Chapter

Severe Testing and Characterization of Change Points in Climate Time Series

James Ricketts and Roger Jones

Abstract

This paper applies misspecification (M-S) testing to the detection of abrupt changes in climate regimes as part of undertaking severe testing of climate shifts versus trends. Severe testing, proposed by Mayo and Spanos, provides severity criteria for evaluating statistical inference using probative criteria, requiring tests that would find any flaws present. Applying M-S testing increases the severity of hypothesis testing. We utilize a systematic approach, based on well-founded principles that combines the development of probative criteria with error statistical testing. Given the widespread acceptance of trend-like change in climate, especially temperature, tests that produce counter-examples need proper specification. Reasoning about abrupt shifts embedded within a complex times series requires detection methods sensitive to level changes, accurate in timing, and tolerant of simultaneous changes of trend, variance, autocorrelation, and red-drift, given that many of these measures may shift together. Our preference is to analyse the raw data to avoid pre-emptive assumptions and test the results for robustness. We use a simple detection method, based on the Maronna-Yohai (MY) test, then re-assess nominated shift-points using tests with varied null hypotheses guided by M-S testing. Doing so sharpens conclusions while avoiding an over-reliance on data manipulation, which carries its own assumptions.

Keywords: severe testing, misspecification testing, abrupt shifts, unit-roots, change-points

1. Introduction

Anybody examining sudden changes in data needs to ask, "Does this mean what I think it means? Are there other explanations?" Further, the evidence needed to overturn the acceptance of a generally held position requires high probative value, supporting the proposed position and addressing the accepted one; and should be convincing to the investigator and others.

This paper addresses this issue by presenting a systematic approach that combines the development of probative criteria with error statistical testing, illustrating it with a specific investigation of climate. The approach is developed from previous work on the philosophy of statistics, which is relatively new to climate work [1–6].

Climate, like many areas of natural science, depends heavily on statistical induction for the interpretation of physically-based behavior. Many popular

statistical tools are generalized tests, framed against broad statistical assumptions that may be challenged by complex physical processes. The implicit assumptions of tests must be considered, as must the linkage between those processes, the accessible data, and statistical models. Where competing alternatives cannot be correctly distinguished by the tests and specific data chosen for that purpose, the data is misspecified with respect to the statistical models or the model selection processes.

The particular aspect addressed here is model specification with respect to the data. Probative criteria drawing from theory and interpretations of physical behavior cannot be applied correctly, if the tests to not adequately represent those criteria, or distinguish between them.

1.1 Illustrative example: abrupt shifts in climate signals

A number of publications now address an area of some controversy – the hypothesis that under greenhouse gas-induced radiative forcing, climate changes in a step-like manner [7–12]. The controversy arises because it is almost universally accepted that the forced response of climate change, especially global mean surface temperature (GMST), responds rapidly to forcing and hence is trend-like; albeit embedded in a very complex "error" process which yields highly structured residuals.

Our paper from 2017 (JR2017) [12] and the PhD thesis of Ricketts (R2019) [13], in addressing this controversy, required the development of automated, reliable and unbiased detection of shifts, and importantly various means of ensuring that presumptive shifts were not artefacts of the detection method and the structured residuals.

We built on the concepts of severe testing [3] and misspecification testing [2], and we adapted a framework of models to connect theory and data [14, 15]. Thus we could severely test two propositions: (*H1*) forced warming and natural variability proceed gradually and independently, with the response to forced warming best represented as trend-like; and (*H2*) forced warming and natural variability interact so that patterns of response may project onto modes of climate variability – either one-way as proposed by Corti et al. [16] or two-way as proposed by Branstator [17] – in either case giving rise to abrupt state-like transitions in the signal.

JR2017 showed that *H2* was preferred to *H1* in all of six tests of a severe testing regime; R2019 also showed that abrupt shifts relate directly to warming; in their extent, frequency and intensity; and more so at finer scale.

1.2 Structure of the rest of the paper

Firstly we very briefly introduce severe testing.

Then we introduce our version of a framework that connects hypotheses about the physical world to statistically based tests that license inductions about models of the world.

Next we spend more time on misspecification testing (M-S), which was proposed as an approach to determining whether the assumptions needed to reliably model the statistical variables are met [2].

2. Severe testing

Severe testing, proposed by Mayo and Spanos, is based on the intuition that "Data x0 in test T provide good evidence for inferring H (just) to the extent that H passes severely with x0, i.e., to the extent that H would (very probably) not have

survived the test so well were H false." [3]. They propose that a severity criterion supplies a meta-statistical principle for evaluating statistical inferences (their page 328), where the severity of testing is not assigned to hypothesis H, but to the testing procedure.

In the preface of [6] we read the following, "If little or nothing has been done to rule out flaws in inferring a claim, then it has not passed a severe test. In the severe testing view, probability arises in scientific contexts to assess and control how capable methods are at uncovering and avoiding erroneous interpretations of data. … A claim is severely tested to the extent that it has been subjected to and passes a test that probably would have found flaws, were they present."

Severe testing is beginning to be picked up by the climate community (e.g. [18–20]). It was applied to an analysis of optimal fingerprint methods in climatology [18] and to address issues of model tuning in climate projections [20]. Severe testing forms a core methodology of JR2017, R2019, and a conference paper [21] (RJ2017).

3. The theoretical mechanistic/statistical inductive (TM/SI) framework

The TM/SI framework borrows from a strong body of earlier work (e.g. [4, 14, 15, 22, 23]) and was outlined in Section 2 of JR2017 to provide support for reasoning about climate where the scientific debate had been muddied by competing claims from outside the science community.

The approach follows Haig [15], and employs the concept of severe testing [3], and in keeping with it, error-statistical methods [4, 23]. It requires a carefully reasoned matching between scientific hypotheses about the physical world, with statistical hypotheses about the observed data.

The theoretical-mechanistic part consists of the physical aspects, components, relationships and measurable quantities.

The statistical-inductive part consists of the process of drawing conclusions about specified hypotheses concerning the system given the physical model, real-world data and statistical tests.

The goal is to construct a chain of reasoning that ties physical hypotheses, *H1 .. Hn*, to statistical hypotheses *h1 .. hn*. That is, features of the world map to defined outcomes of statistical tests (preferably one to one). One to one mapping meets a requirement of severe testing. Misspecification testing assists this mapping.

3.1 The TM/SI structure

Suppes [14] suggested that science employs a values hierarchy of models that ranges from experimental experience to theory, claiming that theoretical models, high on the hierarchy, are not compared directly with empirical data, which are low on the hierarchy. Rather, he said, they are compared with models of the data, which are higher than data on the hierarchy. Following on, Haig describes an egalitarian framework in which three different types of models are interconnected and serve to structure error-statistical inquiry [15].

He describes: Primary models which break a research question into a set of local hypotheses; Experimental models which "structure the particular models", and link Primary models to Data models; which in turn generate and model raw data, and check that the data satisfies the assumptions of experimental models. Although Haig does not fully explain experimental models which "structure the particular models" it seems implicit that they map hypotheses to model components and processes. He leaves to his data models the role of checking that data meets the assumptions of experimental models.

To summarize, the TM/SI was constructed with physically grounded work in mind, and adapts Haig's approach. Physical entities and their relationships about which we propose hypotheses guided by Physical models are the Primary models. These link to the Statistical models which support reasoning with an inductive framework, via Data models which includes Sampling procedures. Sampling procedures guide the accumulation of data on which we reason. All data sampling procedures and statistical tests are framed against *ruling assumptions*. Violation of the ruling assumptions weakens statistical inference.

Physical Model: Concerns the system of physical entities and their interactions. Entities have measurable properties, which are accessed through Sampling procedures.

Statistical model. A mapping between a sampled set of observations and a set of parameterized probability distributions. This is informed by an error model – the theoretical behavior and characteristic distribution of sampling error, generally assumed to be random. If properly specified, the statistical model(s) license(s) valid statistical inductions about hypotheses, generally via statistical model selection from a specified statistical family.

Sampling procedures: Data models which cover the collection of measurements. Measurements are made, and treated (e.g. homogenized), and output to become sample data input to statistical models. The choice of sampling model (random sampling, averaging) influences subsequent induction since sampling error subsumes both random processes and statistical misspecification.

Severe Testing requires that these issues be accounted for so that to the extent possible, when features are present they are detected, and when they are not present they are not erroneously identified.

3.2 Applying the TM/SI to climate

3.2.1 Physical model: surface temperatures

To guide investigation we propose in JR2017 (a) physical model *M1* – a world in which average surface temperatures closely track forced warming, and natural variability is independent, and reflective of the indices of variability and by contrast (b) a physical model *M2* – which mirrors *M1* but in which there is interaction between forcing and natural variability. The *M2* world requires that Earth's surface temperature is additionally reflective of, and tracks, internal physical states of variability modes which may change abruptly, thus imprinting step-like shifts into the temperature records. These shifts mark state changes in the climate system, and represent the major response at decadal time scales of the climate system to the gradually increasing greenhouse forcing. Earth's surface temperature is sampled, but it is understood that this also reflects the overall state of heat transport in the fluid layers.

