed climatic time series such as temperature, however, have empirical features that are inconsistent with the assumptions of stationarity. For such non-stationary variables, the OLS regression/correlation may end up with spurious results. In spurious regression/correlation, the empirical estimates are computed correctly but their relationship implications are nonsensical or unreasonable.

For example, consider pairs of independently simulated random-walk processes in Figure 3. The two variables, namely rw_1 and rw_2, plotted in panel (a) and (b) appear to be related simply because they both trend upwards in the same manner. The scatter plot of these two variables around an estimated linear regression line, in panel (c) of the same figure, also shows a linear relationship between the two variables. However, from prior knowledge of the process that generated rw_1 and rw_2, these two variables are completely independent to each other and there is no actual linear relationship between them. Moreover, a number of such pairs of independent random-walk processes were simulated and the correlation between the pairs was estimated. Panel (d) of Figure 3 shows the frequency histogram of the estimated correlations, where one can see that a considerable number of these independent pairs have correlation (which is supposed to measure strength and direction of linear relationship between variables) as strong as 0.5, in absolute value, and more. These strong correlations are induced due to the existence of trends in the two variables.

Yule (1926) first discussed the risk of regressing a variable that contains trend on another unrelated variable that also contains trend, the so-called ‘nonsense regression’ problem. Granger and Newbold (1974) called such estimates ‘spurious regression’ results and Phillips (1986) explained them in more depth. For regressions that relate integrated (contain stochastic trend) processes, Phillips (1986) showed that the usual t and F -ratio test statistics do not possess limiting distributions but actually diverge as the sample size increase. In such cases, the null hypothesis of no significant relation is rejected too often for a given critical value and the regression is characterized by a higher value of estimated coefficient of determination, R^2, given by the ratio of variation explained by the model to the total variation. In order to avoid such spurious results, when dealing with non-stationary climatic variables, it is highly advised to critically assess the type of non-stationarity and then test for the existence of a stationary long-run relationship between the non-stationary predictand and predictor variables.

5. Investigating stochastic trends in seasonal temperature

The time series plot of MAM temperature for a specific grid point (lon = 65.75°W, lat = 5.25°S) in Figure 4, shows that the trend in seasonal temperature for that location has more stochastic behaviour than deterministic.

Therefore, the unit root test (a test for stochastic trend described in Section 5.1) is required before commenting on forecasting methods for non-stationary time series.

5.1. Unit root test

The unit root test is a tool for assessing the presence of a stochastic trend in a time series. The widely known augmented Dickey–Fuller (ADF) method is used to test for unit root in this study. Dickey and Fuller (1981) proposed different hypotheses to test for a unit root in univariate time series. Sequence of tests, which are based on each of the test regression, as in Equation (2) for instance, treating the initial value T_0 as fixed, are proposed. To balance between minimizing the serial correlation in the error term (due to small lag-length) and maximizing power of the test (overestimating lag-length reduces power of the test), two lagged endogenous variables (T_{t-1} *and* T_{t-2}) have been included in the test regression for all grid points.

$$T_t = \mu + \beta t + \pi_{11} T_{t-1} + \pi_{12} T_{t-2} + \varepsilon_{1t} \tag{2a}$$

$$T_t = \mu + \pi_{21} T_{t-1} + \pi_{22} T_{t-2} + \varepsilon_{2t} \tag{2b}$$

$$T_t = \pi_{31} T_{t-1} + \pi_{32} T_{t-2} + \varepsilon_{3t} \tag{2c}$$

Rewriting Equation (2) in time differenced form, but without affecting its likelihood, one gets:

$$\Delta T_t = \mu + \beta t + \rho_1 T_{t-1} + \gamma_1 \Delta T_{t-1} + \varepsilon_{1t} \tag{3a}$$

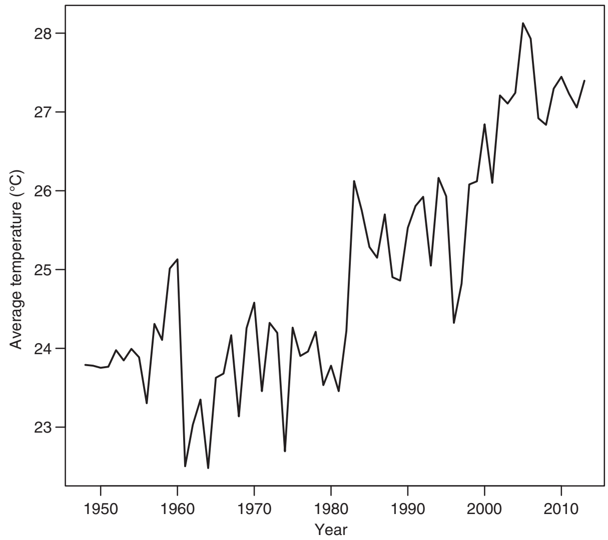

Figure 4. Time series of MAM temperature for a location at longitude = 65.75°W, and latitude = 5.25°S over the period 1949–2012.

$$\Delta T_t = \mu + \rho_2 T_{t-1} + \gamma_2 \Delta T_{t-1} + \varepsilon_{2t} \tag{3b}$$

$$\Delta T_t = \rho_3 T_{t-1} + \gamma_3 \Delta T_{t-1} + \varepsilon_{3t} \tag{3c}$$

where $\rho_i = (\pi_{i1} + \pi_{i2} - 1)$, and $\gamma_i = -\pi_{i2}$. Non-stationary time series that can be made stationary after differencing are known as *difference stationary* (DS) process. The process in (2) is said to contain a unit root, if it is a DS and its counterpart in Equation (3) is not DS. If that is true, then $\rho_i = 0 \Rightarrow \pi_{i1} + \pi_{i2} = 1$.