3.2.2 Sampling considerations

Observed climate data is derived over time using evolving and fallible instrumentation. This dictates the use of a wide variety of strategies to enable intercomparisons. In our analyses we are concerned with annual or monthly averages which, in the case of gridded data, have been further averaged and re-interpolated spatially. We must consider the effects of these procedure.

Averaging implicitly assumes a signal/noise model where mean noise converges on zero at all time points to enhance the signal which is assumed to be represented

equally in all samples. An influence travelling in space and time when averaged will appear as some form of non-stationarity in the time series of the mean.

Temperature records increase in spatial density over time, they are records of opportunity. Conditions over land and ocean differ. To enable inter-comparison with models they are re-interpolated onto regular grids. They are also homogenized prior to gridding to deal with instrumentation changes [24].

Both *M1* and *M2* worlds have time varying temperature records but because forced change in M1 propagates rapidly, averaging does not induce troublesome artefacts. This is not the case for *M2*. A step-like change occurring serially across regions may give rise to trends and auto-regression, and or may obscures more regional signals.

3.2.3 Statistical model(s)

Different statistical models are involved in detection of changes, and in the assessment of the relative merits of *M1* and *M2*. It is important that the probabilities from statistical *feature detection* not be also used for *model selection*.

In break-point analysis the family of segmented linear regression models is used. The choice of specific parameters from within a specified family is termed model selection, and would in our work include the serial selection of specific change-points. The MSBV differs from other approaches in that it does not terminate the search for change-points (feature detection) on the basis of an all of model information criterion such AIC (a model selection criterion), but usually earlier, when no segment can be sub-divided.

In our work, detection of such steps, supported by evidence that they are not artefacts provided by M-S testing supports constitutes support for *M2*, and thus support for *H2*.

4. Misspecification testing

Mayo and Spanos differentiate between model specification and model selection. An adequate model specification licenses primary statistical inference, and with it statistical model selection from the specified family. Serial feature detection in any time series is a form of model selection from a family of related models, reliant on model specification. It must be noted that for our work a series of tests are performed, a single detection test and multiple probative tests, but that as each is against an independent null, this does not involve a multiple-testing issue, instead increasing the overall power of the testing regime.

Chapter 2 of [25] defines experimental error as all extraneous variation outside experimental treatments, and states "Neither the presence of experimental errors or their causes need concern the investigator, provided his [sic] results are sufficiently accurate to permit definite conclusions to be reached". This definition still dominates statistical climatology. Climate data are not generally experimental, but often a feature of interest in climate data is investigated by treating natural variability as extraneous variation. Experimental design requires that statistical models are properly specified, however complex systems being observed may align to many different statistical models and have multiple features of interest, leading to the possibility of misspecification.

Mayo and Spanos [2] (MS2004) introduce a methodology for testing misspecifications in statistical models (M-S testing). Taking this as a point of departure we then propose that a full understanding of the assumptions of statistical models allows one to probe data for features even when available tests are misspecified.

Model specification delineates families of statistical models. For physical problems, the family would be misspecified if the available parameters do not properly reflect the physical processes [13].

In MS2004 the authors use an example of a linear regression model to address a problem of validation in regression models. Three general forms of M-S are recognized:

- Functional form misspecification in which a statistical model includes the correct parameters or variables but inside an incorrect function. For example, as x^2 instead of x^3 or sin(x).
- Missing parameter misspecification in which a parameter/variable is omitted.
- Irrelevant parameter misspecification in which unnecessary parameters are introduced.

4.1 Summary of MS2014

MS2004 says, "A full methodology of M-S testing, as we see it, would tell us how to specify and validate statistical models, and how to proceed when statistical assumptions are violated."

Statistical model specification (goals and assumptions) is different from statistical model selection (from an assumed family of models, with a heuristic (e.g. AIC))). We consider a statistical model M, selected from a family of models.

MS2004 considers firstly the primary questions of statistical inference, whether the assumptions needed to reliably model the data are met; and secondary questions, whether there are influential gaps between variables in a statistical model and primary questions. Primary questions are addressed within the selected statistical model M. Formally, the hypothesis H_{M}:

$$H_{\mathrm{M}} : \text{data } \mathbf{z} \text{ supports the probabilistic assumptions of statistical model } M. \quad (1)$$

Secondary questions are essentially meta-questions conducted outside the model M, they address the suitability of the test, given the data and require auxiliary models and put M's assumptions to the test. Formally this would test

$$\begin{aligned} &H_0 : \text{the assumption(s) of statistical model } M \text{ hold for data } \mathbf{z}, \\ &\text{against all possible assumptions by which } H_0 \text{ could fail } (H_{1.} \ldots H_{\mathrm{n}}). \end{aligned} \quad (2)$$

It is critically important to recognise that this use of multiple tests does not constitute multiple tests of HM. It augments and increases (not diminishes) the confidence in the conclusions.

They present a case study of an empirical relationship between the USA population (y_i) and a secret variable (x_i). They commence with a proposed explanatory model, with $R^2 = 0.995$ and p-value nearly zero. Here, $\hat{u}_i$ represents the estimated error process.

$$M0 : y_i = 167.115 + 1.907x_i + \hat{u}_i \quad (3)$$

Concerning H_{M}. An assumption of the regression is that errors $\hat{u}_i$ are normally distributed, independent and identically distributed (NIID). Is this met? A runs test suggests not, and a parametric Durbin-Watson test suggests autocorrelation.

$$M1 : y_i = \beta_0 + \beta_1 x_i + u_i, u_i = \rho u_{i-1} + \varepsilon_i. \tag{4}$$

However an alternative AR(2) model is then shown to explain more variance *without* the $\beta_1 x_i$ term.

$$M2 : y_i = \beta_0 + \beta_1 y_{i-1} + \beta_2 y_{i-2} + \hat{u}_i \tag{5}$$

Probing the model M0 shows it to be misspecified due to an irrelevant variable. The secret variable x_i is the number of shoes owned by Spanos's grandmother!

4.2 Application to climate data

4.2.1 Abrupt changes in previous literature

In some papers step-like changes are introduced *en passant*, on the way to revealing or locating in time various phenomena. For instance the delineation of the Pacific Decadal Oscillation [26–28], or reduction in South-Western Western Australian rainfall [29]. In the last decade an astonishing number of papers addressed the so-called hiatus, many purporting to show that it never happened [30] or was simply routine variability [31, 32], or a methodological/statistical error [33], or suggesting that natural variability, internal variability and extrinsic factors combined with forced warming [34]. However others, one way or another, simply incorporate it as fact [35, 36].

From these and other papers and some personal communication, the objections/challenges to the existence of abrupt changes (including but not limited to the so-called hiatus) appear to be

1. Physical implausibility of step like changes in average temperatures.

2. Overcooking. In general, that warming is in fact more or less constant and positive, and more or less smoothly changing natural variability is imposed on it, with the result that a test for shifts is deceived by increases and decreases in the derivative of the sum.

3. Overcooking worsened by autocorrelation. As above but with at least some natural components following an autocorrelation model.

4. Model misspecification by virtue [sic] of step methods applied to trending data.

5. Non-determinism. Red noise/unit root processes masquerading as natural variability and/or as one off deterministic events. Non-determinism implies that detected events cannot be attributed to a deterministic physical model.

6. Presence of one or more sub-detection threshold deterministic events. This is a particularly nasty issue because (a) it affects detection of many phenomena, (b) it may deceive autocorrelation tests and unit-root tests as well as trend tests.

7. Conflict with objectors favoured model/approach.

Not all of these concern statistical M-S. Objection 1, physically implausibility of discontinuities in surface temperatures [37] can only result from an underlying

assumption of a physical model where heat is dispersed rapidly and uniformly. Objection 7 is regrettable but not uncommon and not further considered.

4.2.2 Approach

Bearing in mind *M1* and *M2*, an important step is determining precisely what information is of primary importance. What variables are of interest and what features are important?

Step 1: As argued in JR2017, step-like shifts in temperatures are a feature of the abrupt state transitions of *M2* rather than the smooth transitions of *M1*.

Step 2: Matching alternative hypotheses, including those represented by the objections, to appropriate statistical tests. Consideration of the implicit choices made.

1. Physical implausibility is not considered further here.

2. Overcooking alone and ...

3. Overcooking with autocorrelation. The challenge here is to the meaning of abrupt shifts. If detection tests are finding the point of maximum (or minimum) derivatives of quasi-sinusoidal variability then the residuals of a segmented model will be heteroskedastic whether or not autocorrelation is present.

4. Step-change methods applied to (and deceived by) constantly trending data. In general segmenting such a process will yield segments which testing against a step and trend model will reveal to be co-linear.