The rule of thumb (Pfaff, 2008) is to start testing for unit root from Equation (3a) and proceed to Equation (3c). If the null hypothesis of *unit root* based on Equation (3a) is rejected, then, there is no need of proceeding with tests for different versions of unit root hypotheses. Here, we present results of the test based on Equation (3c). Thus, we test the null hypothesis $H_0 : \rho_3 = 0$ versus alternative, $H_A : \rho_3 < 0$. A t -type test statistic, τ, whose distribution is slightly non-symmetrical (see Dickey and Fuller (1979); Hamilton (1994) for details) is used to test this hypothesis.

The null hypothesis of unit root is rejected if the test statistic is less than the critical value. We use a 95% critical value, −1.95, given in Table B.6 (Case 1) of Hamilton (1994). This test is applied for both predictor and preditand time series of seasonal average temperatures at each grid point. This is a test to check if a given univariate time series, say T_t , is stochastically non-stationary or not (see Figure 4 for an example of time series with stochastic trend). For the ADF test, which we used in this article, the null hypothesis is that ‘there is a stochastic trend in the data (T_t).’ Under this null hypothesis, the test statistics will have different asymptotic distributions based on the type of model (Equation 3(a)–3(c)) on which the test is based on. These distributions are different from the traditional t (for Equation 3(c)) and F (for Equation 3(a) and (b)) distributions. Therefore, here we are testing if the time

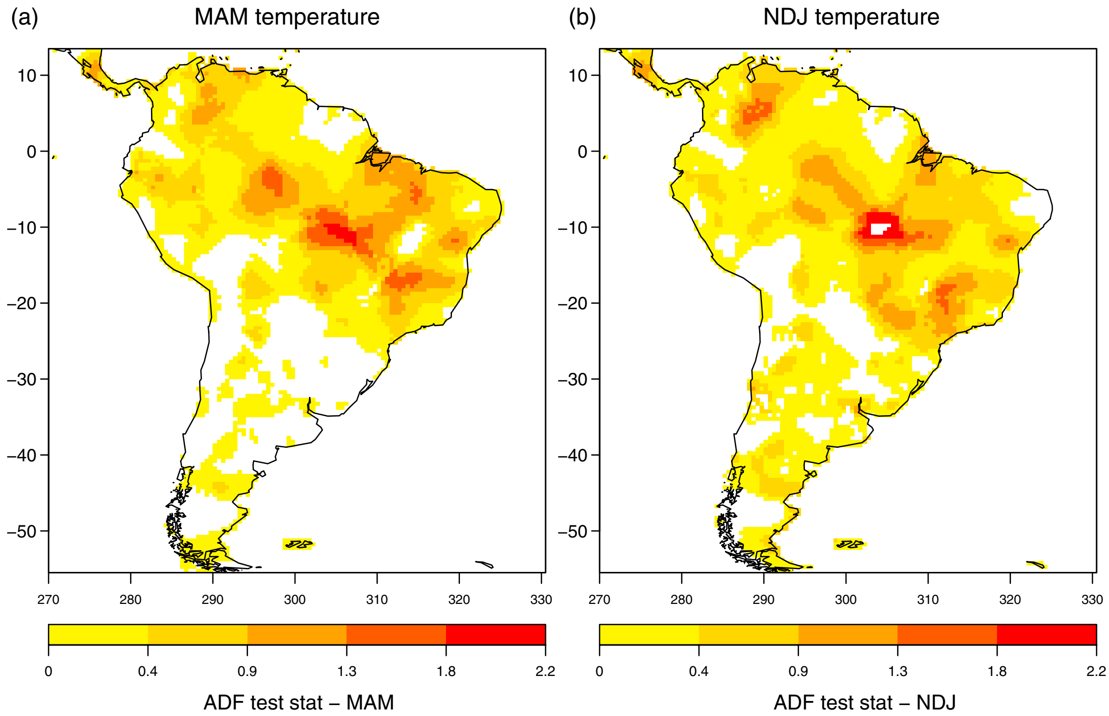

Figure 5. ADF unit root test statistic for (a) MAM and (b) NDJ temperatures: Map of significant ADF test statistics at 5% level. For the areas in white either the ADF unit root test statistics was found to be less than the critical value of −1.95 indicating non existence of stochastic trends or the test statistic is not computable due to missing numerical data.

series of seasonal temperatures (T_t) at each grid point is stochastically non-stationary.

As Figure 5(a) and (b) shows the ADF test statistic is not less than the critical value of −1.95 at most parts of South America. This implies that seasonal average temperatures contain unit roots (stochastic trends) in most parts of the continent. Such time series are called integrated of order 1, denoted $I(1)$, if they can be made stationary after differencing once and $I(d)$ if one needs d times differencing to make them stationary.

Therefore, use of the OLS regression such as (1) and correlation skills without assessing the existence of a stationary long-run relationship (cointegrating) between the predictand and predictor temperatures, discussed in Section 6, may lead to a spurious result problem for those locations with unit root.

6. Cointegration modelling

The recognition that most time series are non-stationary profoundly altered the methodologies of econometrics, introducing the concept and tools associated with integrated–cointegrated data (Hendry and Juselius, 2001). Components of an integrated process $\mathbf{z}_t$ are said to be cointegrated of order d and b, denoted $\mathbf{z}_t \sim \mathrm{CI}(d, b)$, if (1) each component of $\mathbf{z}_t$, when considered individually, is integrated of order d, $I(d)$; and (2) there exists at least one (and possibly r) cointegrating vectors $\beta \neq 0$ such that $\beta' \mathbf{z}_t$ is integrated of order $d-b$, $I(d-b)$, for $b = 1, 2, \cdots, d$, $d = 1, 2, \cdots$ (Engle and Granger, 1987).

Thus, cointegration implies that certain linear combinations of the variables are integrated of lower order than the process itself. For instance, two time series y_t and x_t that are both integrated of order one, $I(1)$, are said to be cointegrated if there exists a parameter β such that $u_t = y_t - \beta x_t$ is a stationary, $I(0)$, process.

To find out which linear relation is stationary and which is not, a time series modelling of the full system of equations is required. To this end, one can use a VAR model, where each variable is explained by its own time-lagged values and the time-lagged values of all other variables in the system.

Let $\mathbf{z}_t$ be a 2×1 vector of two temperature series, T_{MAM_t} (next MAM temperature) and $T_{\mathrm{NDJ}_{t-1}}$ (previous year's NDJ temperature), $\mathbf{z}_t = \begin{pmatrix} T_{\mathrm{MAM}_t} \\ T_{\mathrm{NDJ}_{t-1}} \end{pmatrix}$. A second-order vector autoregression time series model, VAR(2), is then given by

$$\mathbf{z}_t = \Pi_1 \mathbf{z}_{t-1} + \Pi_2 \mathbf{z}_{t-2} + \in_t \tag{4}$$

where $\in_t$ is vector of random-error terms which are assumed to be independent and normally distributed with zero mean vector and variance matrix Σ (Johansen, 1991), Π_1 and Π_2 are 2×2 matrices of unknown coefficients. For vector of integrated variables $\mathbf{z}_t$, estimates of these parameters may not have their asymptotic properties. Vector error-correction form (discussed in Section 6.1) of Equation (4) makes the series stationary and keeps the valuable long-run information by including a lagged error-correction term in the model (Granger, 1981; Engle and Granger, 1987).

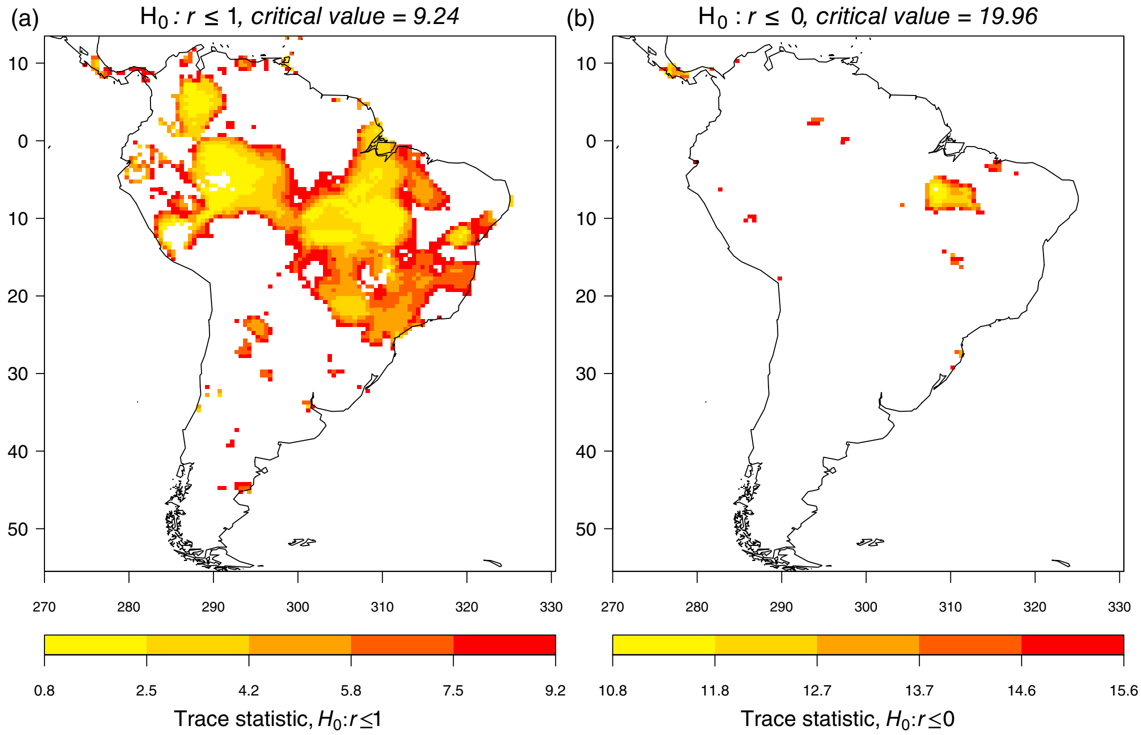

Figure 6. Trace test for rank of cointegration between the MAM and NDJ temperatures: Map of significant test statistics at 5% level of significance. Panel (a) shows map of significant test statistics for the hypothesis of one cointegrating stationary relationship between the two temperatures and panel (b) shows those for the hypothesis of no cointegrating stationary relationship between them. For the areas in white one of the following holds: the two temperatures are not jointly unit root or the trace test statistics are not significant at 5% or missing numerical data.