5. Non-determinism relates to the interpretability of change-points and their relationship to any physical model.

6. Undetected deterministic events (events below detection thresholds or misspecification). The primary issue with these is that the error series is not random and tests assuming such are ill founded. This includes autocorrelation and trend tests.

4.3 Types of tests

In what follows, three classes of testing useful during analysis of step-like change-points have been identified. The first two involve testing the segment of data within which a change-point is found. The third asks if a multiple change-point model is adequate.

1. Does the requirements of the detection test (specifically that it should be sensitive to small shifts while precise in the timing) open the door to deception? Do individual change-points remain if more parameters are allowed?

2. Are changes reasonably regarded as deterministic or is there evidence of non-determinism which would support objection 5?

3. Does the full set of change-points explain the necessary degree of variation? Are the residuals homoskedastic?

4.3.1 Detection test

Complex climate time-series data is almost certainly misspecified for *any* change-point detection test – thus the goal is adequate applicability to questions of interest. In testing for multiple change-points, many methods, including the MSBV, examine data only between presumptive lower and upper bounding points and restart estimation of the distribution parameters. The assumptions of the basic detection methods used must be considered.

Issues potentially arising include false detection, timing errors, and false negatives. Timing error includes misplacement and imprecision. The MSBV incorporates a resampling strategy [38] which reduces imprecision. False positives (deterministic and non-deterministic) can be uncovered by post-detection assessments, but false negatives introduce down-stream non-stationarities that interfere with detection of later change-points. Combining tests with differing assumptions and different nulls probes for both non-deterministic and deterministic causes of false results including sub-detection threshold events.

Analysis of covariance (ANCOVA) is used post-detection of a change-point by MSBV (which does not consider trend) to ensure that the presence of the change-point provides explanatory power in an unconstrained disjoint linear statistical model which allows trend. It does not attempt to locate an alternative change-point – the Zivot-Andrews test however, see below, does this in passing. ANOVA tests for change of trend and change of level are obtained in passing but in R2019, final *p*-values for change-points are obtained from only from ANCOVA.

4.3.2 Tests for heteroskedasticity for segmentation of data with change-points

The full set of change-points in an entire sequence is tested here by the studentized Breusch-Pagan test (hereafter SBP test) for homoskedasticity of the residuals of the disjoint multi-segment model (JR2017 utilised the equivalent White's test [39]). An adequate model explanation of a time series, under the assumption of i.i.d. error, should have a featureless residual. This test has a null of homoskedasticity, rejected in favour of heteroskedasticity at low probabilities.

4.3.3 Tests for stationarity in a segment

Our detection test and the subsequent probability assignments by ANCOVA or ANOVA, and the further misspecification testing all assume serial independence either in the null or contrast hypothesis.

In these tests the segment containing a provisional change-point is tested for features that may deceive tests for shifts and trends. The MY test, ANCOVA, and where used ANOVA tests, have ruling assumptions of serial independence. The MSBV, and other multiple break tests assume some form of censorship between provisional data segments (determination of change-points within provisional bounds includes only the data within the bounds); but tests of the overall model assume homogeneity of error, thus of variance (e.g. the Akaike Information Criterion or AIC). The SBP also assumes this. All of these above tests are formalised as null hypothesis statistical tests (NHST) and as such they each are subject to their own ruling assumptions. The ruling assumptions are incorporated in the interpretation of the tests.

Autocorrelation in climate time-series is variously treated; some propose its estimation and removal [40], some warn against this idea [41]. Some treat it as a short term process and a cause of deception in change-point analyses [42], others have treated it as a persistent signal [43]. In climate signals, autocorrelation often

appears to be time varying. Therefore we apply the MSBV without adjustment for autocorrelation and perform post-detection analysis to determine whether the detection test is likely to have been interfered with. In general, regression based statistical tests assume the absence of deterministic step-like changes and of unit-root, or red-noise progression.

The term unit-root refers to processes with a characteristic equation that has a value of one. If a unit characteristic is moving average, the error is integrated order zero or I(0), if it is auto-regressive, it is integrated order one, I(1). I(0) processes tend to revert to a mean, I(1) processes follow a martingale [44], and is dominated by red-noise. The integration order defines the number of successive differencing operations required to produce a trend-stationary series.

4.3.4 Residuals compared to initial data

In our work, both the raw data, and the residuals after removal of internal steps and trends, are tested. The rationale for testing both derives from the formulation of the tests themselves, since in these tests, the deterministic and non-deterministic components are separately parameterised. The set of tests chosen are from the econometric literature, and each is framed as a null hypothesis significance test (NHST) with its own specific assumptions. Each test poses either H_0 or H_1 as presence of an assumed non-deterministic unit root progression (see Chapter 2) in data, and the alternatives are chosen from a small range of deterministic features. Crucially, each must be interpreted in the light of its own ruling assumptions.

4.3.5 The full process applied to a single time-series

a. The MSBV is applied to delineate provisional change-points. The resulting statistical model would be accepted as the best estimate (i.e. further testing of change-points not warranted) if the time-series of the residuals was *known* to be i.i.d., *and* underlying physical processes were fully deterministic, and fully reflected in the time series. However this should not be simply assumed.

b. The segment containing each provisional change-point is tested to ensure that to a feasible extent, physically plausible types of deception are not present, and that change-points are deterministic, not stochastic quirks.

c. The set of detected change-points is treated as a disjoint segmented model and the residuals examined for evidence of a misfit of model to data.

The program of tests thus sharpens the error-statistical reasoning component of the TM/SI framework.

4.3.6 Deceptive features detectable with unit root tests

The application of deterministic methods such as OLS to non-deterministic data progression such as a random walk is a misspecification; the results may be deceptive with meaningless shifts and/or trends. Unit root (UR) tests probe the data for features that can superficially imitate deterministic structural changes by cumulative random walks, a red or near red progression. It has been shown by Monte Carlo methods that a test for deterministic trends will find deterministic trends in about 85% of realizations that contain only a stochastic (UR) trend [45]. However combinations of UR tests may also be used to detect both stochastic and deterministic non-stationary sequences, due their varied ruling assumptions and constructions.

Because multiple UR tests are performed, and because they each have differing ruling assumptions, the tests are interpreted in terms of evidence for and against stationarity in the underlying processes. In the econometric literature an exogenous change is one imposed upon a model from outside the model. We elected to retain this word where concepts were derived from economic papers as meaning an abrupt and deterministic change in a deterministic time-series.

4.3.7 Unit roots, non-stationarity, and climate

Transient unit root behaviour, if it occurred, could indicate some sort of regime change, temporarily decoupled from normal forcings. If, in addition, measured noise was not persistent this would show I(0) behaviour; or, if it were fully persistent, as I(1) behaviour. In regional signals in which this occurs, the region may also have become coupled to other sub-systems [46]. This could indicate that the underlying physical model is incomplete and that a missing variable misspecification has resulted. On the other hand, persistent unit root behaviour means that a deterministic change-point analysis is suspect.

The Earth system is constrained so that the overall temperature cannot solely follow a pure random walk – at worst it would follow a Brownian bridge (i.e. sequences where the end-points are meaningful and accepted as deterministic but the path is apparently a random walk [47]). However the composition of summary deterministic signals, such as the GMST, involves manipulations that can produce data that existing unit root tests will identify as containing unit roots, and furthermore deceive deterministic tests in much the same way as random walk data. This issue was extensively examined in R2019 and is addressed later.

4.3.8 Detecting unit root presence

Random walk progression may be present in climate data because of transient physical conditions, or because the data is unrelated to the physical processes assumed (M-S due to irrelevant variables). Additionally there may be features in the data that do not correspond to any of a shift, a trend change, or unit root behaviour (M-S due to missing variables), and UR tests are potentially sensitive to this. This source of deception must also be dealt with. Other features may be present in the data but not detected. For instance, a step-like shift well above a detectability threshold may be present together with a number of small, deterministic shifts below detectability, and this latter may be taken to be evidence of stochastic drift by a UR test.

The unit root based tests used here all inherit in one form or another the Dickey Fuller (DF) model [48].

$$Y_t = \mu + \beta t + \rho Y_{t-1} + e_t \tag{6}$$

ρ represents the portion of the signal (Y_{t-1}) carried forward by autocorrelation, β represents the (deterministic) linear trend, μ represents the intercept, and e_t is the i.i.d. error with zero mean and a constant variance σ^2. If $\rho = 0$ this describes a deterministic trend with no autocorrelation, if $0 > \rho < 1$ there is a deterministic trend with a degree of autocorrelation, and if $\rho = 1$, regardless of other parameters it contains a unit root. If all other parameters are zero and ρ equals one, then there is no deterministic trend, no offset, and Y_i is a random walk. This formulation is modified and sometimes rearranged in different ways by the three UR tests used here.