6.1. The vector error-correction model

The unrestricted VAR model in Equation (4) can be reparameterized in terms of differences, lagged differences, and levels of the process – the vector error-correction model (VECM). The VECM form of VAR(2) is obtained by subtracting $\mathbf{z}_{t-1}$ from both sides of Equation (4) and by adding and subtracting $\Pi_2\mathbf{z}_{t-1}$ to and from the right-hand side of the same Equation. We then obtain:

$$\Delta\mathbf{z}_t = \Pi\mathbf{z}_{t-1} + \Gamma\Delta\mathbf{z}_{t-1} + \in_t \tag{5}$$

where $\Pi = (\Pi_1 + \Pi_2 - \mathbf{I}_{2\times 2})$, $\Gamma = -\Pi_2$ and other notations and assumptions remain the same as in Equation (4). The Π matrix describes the *long-run* relationship between variables (T_{MAM_t} and $T_{\mathrm{NDJ}_{t-1}}$) and Γ describes *transitory* effects measured by the lagged changes of the variables. The VECM form makes tests for cointegration simpler in that the left-hand and the right-hand sides have different orders of integration which cannot be balanced unless there are cointegrating relations in levels on the right-hand side of Equation (5).

For Equation (5) to hold at locations where the two temperature series (components of $\mathbf{z}_t$) are $I(1)$, the long-run matrix Π should be of a reduced rank, say r. For a bivariate model, if $r=2$, then Π is of full rank and invertible. In this case, $\mathbf{z}_t$ itself is stationary and standard inference applies. But for the system of integrated processes $\mathbf{z}_t$, the Π matrix should be non-invertible and thus, this case does not hold. If $r=0$, then there are 2 unit roots in $\mathbf{z}_t$ and it is not possible to obtain stationary cointegration relations between the levels of T_{MAM_t} and $T_{\mathrm{NDJ}_{t-1}}$. In this case, the two $I(1)$ variables do not have a common stochastic trend and, hence, do not have similar temporal behaviour over time (Juselius, 2006). If $0<r<2$, then there is $r=1$ stationary linear combination of the two $I(1)$ series, which is the cointegrating relation. In this case, the long-run coefficient matrix can be given as $\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}'$, where α (vector of adjustment coefficients) and β (cointegration vector) are both 2×1 vectors. See Turasie (2012, page 56–59) for details on the likelihood estimation of the cointegration vector for a bivariate VAR(2) model.

Statistical test for the rank of Π is the same as a test for existence of cointegration. Because VECM exists if and only if Π has less than full rank but is not equal to zero, and by the *Granger Representation Theorem* (Engle and Granger, 1987) VECM implies cointegration and vice versa.

6.2. Test for cointegration rank

The number of linearly independent cointegration vectors is referred to as cointegrating rank (Johansen, 2000). A formal test for the cointegration hypothesis can be formulated as a reduced rank test on the Π matrix, H_0: $r \leq q$, for some constant number q ($q=0, 1$, for bivariate model), and the alternative is $r>q$. The likelihood ratio-based *trace* statistic introduced by Johansen (1988) is used to test rank of Π matrix in this study. For our bivariate case, the null hypothesis of *at most one cointegration vector* is rejected in favour of a *more than one* alternative if the estimated trace statistic is greater than critical value provided, for instance in Table 1* of Osterwald-Lenum (1992).

Figure 6(a) shows that at the most one cointegrating relation hypothesis is not rejected in almost all locations where both T_{MAM_t} and $T_{\mathrm{NDJ}_{t-1}}$ are unit root, whereas Figure 6(b) shows that the zero cointegration hypothesis is rejected at almost all (except few grid points) locations where the two temperatures are unit root. These together imply that there is one stationary long-run relationship between the two non-stationary seasonal temperatures.

The hypothesis of one cointegrating relation between previous NDJ and next MAM temperatures is not rejected at most parts of the continent. The critical value of this test is 9.24. Then, one can use this estimated long-run linear relationship to forecast average MAM temperature using average NDJ temperature as predictor.

6.3. Cross-validated cointegration forecasts

For locations where the Π matrix does not have full rank, Equation (5) can be rewritten as

$$\begin{aligned}\Delta \mathbf{z}_t &= \alpha\beta' \mathbf{z}_{t-1} + \Gamma \Delta \mathbf{z}_{t-1} + \in_t \\ &= \alpha\left(T_{MAM_{t-1}} - \beta_1 T_{NDJ_{t-2}}\right) + \Gamma \Delta \mathbf{z}_{t-1} + \in_t \qquad (6)\end{aligned}$$

and estimated leaving 1 year out at a time using Johansen's maximum likelihood estimation (normalized to T_{MAM}). We estimated Equation (6) including an intercept term in the long-term part of the model so as to facilitate comparison with the lagged regression model (discussed earlier) with intercept. After establishing linear relationship between the two temperatures, forecast variances are computed in a similar way to that of the lagged regression forecast. That is, since the residuals from estimating model (6) account for the whole error-correction model including the long-term (cointegrating) and short-term components, we use the estimated long-term relationship between the two temperatures and compute residuals of the estimated cointegrating relationship by taking difference between the fitted and observed values of the predictand in a similar way to that discussed in Section 3.2.