It is important to note that time-series of successive differences of a step-change in an otherwise stationary time series will contain only one out of range difference. Hence the DF model is intrinsically insensitive to deterministic step changes. Another important property of a unit root process is that the variance of the process increases over time, whereas the variance of a stationary process is constant. This gives a second strategy for determining unit root like behaviour – testing for diverging variance. The Kwiatkowski-Phillips-Schmidt-Shin test (KPSS), [49] examines the properties of the variance rather than the fitted parameters, and it primarily focussed on determination of stationarity. As a result it is more sensitive to exogenous changes.

5. Proposed tests and strategies

The unit root methods used are all coded in R and are, (a) a development of the DF test, the Augmented Dickey-Fuller test (ADF), which takes H_0 of a I(1) unit root against an alternative H_1 of a presumption of no unit root (in this implementation trend and multiply lagged autocorrelation is allowed for), (b) two variants of the KPSS, which takes a H_0 of stationarity (or trend-stationarity) rejecting it in favour of an alternative H_1 of a presumption of unit root, and (c) the Zivot-Andrews test (ZA) [50], which takes a H_0 of I(1) unit root behaviour with a possible endogenous drift against an alternative H_1 of trend-stationarity with exogenous structural change. A trend change or a step change would constitute an exogenous structural change.

Use of a combination of UR tests is not new. The combination of ADF and of KPSS testing has been used before in order to add precision to an analysis of monthly inflation expectations (e.g. [51] Appendix B).

The tests are being applied to data within which a single presumptive deterministic, exogenous, step-like changes was detected. No such change is allowed for in the KPSS and ADF tests, the presumption of unit-root in H_0 or H_1 of the above tests is reinterpreted as evidence of non-stationarity. Evidence of unit-root like behaviour is then sought by examination of the residuals after the removal of the deterministic internal trends and shifts detected in the data.

In general, where evidence of a unit-root is detected, it may be due to undetected deterministic features, and hence will be initially treated as evidence of either deterministic non-stationarity or stochastic non-stationarity.

For all of the above tests, the R implementations take published critical values of the test statistic at the 0.01, 0.05, and 0.1 levels. The KPSS implementation interpolates the test statistic against these values to give probabilities between 0.01 and 0.1, the ADF and ZA implementations simply give the critical values and the test statistic.

None of the tests proposed consider unit root presence or absence when possible structural breaks (such as shifts or trend changes) exist under both the null and alternate hypotheses. The problem is under active consideration [52–54].

5.1 ADF

The ADF test is a variation of the Dickey Fuller test for trend stationarity in the possible presence of unit root. It has a null hypothesis of unit root against an alternative of stationarity after compensation for auto-correlation [48, 55]. The ADF test has relatively low power, and in this type of application a finding of a UR may be because of a single deterministic permanent shift or trend-change [56], as noted above.

Eq. (6) is expanded to allow for multiple lags in the case of the Augmented Dickey Fuller (ADF) test, taking advantage of the recursive nature of the formula. This is more explicit below where k multiple lags are included as $\sum_{j=2}^{k} \rho_j \Delta y_{t-j+1}$. The difference series is then computed,

$$\Delta Y_i = b_0 + b_1 t + (\rho_1 - 1) Y_{i-1} + \sum_{j=2}^{k} \rho_j \Delta y_{t-j+1} + e_t \tag{7}$$

A unit root exists if $\rho_1 = 1$. The number of lags can be specified by the user or, as here, selected by using an information criterion.

The ADF test implementation used is programmed in R, available in the package 'urca' [57], and estimate and removes auto-correlation then applies a DF test. The code allows for three variants, are available, (a) a unit root, (b) a unit root with drift, and (c) a unit root with drift and a deterministic time trend – which corresponds to the model of Eq. (7) (above) and which we use. We select suitable autocorrelation lags on the basis of an information criterion, using the call "ur.df (ys, type = "trend", lags = 7, selectlags = "AIC")" following Hacker [58]. The resulting possible reduction in power in the test (inability to distinguish unit root from near unit root) is compensated by other tests in the suite. The test assumes no exogenous change, and H_0 may be accepted in the presence of one ([59], page 76).

5.2 KPSS

There are two variant of the KPSS test used here to test for level and trend stationarity. These tests invert the sense of the testing with respect to the ADF test, rejecting an H_0 of stationarity in favour of H_1, a presumption of a unit root. In this case a regime shift may well appear as H_1, with a step change being non level stationary and a trend change being non trend stationary. We use the R package 'tseries' [60] and invoke the two tests as kpss.test(ys), to test for level stationarity (henceforth KPSS-L) and kpss.test(ys,null = "Trend") to test for trend stationarity (henceforth KPSS-T).

KPSS tests are designed to give weight to stationarity. Assuming that the time-series can be decomposed into the sum of a deterministic trend, a random walk and a stationary error, the model of Eq. (6) is re-parameterised as follows with r_t representing the random walk

$$\begin{aligned} Y_t &= r_t + \beta t + u_{1t} \\ r_t &= r_{t-1} + u_{2t} \end{aligned} \tag{8}$$

Where u_{1t} is a stationary process, and u_{2t} is an i.i.d. process with zero mean and a variance σ^2.

If $\sigma^2 = 0$ then r_t is constant and the stationary processu_{1t} dominates. If not, then a unit root enters via u_{2t} and r_t is a random walk. Under a random walk, variance increases with time. Therefore this expectation is tested by estimating the variance using the Newey-West estimator [61] s^2 . To test for trend stationarity, a residual series ($\{e_1..e_n\}$) is given by residuals of an OLS linear regression ($\{e_1..e_n\}$).). To test for level stationarity the residual series is replaced by $e_t = y_t - \bar{y}$. Then for both cases, partial sums of residuals are defined as $S_t = \sum_{i=1}^{i} e_i$ and for T samples, the test statistic is given as

$$LM = \frac{\sum_{i=1}^{T} S_i^2}{s^2 T^2} \tag{9}$$

Both the ADF test and the ZA test below, perform by estimating an auto-regression parameter by OLS, whereas the KPSS tests examine the properties of the variance of the time series (KPSS-L) or of the difference series (KPSS-T).

5.3 Zivot-Andrews test

The previous tests are confounded by deterministic/exogenous change (steps or shifts), and additionally a combination non-deterministic and deterministic change must be detected.

The Zivot-Andrews test (ZA) [50] tests for the presence of a unit root (with a possible deterministic/exogenous change) against an alternative of stationarity with at most one exogenous change. An advantage is that the test also returns a time of a possible exogenous change [62] – but note that an exogenous change can be any of step, transient or trend change.

The code is in the R package "urca", called as "ur.za(ys, model = "both")", which allows for changes in trend or steps. H_0 is UR without exogenous change. H_1 is trend-stationary with a possible exogenous change at an unknown time.

The ruling assumptions are (a) that there is at most one exogenous structural change (b) in a multivariable model, that only one exhibits unit root. In either of these cases other tests are preferred [54]. Here, we are testing a single variable with intervals bounded by breaks within which we have already detected exactly one break, whilst others may be below a detectability threshold. It has been previously shown that rejection of the null of a unit root could be due to a structural break even in the presence of unit root [63], whilst the presence of more than one break in the absence of a unit root may lead to the acceptance of the H_0 of UR [64].

Acceptance of H_0 does not imply merely UR, but rather, UR without exactly one deterministic break, [56], and thus H_1 means not UR or not a single break. Given we know there is a break (detected by MSBV, confirmed by ANCOVA), H_1 is reinterpreted as not UR, or more than one break.

The model used here is that documented by Zivot and Andrews (50) as Model (C). The model follows the ADF approach and its equation contains more complex parameters for: intercept and change of intercept (a step-like change), $\hat{\mu} + \theta DU_t(\hat{\lambda})$; and trend and change of trend, $\hat{\beta}t + \hat{\gamma} DT_t^* (\hat{\lambda})$. The remaining parameters are similar to the ADF; autocorrelation with lags, $\hat{\alpha} y_{t-1} + \sum_{j=1}^{k} \hat{c}_j \Delta y_{t-j}$ and the presumed i.i.d. error ...

$$y_t = \hat{\mu} + \theta DU_t(\hat{\lambda}) + \hat{\beta}t + \hat{\gamma} DT_t^* (\hat{\lambda}) + \hat{\alpha} y_{t-1} + \sum_{j=1}^{k} \hat{c}_j \Delta y_{t-j} + e_t \tag{10}$$

Circumflexes above represent estimates of parameters. $\hat{\lambda}$ is a value that is minimised during the search for the most likely time of a break, $DU_t(\hat{\lambda}) = 1$ if $t > T\lambda$, the time of change, 0 otherwise, and $DT_t^* (\hat{\lambda}) = t - T\lambda$ if $t > T\lambda$, 0 otherwise. Parameters estimated include the time of change and each of the parameters of the above model. $\hat{\lambda}$ is estimated so as to minimise the one side t-statistic for $\alpha = 1$, which in turn leads to rejection of the null. One should note that in the absence of any deterministic change-point the test functioned as a stationarity test when empirically assessed.

5.4 Empirical quantification of false determination rates

All of these tests are posed as null hypothesis tests. As such they only reject the null hypothesis at a particular level once sufficient evidence is found against it, and when the data size is limited, the power (the probability of correctly rejecting the null hypothesis) is similarly reduced. Therefore, in R2019 (page 100) the four tests were each tested separately for their false positive and false negative rates using a Monte Carlo method.

This aids interpretation since data segments vary in length. Before proceeding further, one may ask how meaningful the nominal p-values are, or as in R2019, one can determine the minimum data length required to allow acceptance of a finding of both UR and non-UR separately, for each test.

5.5 Applying these UR tests

Assuming an objective change-point method has been used bounded between two objectively determined change-points. Do the assumptions of the detection method hold for the segment of data and to what extent?

These tests are all applied to the segments of data within which a single change-point has already been provisionally identified. The change-point itself is not otherwise considered. However, since the climate data being tested provisionally contains a deterministic change and only the ZA test is formulated with this as a ruling assumption, findings of non-stationarity may be caused by the presence of additional deterministic change-points below detection thresholds.

Level stationarity is not simply a zero trend, since data with zero trend may be either deterministically or stochastically level, and even if deterministic may not be linear. A deterministic change-point detection method may return indeterminate change-points given non-linear trend. The residuals around stochastic trend will retain a UR characteristic. Trend stationary data has level stationary residuals, as do discontinuous trend stationary data fitted appropriately.

A segment of data with a valid change point should not be found to be level stationary, it should not be in a segment with unit root behaviour, and if it shows trending behaviour this should not be due to a drifting unit root. It should also have low p-values by ANCOVA.

5.5.1 Level stationarity

The KPSS-L test is used here with an expectation that segments of climate data in which a change-point occurs contain a step-like shift but may also contain a change of trend. Hence it is used as a cross check. Further, once the deterministic internal shift and trend components are removed the residual should be both level and trend stationary. Level non-stationarity in the segment and level stationarity in the residuals supports the existence of a change-point.

5.5.2 Trend stationarity

Data with a provisional change of trend is expected to be non-trend-stationary. Data with constant trend and a step-like change may show as trend stationary depending on the assumptions of the specific test. The KPSS-T test and the ADF test as formulated here may return different results in the presence of a step-like shift and no trend change, with the ADF test showing trend stationarity and the KPSS showing non-stationarity.

5.5.3 Unit root/non-stationarity in the absence of any deterministic change

The presence of a unit root may cause the data to mimic either a step-like change or a change of trend. In either case the MSBV can return a step-like change. All four unit-root tests are expected to detect this, with the ADF being less powerful, partly due to a potential to overfit autocorrelation lags. Since the detection method has provisionally detected a change-point, tests on the residuals would likely all show non-stationarity, and similarly testing of the segment itself. The ZA test would likely be the most powerful.

5.5.4 Unit root/non-stationarity in the presence of deterministic change

This is a complex issue. The combination of UR and deterministic trend is potentially explosive [58]. On the other hand the climate system is physically bounded and so at worst the combination may appear as step-like. If tests support unit root in both the segment data and its residuals, either a genuine unit root is present or multiple deterministic changes are. Data with apparent UR that disappears in the residuals is consistent with a single deterministic change. However data with multiple change points is misspecified for all tests.

5.5.5 Misspecification due to use of averaging

As discussed, the TM/SI identifies the sampling model as a point of consideration, and data conditioning may itself be a misspecification to a given investigation. Climate data is not homogenous. Averaging is often assumed to increase the signal to noise ratio (S/N) but more localised features may fall below detectability thresholds. For step-like changes occurring at different times in different components, the steps are diluted but potentially, autocorrelation is induced. Further, if the changes differ slightly in time over a number of components then the deterministic shift-like changes may be confused with either stochastic or deterministic trend. Similarly trend changes: Only if the step-like or trend change happens simultaneously across all processes will the S/N increase. Data conditioning methods that imposes or presume smoothing may turn steps into trends. If autocorrelation is present as part of the signal in the component's data together with trend, the situation is still more complex.

Table 1 summarises the conditions that can be diagnosed with these UR tests. **Tables 2** and **3** provide interpretations.

5.6 The averaging of multiple datasets with autocorrelation

The "order" of an AR process is the number of lags, and also the polynomial order required to fit the error terms. The Dickey-Fuller equation (Eq. (6)) describes an autoregressive single lag, i.e. an AR(1) process. The sum of two AR(1) processes is most compactly represented as an autoregressive-moving average (ARMA) process of greater order, ARMA(2,1) [65].

If p and q are the lag order of processes, then two AR processes combine into an ARMA process, where the first parameter of the ARMA is the order of the AR part, and the second is the order of the moving average (MA) part.

$$AR(p) + AR(q) = ARMA(p+q, \max(p,q)) \tag{11}$$

Properties of Test	**ADF (trend and drift)**	**KPSS (level stationarity)**	**KPSS (trend stationarity)**	**ZA**
Ruling assumptions	No exogenous change.	No exogenous change.	No exogenous change.	Not combined exogenous change and unit root. At most one exogenous change (shift or trend change).
Null hypothesis, H0	I(1) Unit Root after allowing for autocorrelation and trend.	Stationarity	Trend stationarity	I(1) Unit Root with drift and no exogenous change
Contrast hypothesis, H1	Presumption of trend stationarity	I(0) Unit Root	I(1) Unit Root	Deterministic with possible exogenous change at a date
When at most a single exogenous change is present				
If exogenous change and UR present	Accept H0, i.e. UR	Prefer H1, i.e. UR	Prefer H1, i.e. UR	May prefer H1, i.e. exogenous change
If exogenous change but UR not present	May accept H0, i.e. UR	If constant trend and step-change then will accept H0, stationarity. Otherwise prefer H1, i.e. UR	If step-change only then will accept H0, trend stationarity. A strong trend change will prefer H1, i.e. UR	Prefer H1, i.e. exogenous change
Unit root but no exogenous change	Accept H0, i.e. UR	Prefer H1, i.e. UR	Prefer H1, i.e. UR	Accept H0, i.e. UR
When multiple exogenous changes are present				
Plus Unit Root	Accept H0, i.e. UR	Prefer H1, i.e. UR	Prefer H1, i.e. UR	Accept H0, i.e. UR
No Unit Root	May accept H0, i.e. UR	Prefer H1, i.e. UR	May prefer H1, i.e. UR if exogenous trend changes present	May accept H0, i.e. UR
After removal of all exogenous change				
If Unit Root	Accept H0, i.e. UR	Prefer H1, i.e. UR	Prefer H1, i.e. UR	Accept H0, i.e. UR
No Unit Root	Prefer H1, trend stationarity	Accept H0, stationarity (unless residual trend remains)	Accept H0, trend stationarity (unless residual trend remains)	Prefer H1, exogenous change (even if there is none)
After removal of main exogenous change but with exogenous change still present				
If Unit Root	Accept H0, i.e. UR	Prefer H1, i.e. UR	Prefer H1, i.e. UR	May prefer H1 if exactly one exogenous change remains. H0 of more than one.

Properties of Test	ADF (trend and drift)	KPSS (level stationarity)	KPSS (trend stationarity)	ZA
No Unit Root	Prefer H1, trend stationarity (even if residual trend remains)	Accept H0, level stationarity	May prefer H1	Prefer H1, exogenous change

Table 1.
Unit root tests used and their main assumptions (reproduced from R2019, Table Ch4.1.2). Possibilities not formally considered may deceive these tests by supporting either the null or contrast hypotheses.

Initial data with a presumptive step change	Residual with internal step and trends removed	Interpretations
H_0 rejected, accept as Exogenous/Stationary	H_0 rejected, accept as Exogenous/Stationary	There is a deterministic change with stationary residual.
	H_0 not rejected, accept as Endogenous/Non stationary	There is a deterministic change with non-stationary residual
H_0 not rejected, accept as Endogenous/Non stationary	H_0 rejected, accept as Exogenous/Stationary	Residual is non-stationary with two deterministic changes
		Residual is stationary with two deterministic changes
	H_0 not rejected, accept as Endogenous/Non stationary	Residual is non-stationary with zero exogenous changes: step-change is false positive
		Residual is stationary apart from two or more undetected change-points
		Residual is non-stationary with more than two deterministic change-points

Table 2.
Reproduced from R2019 Table Ch4.1.3: Expected outcomes of the Zivot Andrews test, given data with a presumptive step-like change plus a variety of additional conditions. The first and second columns define results of the tests on the initial data segment and the residual with internal step and trend removed. The last column lists interpretations of the pairs of results.