Forecasts of T_{MAM} using previous year's T_{NDJ} are then made using the estimated long-term relationship between the two temperatures. Correlation skills of the resulting retrospective forecasts (for only locations where the two temperatures are unit root and have one stationary linear relationship between themselves) are shown in Figure 7.

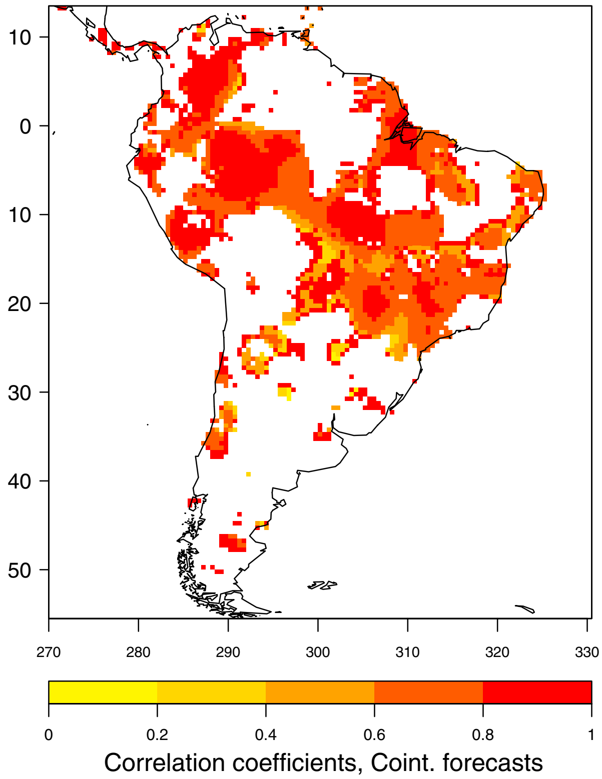

Figure 7. Map of correlation skill for the 1949–2012 period. Cointegration forecasts of MAM temperature using previous NDJ temperature as predictor. For the areas in white one of the following holds: the predictor and predictand temperatures are not jointly unit root or the two temperatures have no cointegrating relationship or the correlation coefficients between the forecast and observations are not significantly different from zero at 5% level or missing numerical data.

7. Forecast skill comparison

The cointegration method is designed to take account of non-stationarity in studying long-run relationship between non-stationary variables and hence it is proper (recommended) for forecasting climatic variables presenting such behaviour. Once existence of stationary long-run relationship between the predictand and predictor variables is identified, the OLS regression and correlation skill results are also acceptable (not spurious). But still there may exist differences in forecast performance of these approaches. Therefore, one needs to compare the forecast performance of the three methodological approaches considered in this study. To make a sensible and consistent performance comparison among these forecasts, we only consider those grid points where the predictand and predictor temperatures are unit root and have a cointegrating long-run relation.

7.1. Differences in correlation skill

In this section, we show pairwise comparisons of the correlation skill of cointegration forecasts against those of the OLS regression and persistence forecasts. First, the correlation skill of each method is computed using the respective forecasts and the observed series. Then differences between correlation skill of cointegration forecasts and each of the other two are compared.

The significance of the difference between correlation skill is tested using Pearson and Filon's (Pearson and Filon, 1898) z statistic, which is given as:

$$z = \frac{\sqrt{n}\left(r_{oc} - r_{or}\right)}{\sqrt{\left(1 - r_{oc}^2\right)^2 + \left(1 - r_{or}^2\right)^2 - 2k}} \qquad (7)$$

where $k = r_{\mathrm{cr}}\left(1 - r_{\mathrm{oc}}^2 - r_{\mathrm{or}}^2\right) - \frac{1}{2}\left(r_{\mathrm{oc}} r_{\mathrm{or}}\right)\left(1 - r_{\mathrm{oc}}^2 - r_{\mathrm{or}}^2 - r_{\mathrm{cr}}^2\right)$ (as explained in Steiger, 1980), and r_{oc}, r_{or}, and r_{cr} are the correlation coefficients between observations and