Treating the result of $AR(1) + AR(1) = ARMA(2, 1)$ as an AR(1) process may be deceptive. And yet in many analyses, the issue of the composition of the data is at best brushed off, and autocorrelation is in general approximated as AR(1). In R2019 apparent unit root-like behaviour in some zonal ocean temperature data sets resolves to deterministic shifts at different times in sub-sectors of those zones, and this affects the determination of change-points.

5.7 Reasoning about change-points

It is possible to examine the data segment and its residual and to have greatly increased confidence that the change-point methods are adequate to the task, and to broadly classify change-points detected as potentially affected by (a) misspecification of the detection tests with data out of its applicability range, (b) random-walks, (c) presence of undetected change-points, (d) some forms of model family misspecification.

Initial data with a presumptive step change	Residual with internal step and trends removed	Interpretations
KPSS-T H_0 not rejected accept as Stationary. *ADF* H_0 rejected accept as Stationary.	*KPSS-T* H_0 not rejected accept as Stationary. *ADF* H_0 rejected accept as Stationary.	Residual is stationary, the single change-point did not have a trend change
	KPSS-T H_0 rejected accept as Non stationary. *ADF* H_0 not rejected accept as Non stationary.	Location of a single change-point misidentified so that the trend is also miscalculated
KPSS-T H_0 rejected accept as Non stationary. *ADF* H_0 not rejected accept as Non stationary.	*KPSS-T* H_0 not rejected accept as Stationary. *ADF* H_0 rejected accept as Stationary.	Residual is stationary and change-point included a trend change
	KPSS-T H_0 rejected accept as Non stationary. *ADF* H_0 not rejected accept as Non stationary.	The data segment is non-stationary and the provisional change-point may be a false positive.
		Residual is non-stationary. The initial segment contained a step and/or trend change.

Table 3.
Reproduced from R2019 Table Ch4.1.4: Expected outcomes of the KPSS-T and ADF tests, given data with a presumptive step-like change plus a variety of different conditions.

In Chapter 4 of R2019 the KPSS-T, ADF and ZA tests are combined to provide a classification scheme for change-points (**Table 4**). Using this classification scheme it became straight-forward to determine that regime changes over land and ocean

Classification	Reasoning and interpretation
Single, non-stationary	We accept the single exogenous change, but the residuals are not stationary, leaving open the possibility of undetected features. The ZA has reverted from Exogenous/stationary to Endogenous/non-stationary in the residuals, consistent with a single exogenous change plus a presumptive unit root. The presumptive unit root in the residuals is not reliably separable from multiple change-points below detectability.
Single, Stationary	We accept the step-change detected by the MSBV as the single exogenous change with no stochastic trend. The residuals are stationary supporting the single change-point. The ZA test does not change from exogenous/stationary
Single, N/A	We accept the step-change detected by the MSBV, without a valid ZA result, noting that there is insufficient data to probe further.
Non-stationary	We have evidence that the data segment contains sufficient non-stationarity as to cast doubt on the MSBV. The ZA test does not revert from endogenous/non-stationary and neither do the other tests. Hence the removal of a single change-point has had no apparent effect. Multiple change-points on top of a non-stationary background is too complex a situation to detect with these tests.
Multiple, Stationary	We may be dealing with a pair of exogenous changes. The ZA reverted from non-stationary to stationary with other tests consistent with this. Potentially a single additional undetected change-point, since two exogenous changes may be classified as an endogenous change in the ZA.
Stationary	Possible false positive or weak change in stationary data
N/A	Not classifiable/indeterminate

Table 4.
Extended from R2019 Table Ch4.1.5: classifications of data segments.

differ in complexity. Sharpening the testing, it also further supported the principal findings of R2019, that abrupt shifts relate directly to warming; in their extent, frequency and intensity; and more so at finer scale. For this paper the last two additional classes apply when ANCOVA does not support a change-point. An example is provided in the appendix.

6. Examples

Figure 1 below, illustrates the difference between analysis, commencing with the MSBV, of global mean temperature and the area averaged Northern mid-latitude (NML) temperature. While the step change after 1996 is very obvious in the zonal data, the change is sometimes disputed in the global signal. **Table 5**, adds strong support to the contention that the so-called hiatus was a significant event, but not on the basis of trends. The 1988 event in the NML corresponds to an atmospheric reorganisation and extensive biophysical changes regionally [10]. All of the change-points occur in data which is otherwise stationary.

Figure 2 below, illustrates the contribution to reasoning about the nature of decadal climate regimes which follows from a reasoned classification scheme. If the global temperatures are averaged over smaller areas, and then step-change points are calculated it becomes more likely that the data will present as stationary. This shows that the zonal data are not homogeneous with respect to regime shifts; that regime shifts are more regional. Note also the difference between land (almost always stationary) and ocean (less so), supporting the ocean as being more complex. There is also a tendency for land shifts to be a year or two delayed.

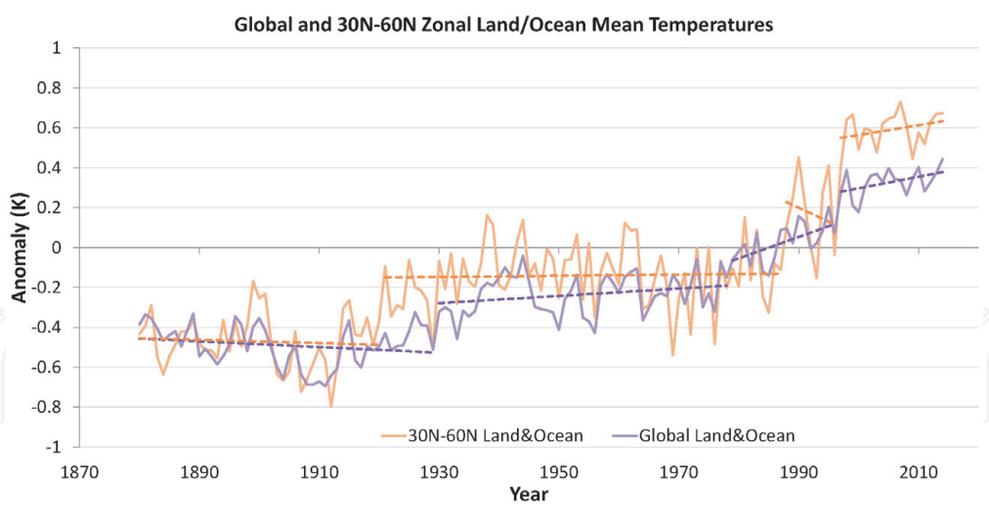

Figure 1.
Adapted from R2019, Figure Ch3.11. Step-like shifts in the Northern extra-tropics compared to those detected in global data.

7. Conclusions

The principal contribution of this paper is to expand on the use of misspecification testing to strengthen reasoning about abrupt shifts in time-series. We focus on climate data records and statistical model specification with respect to the data. Probing the misspecification of statistical models helps ensure that tests better represent probative criteria, and better distinguish between them.

Zone	MSBV			KPSS-L		KPSS-T		ADF		Zivot Andrews		ANOVA/ANCOVA			
	First Changed Year	Internal Shift (°C)	Internal Trend Change (^{0}C/Yr)	Data segment	Residuals	Data segment	Residuals	Data segment	Residuals	Data segment	Residuals	First Changed Year	ANOVA-Internal Shift(Pr)	ANOVA-Trend Change (Pr)	ANCOVA-Change-point(Pr)
30 N–60 N	1921	0.34	0.001	NS	S	NS	S	NS	S	S	S	1921	***	—	***
30 N–60 N	1988	0.37	−0.014	NS	S	NS	S	NS	S	S	S	1964	**	—	***
30 N–60 N	1997	0.43	0.019	NS	S	S	S	NS	S	NS	S	1997	***	—	**
Global	1930	0.25	0.003	NS	S	NS	S	NS	S	S	S	1914	***	*	***
Global	1979	0.12	0.009	NS	S	NS	S	NS	S	S	S	1946	**	*	***
Global	1997	0.16	−0.005	NS	S	S	S	S	S	S	S	1997	***	—	**

Table 5.
*Adapted from R2019 Table A4.1.30: For the UR tests red text denotes results of tests where the data length may affect precision. NS = non-stationary, S = stationary. *** $p <= 0.001$, ** $0.001 > p <= 0.01$, * $0.01 < p >= 0.05$.*

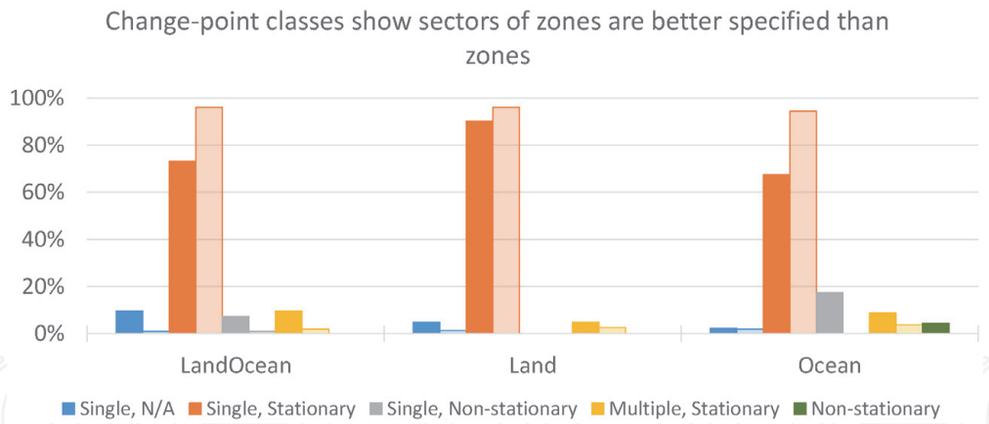

Figure 2.
Adapted from R2019, Figure Ch5.5, Data becomes less complex at finer scale, as evidenced by classification. The proportions of each class of change-point are shown for the same data averages over 30 degree zones (saturated colors) and 45 degree sectors of the same zones (unsaturated colors).

The TM/SI framework has been suggested as a variation on previously discussed inductive frameworks. While it assists adequate testing of physical hypotheses; none the less, climate is a complex system [66]. Thus the ongoing assessment of testing procedures, and with it model specification.

Misspecification testing supports severe testing. Severe testing is strengthened by improved one to one mapping between physical features and statistical test outcomes. The tests outlined here assist a probative analysis, firstly by adding nuance to the findings, and secondly by providing the basis of a change-point classification, they assist strong reasoning. They have been selected because they are individually automatable and complementary, and the utility of this has been indicated in the case study. The chain of reasoning involved in the use of multiple tests is complex but the final classification scheme is compact and as seen, informative.

A basis has also been established for potentially detecting signatures of a data composition misspecification whereby features emerge or submerge in composited data due to averaging of signals (especially ones moving in time and space). The signature is a reduction in non-stationarity when signals are decomposed or segmented using the MSBV as seen in **Figure 2**. The same issue also affects both autocorrelation and trend analysis simply because step-like dislocations in data are generally deceptive for the regressions embedded in many general methods.

Acknowledgements

R.N. Jones is a Professorial Fellow of Victoria Institute for Strategic Economic Studies in Melbourne. J. H. Ricketts was the holder of a Victoria University postgraduate research scholarship. The anonymous reviewers of a joint paper and a joint conference paper, and two thesis reviewers, all contributed substantial improvements and assisted development of the ideas expressed here. We would like to acknowledge with both gratitude and sorrow the lasting influence of our colleague and friend, Dr. Penny Whetton, who passed away too soon.

Notes

A number of tables and figure are adapted from the PhD thesis of JH Ricketts [13], mostly chapter 4. A peer reviewed joint paper [21] and a joint conference paper [21] are also sourced.

Koninkliijk Nederlands Meteorologisch Instituut (KNMI) make available the KNMI Climate Explorer and this was a valuable resource.

Other data has been sourced from, Met Office Hadley Centre, NASA, Goddard Institute for Space Studies and United States National Climatic Data Center.

A. Appendix

During sensitivity testing of the detection and characterization tests in R2019 simulations were run, including assessments of (a) the effects of shifts single and multiple shifts below detection thresholds, (b) multiple shifts close in time, (c) high levels of autocorrelation, (d) state switching between deterministic and stochastic data, and (e) curvilinear trends. This illustrative example is an extension of one part of that work.

A.1 Synthetic climate-like data

Following R2019, a suite of four artificial multi-step time series ('A' to 'D') was constructed and analyzed by MSBV then validation tests were run against both the shifts as detected by MSBV and as originally defined.

A is an artificial 200 year annual temperature consisting of random data (and a standard deviation, σ, of 0.44) with lag 1 autocorrelation of 25%, lag 7 autocorrelation of 10%, centered about zero, plus a quadratic trend curve rising 2.1 degrees. The degree of autocorrelation is consistent with the findings of [67].

Eight shifts random shift level (mean 1.5 σ) are added at defined times (Shifts of 1.5σ are less than MSBV reliability threshold) (**Table 6**).

To assess the suite presence of UR with deterministic trends plus shifts, shifts without trends, and UR alone, red-noise (summed white noise, μ = 0, σ = 0.44) was added to A to produce set B, set C is the defined steps plus red-noise and D is red-noise only.

A.1.1 Studentized Breusch-Pagan test for heteroskedasticity

The studentized BP test was run for the disjoint regression of breaks detected (Break model), and also for the breaks as defined (**Table 7**). A linear model and a

Year	**1954**	**1982**	**1998**	**2029**	**2035**	**2054**	**2070**	**2096**	**Total**
Shifts in K	0.57	**0.34**	0.72	0.85	1.00	0.61	0.94	**0.31**	5.34
(σ)	(1.30)	**(0.77)**	(1.64)	(1.93)	(2.27)	(1.39)	(2.14)	**(0.70)**	(12.14)

Table 6.
Adapted from R2019, Table Ch4.1.6 Synthetic Data Timing and extent of Shifts. Total Rise is shown both as anomaly and as standard deviations. Shifts of <0.5 are not guaranteed to be found by MSBV and are bolded.

Dataset	**Break model**	**Defined**	**Linear model**	**Quad model**
A.	0.6802	0.6959	0.0241	0.0001
B.	0.0034	0.0270	0.0000	0.9108
C.	0.0000	0.0039	0.0000	0.5205
D.	0.2870	0.0024	0.0000	0.0457

Table 7.
Studentized Breusch-Pagan Test results. Green denotes 0.01 < p < 0.05, red 0.01 > p, black p > 0.05. A null hypothesis of homoscedasticity is rejected for low p-values.

quadratic model were also run for comparison. Data sets A and D appear to have homoskedastic residuals for their breaks given the detected shifts, and yet A is deterministic and D is non-deterministic. Datasets B and C, on the other hand appear homoskedastic given a quadratic model. Note that the SBP operates under an i.i.d assumption which is violated by sets B, C and D.

The breaks returned by the MSBV, and breaks defined, for A both form an adequate model. The breaks returned by the MSBV for D form an adequate model whereas the defined set – not present in D – does not. B contains a curvilinear trend, plus along with C, shifts which also induce an apparent curvilinearity. As can be seen from **Figure 3** the MSBV does quite well at locating the change-points.

Note: The data tested, residuals of detected change-points, will almost always appear to be deterministic when tested by the UR tests. This is because the residual