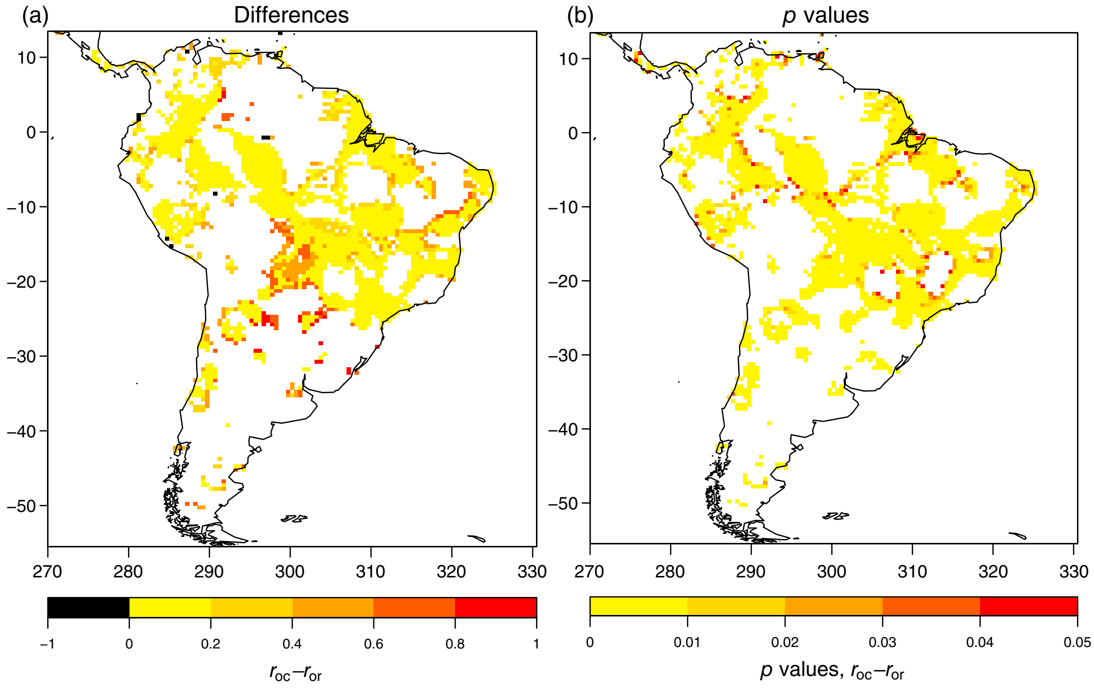

Figure 8. Differences between correlation skill of cointegration and the OLS regression forecasts using previous NDJ temperatures as predictor: Map of (a) the differences between correlation coefficients (cointegration−OLS regression) and (b) the respective p values of the test. For the areas in white one of the following holds: the predictor and predictand temperatures are not jointly unit root or the two temperatures have no cointegrating relationship or the difference between correlation skill of the two methods are not significantly different from zero at 5% level or missing numerical data were identified.

cointegration forecasts, between observations and regression forecasts and between the two forecasts, respectively. This statistic has an asymptotic standard normal distribution and provides a large-sample statistic for testing the equality of two correlations with one series, the observed temperature, in common (Steiger, 1980). Replacing r_{or} and r_{cr} by r_{op} and r_{cp} respectively, we test significance of differences between correlation skill of cointegration and persistence forecasts.

The positive difference between the two correlations, $r_{oc} - r_{or}$ and $r_{oc} - r_{op}$, shown in Figures 8(a) and 9(a), at most of the locations under consideration indicates that the cointegration forecasts of MAM temperatures, using previous NDJ temperature as predictor, has better correlation skill than its regression and persistence counterpart. Figures 8 and 9 show these differences and the respective p values of the test for locations where there are significant differences between pairs of correlations. We reject the null hypothesis of no significant difference in favour of a two-sided alternative for smaller p values.

7.2. Brier scores and reliability diagrams

Assuming Gaussian distribution for seasonal average temperatures, the probability forecast values for the event at each grid point are estimated by using forecast density $p(x) = \frac{1}{\sqrt{2\pi}s} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{s}\right)^2}$, where $\bar{x}$ is mean forecast for a given grid point and forecast period (season) and s^2 is forecast variance of the corresponding location and time.

One condition for the validity of probabilistic forecasts for the occurrence of an event is statistical consistency between a priori predicted probabilities and a posteriori observed frequencies of the occurrence of the event under consideration (Jolliffe and Stephenson, 2003). This consistency is well visualized using the reliability diagram.

Figure 10 shows reliability diagrams of forecasts of upper tercile MAM temperature using the three methods considered in this study, for grid points where the two temperatures are unit root and cointegrated. It compares observed relative frequencies with forecast probabilities. Reliable (well-calibrated) forecasts should show a curve close to the diagonal (45 ^{o}) line. Here, all the three investigated forecasting methods are close to the diagonal. However, the cointegration forecast (panel 10c) is the closest to the diagonal of all examined forecasts. This also suggests that forecasts produced with cointegration slightly better agree with the observed frequencies than the other two forecasts.

8. Summary

Different empirical seasonal forecast methods and their potential skill are investigated in this study. Simple persistence and lagged regression methods, which are more familiar in the seasonal forecast literature, were compared to the cointegration method for locations where seasonal temperatures were found to be non-stationary. Before producing and comparing empirical forecasts using these methods, temporal behaviours of seasonal

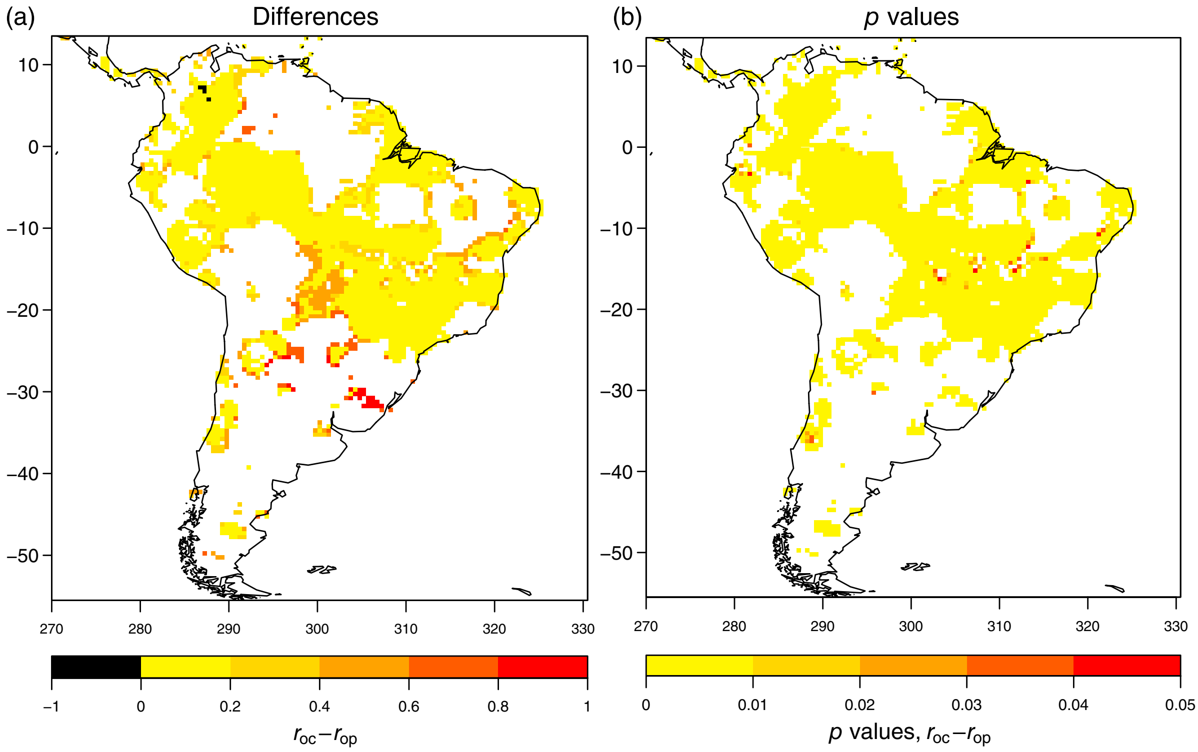

Figure 9. Differences between correlation skill of cointegration and persistence forecasts using previous NDJ temperatures as predictor: Map of (a) the differences between correlation coefficients (cointegration−persistence) and (b) the respective p values of the test. For the areas in white one of the following holds: the predictor and predictand temperatures are not jointly unit root or the two temperatures have no cointegrating relationship or the difference between correlation skill of the two methods are not significantly different from zero at 5% level or missing numerical data were identified.

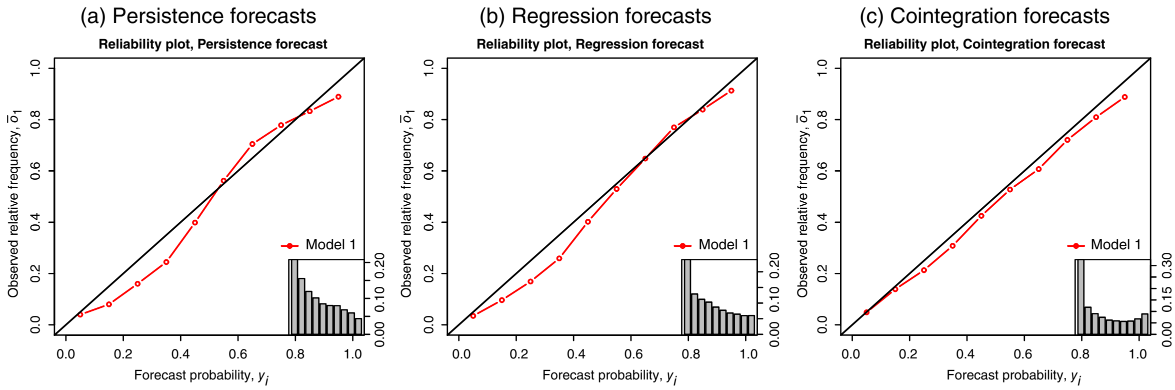

Figure 10. Reliability curves for persistence, lagged regression and cointegration forecasts of the upper tercile MAM temperature using previous NDJ temperature as predictor.

temperatures for the South American region were studied. The results have shown that these temperatures are characterized by trends, dominated by stochastic trends, through time at most parts of the continent.

We have shown the impact of trends on OLS regression and correlation outputs using independently simulated synthetic data. As to these results and literature cited in this article, existence of stochastic trends may affect regression/correlation outputs which in turn might result in misleading interpretation. This article, therefore, takes this concern into account and (1) studied existence of true long-run relationship (cointegration) between the predictor and predictand temperatures before inferring on correlation skill of different empirical methods and (2) compares the potential skill of a newly introduced cointegration model to those of the widely known persistence and lagged regression models for regions where the two temperature time series (predictor and predictand) are unit root (i.e. both have stochastic trends).

This study has shown that the previous and next season temperatures have a stationary linear relationship in the long run at almost all locations where both temperatures are unit root. Cointegration forecasts have shown to perform better than the other two forecasts (persistence and lagged regression) for those locations where the predictor and predictand temperatures are cointegrated. This has been demonstrated by differences between correlation skill and reliability diagrams for the three investigated forecast methods. The cointegration method has been shown to perform well for locations where the predictor and predictand climate variables are non-stationary. The robustness of these results was checked by performing similar analysis using previous January temperature as predictor to next MAM temperature forecast (results not shown here). However, the overall comparison indicated that previous NDJ has a better explanatory power over previous January in predicting next MAM temperatures.

One of the main reasons for the better performance of cointegration method is that the modelling procedure accounts for the existing non-stationarity in the data over time. On the other hand, both the regression and persistence approaches assume stationarity. Application of these later methods to non-stationary data may result in less accurate model estimation. Studies have shown that, for example, the OLS regression (which assumes stationarity over time) estimates are negatively biased for non-stationary process (Turasie, 2012). Thus, forecasts based on such inefficiently estimated models are less reliable as seen in this study.

Under the current global warming scenario, most of the climatic variables including temperature in particular cannot be assumed to be stationary through time. Thus, under such situation, assuming stationarity and employing models that cannot handle the inherent non-stationarity might end up in misleading results. The cointegration approach used in this study accounts for non-stationarity in the data and estimates model parameters efficiently so that the model can predict the non-stationary system out of the range as accurately as possible. Therefore, the used cointegration method appears to be an ideal tool for modelling and predicting climatic variables under the global warming scenario.

Acknowledgements

This study was partly supported by the SPECS (Seasonal-to-decadal climate Prediction for the improvement of European Climate Services) project (grant agreement No. 308378) funded by the Seventh Framework Programme (FP7) of the European Commission. This project aims to deliver a new generation of climate prediction systems for seasonal-to-decadal time scales. We would like to thank the SPECS project for all support provided during the development stage of this study. AAT was supported by Science Faculty, University of Wits, and would like to thank them. CASC was supported by Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq) process 306863/2013-8.

References

Barnston AG, He Y. 1996. Skill of canonical correlation analysis forecasts of 3-month mean surface climate in Hawaii and Alaska. *J. Clim.* **9**(10): 2579–2605.

Barnston AG, Smith TM. 1996. Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Clim.* **9**(11): 2660–2697.

Coelho CAS, Pezzulli S, Balmaseda M, Doblas-Reyes FJ, Stephenson DB. 2004. Forecast calibration and combination: a simple Bayesian approach for ENSO. *J. Clim.* **17**(7): 1504–1516.

Coelho CAS, Stephenson DB, Balmaseda M, Doblas-Reyes FJ, van Oldenborgh GJ. 2006. Toward an integrated seasonal forecasting system for South America. *J. Clim.* **19**(15): 3704–3721.

Colman AW, Davey MK. 2003. Statistical prediction of global sea-surface temperature anomalies. *Int. J. Climatol.* **23**(14): 1677–1697.

Dickey DA, Fuller WA. 1979. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **74**(366a): 427–431.

Dickey DA, Fuller WA. 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* **49**(4): 1057–1072.

Engle RF, Granger CWJ. 1987. Co-integration and error correction: representation, estimation and testing. *Econometrica* **55**(2): 251–276.

Fan Y, Van den Dool H. 2008. Global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.* **113**: D1.

Granger CWJ. 1981. Some properties of time series data and their use in econometric model specification. *J. Econ.* **16**(1): 121–130.

Granger CWJ, Newbold P. 1974. Spurious regressions in econometrics. *J. Econ.* **2**(2): 111–120.

Hamilton JD. 1994. *Time Series Analysis*. Princeton university press: Princeton, NJ.

Hendry DF, Juselius K. 2001. Explaining cointegration analysis: Part II. *Energy J.* **22**(1): 75–120.

Jan van Oldenborgh G, Balmaseda MA, Ferranti L, Stockdale TN, Anderson DLT. 2005. Evaluation of atmospheric fields from the ECMWF seasonal forecasts over a 15-year period. *J. Clim.* **18**(16): 3250–3269.

Johansen S. 1988. Statistical analysis of cointegration vectors. *J. Econ. Dyn. Control.* **12**(2): 231–254.

Johansen S. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**(6): 1551–1580.

Johansen S. 2000. Modelling of cointegration in the vector autoregressive model. *Econ. Model.* **17**(3): 359–373.

Jolliffe IT, Stephenson DB. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons.

Juselius K. 2006. *The Cointegrated VAR Model: Methodology and Applications*. Oxford University Press.

Landman WA, Mason SJ. 2001. Forecasts of near-global sea surface temperatures using canonical correlation analysis. *J. Clim.* **14**(18): 3819–3833.

Osterwald-Lenum M. 1992. A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics 1. *Oxf. Bull. Econ. Stat.* **54**(3): 461–472.

Pearson K, Filon LNG. 1898. Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philos Trans R Soc Lond A* **191**: 229–311.

Pfaff B. 2008. *Analysis of Integrated and Cointegrated Time Series with R*. Springer.

Phillips PCB. 1986. Understanding spurious regressions in econometrics. *J. Econ.* **33**(3): 311–340.

Steiger JH. 1980. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **87**(2): 245–251, doi: 10.1037/00332909.87.2.245.

Turasie AA (2012). *Cointegration Modelling of Climatic Time Series*. PhD thesis. University of Exeter, UK.

Van den Dool H. 2006. *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press.

Yule GU. 1926. Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *J. R. Stat. Soc.* **89**(1): 1–63.