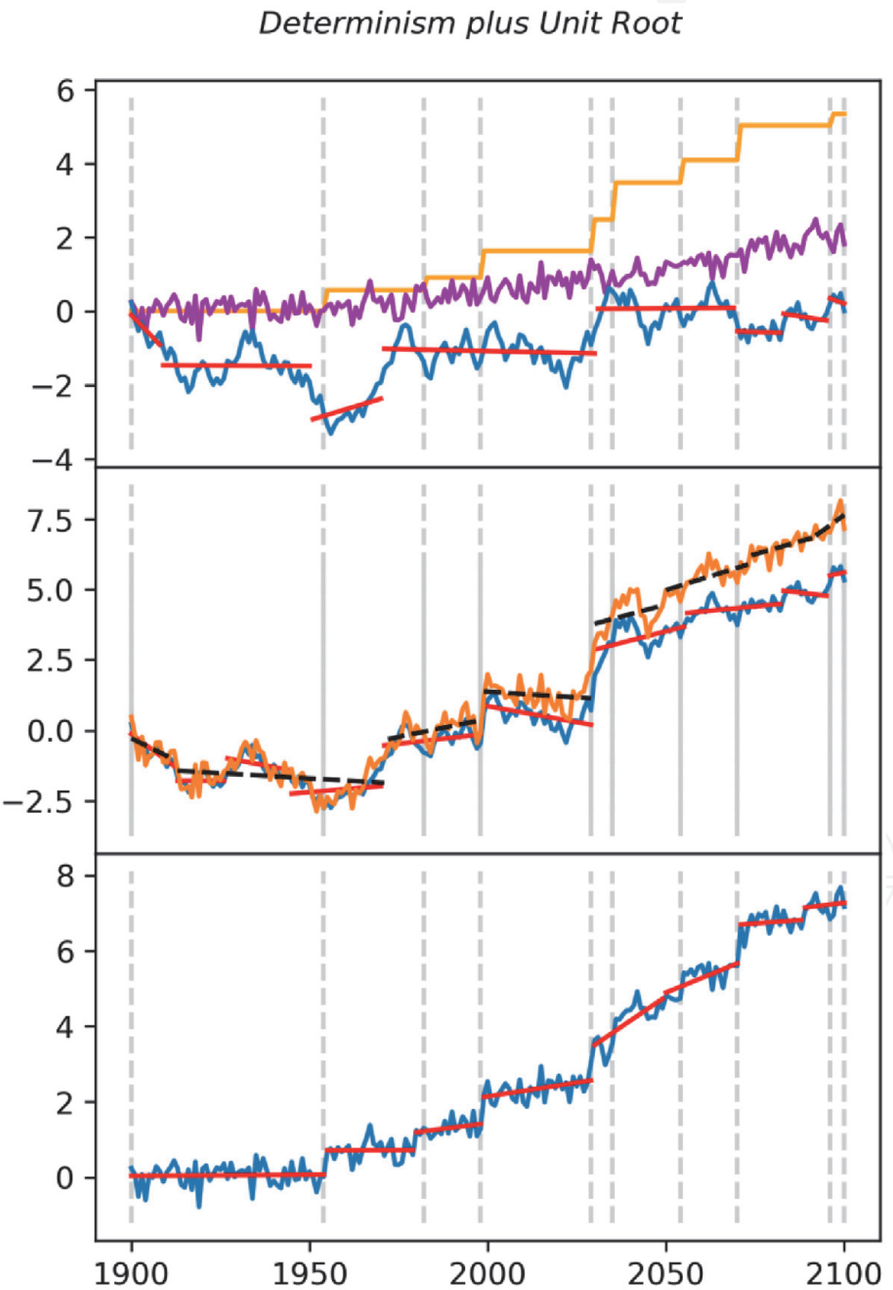

Figure 3.
Top: Blue is data set D, red is break-segments determined by MSBV. Magenta is auto-correlated noise plus quadratic trend. Orange is the defined shifts. Middle: Dataset C is shown as orange (black breaks), dataset B is blue (red breaks). Bottom: dataset A in blue, red is break-segments determined by MSBV. Vertical grey reference lines indicate defined shifts.

Defined	A	B	C	D
		1912(++S)	1912(NS)	1908(+NS)
1954	1954(+S)		1926(NS)	
			1944(+NS)	1950(+NS)
		1971(++S)	1970(+NS)	1970(NS)
1982	1979(+S)			
1998	1998(+S)	1998(+S)	1998(NS)	
2029	2029(+S)	2029(++S)	2029(+NS)	2030(NS)
2035				
2054	2049(S)	2049(S)	2055(+NS)	
2070	2070(+S)	2073(S)		2069(+NS)
	2088(S)		2082(++S)	2082(+)
2096		2091(−)	2095(+)	2095(+)

Table 8.
Changes defined and years detected in each dataset. Annotations denote segment classification, + is single change-point, ++ possible multiple changes, S is stationary, NS is non-stationary. Red denotes ANCOVA $p < =0.05$.

of deterministic signal is expected to be deterministic, whereas the residual of a purely non-deterministic signal from which a deterministic components has been subtracted *acquires* a deterministic component and appears mean-reverting, i.e. I (0). There is no current method for dealing with multiple deterministic changes in a UR time series, and blended series such as B and C will not meet the criteria of a UR series. In fact looking at D only through the lens of the SBP and UR tests of the residuals does not distinguish it from a deterministic time series like A. The difference only becomes apparent when the individual change-points are tested (see **Tables 8** and **9**).

A.1.2 Analysis of individual change-points

The full analysis results are available on-line at https://cdn.intechopen.com/public/docs/230558_files.zip.

Set A. One pair of defined change-points violated an assumption of the MSBV that rejects shifts within a seven year refractory period (defined as 2029, 2035), selecting only 2029 which registers as a strong shift embedded in stationary data with an internal trend (notably the ZA test of the residuals locates 2035). When the data is broken up according to the defined shifts, 2035 registers as a strong shift in non-stationary data, and evidence for the internal trend weakens. The defined small shift in 2054 following 2035 was attributed to 2049 after 2029 but not supported by ANCOVA and the segment was classified as non-stationary. The ZA suggests a change in 2034 but non-stationarity in the residuals. All other change-points were detected as defined and classified as single change-points in trend-stationary data.

Datasets B though D represent increasingly UR dominated data. For B (combining deterministic trends and red noise), the only detected shift that is classified as a single shift in stationary data is 1998, all prior being classified as having possible multiple sub-detection shifts, and all following being rejected by ANCOVA although the segments are classed as stationary. Sets C (UR with shifts) and D (UR only), show that the MSBV by itself is vulnerable to non-determinism.

DataSet	Breaks	Single, Non-stationary	Non-stationary	Single, N/A	Stationary	Multiple, Stationary	Single, Stationary	N/A	Sum
A.	Found	0	1	0	1	0	5	0	7
A.	Defined	0	1	0	2	0	5	0	8
B.	Found	0	0	0	2	3	1	1	7
B.	Defined	0	2	0	2	1	3	0	8
C.	Found	4	3	1	0	1	0	0	9
C.	Defined	3	3	0	0	1	1	0	8
D.	Found	3	2	2	0	0	0	0	7
D.	Defined	4	2	1	1	0	0	0	8

Table 9.
Numbers of change-points assigned to each class. Note that C and D differ from A and B by having non-stationary residuals, where as B differs from A by displaying evidence of undetected multiple change-points.

The principal indication that a change-point dominated time-series has an underlying difference stationarity (i.e. red, or brown noise) is given by examination of the segmentation and not the residuals.

Author details

James Ricketts* and Roger Jones
Victoria University, Melbourne, Australia

*Address all correspondence to: james.ricketts@live.vu.edu.au

References

[1] Mayo DG. An error-statistical philosophy of evidence. The nature of scientific evidence: Statistical, philosophical and empirical considerations. 2004:79-96.

[2] Mayo DG, Spanos A. Methodology in Practice: Statistical Misspecification Testing. Philosophy of Science. 2004;71 (5):1007-25. doi: 10.1086/425064.

[3] Mayo DG, Spanos A. Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. The British Journal for the Philosophy of Science. 2006;57(2):323-57.

[4] Mayo DG, Spanos A. Error statistics. Handbook of the philosophy of science. 2011;7:153-98.

[5] Spanos A, Mayo DG. Error statistical modeling and inference: Where methodology meets ontology. Synthese. 2015;192(11):3533-55.

[6] Mayo DG. Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars: Cambridge University Press; 2018.

[7] Jones RN. Detecting and attributing nonlinear anthropogenic regional warming in southeastern Australia. Journal of Geophysical Research: Atmospheres (1984–2012). 2012;117 (D4).

[8] Belolipetsky P. The Shifts Hypothesis-an alternative view of global climate change. arXiv preprint arXiv: 14065805. 2014.

[9] Belolipetsky P, Bartsev S, Ivanova Y, Saltykov M. Hidden staircase signal in recent climate dynamic. Asia-Pacific J Atmos Sci. 2015;51(4):323-30. doi: 10.1007/s13143-015-0081-6.

[10] Reid PC, Hari RE, Beaugrand G, Livingstone DM, Marty C, Straile D, et al. Global impacts of the 1980s regime shift. Global change biology. 2015.

[11] Bartsev S, Belolipetskii P, Degermendzhi A, editors. Multistable states in the biosphere-climate system: towards conceptual models. IOP Conference Series: Materials Science and Engineering; 2017: IOP Publishing.

[12] Jones RN, Ricketts JH. Reconciling the signal and noise of atmospheric warming on decadal timescales. Earth Syst Dynam. 2017;8(1):177-210. doi: 10.5194/esd-8-177-2017.

[13] Ricketts JH. Understanding the Nature of Abrupt Decadal Shifts in a Changing Climate. Melbourne: Victoria University; 2019.

[14] Suppes P. Models of data. In: Nagel E, Suppes P, Tarski A, editors. Logic, methodology, and philosophy of science: Proceedings of the 1960 International Congress; Stanford, CA: Stanford University Press.; 1962. p. 252-61.

[15] Haig BD. Tests of Statistical Significance Made Sound. Educational and Psychological Measurement. 2016: 0013164416667981.

[16] Corti S, Molteni F, Palmer T. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. Nature. 1999;398 (6730):799-802.

[17] Branstator G, Selten F. "Modes of variability" and climate change. Journal of Climate. 2009;22(10):2639-58.

[18] Katzav J. Severe testing of climate change hypotheses. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics. 2013;44(4):433-41. doi: http://dx.doi.org/10.1016/j.shpsb .2013.09.003.

[19] Katzav J. Should we assess climate model predictions in light of severe tests? EOS, Transactions American Geophysical Union. 2011;92(23):195-.

[20] Katzav J, Dijkstra HA, de Laat ATJ. Assessing climate model projections: State of the art and philosophical reflections. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics. 2012;43(4):258-76. doi: http://dx.doi.org/10.1016/j.shpsb.2012.07.002.

[21] Ricketts JH, Jones RN. Characterizing change-points in climate series with a severe approach. In: Syme G, Hatton MacDonald D, Fulton B, Piantadosi J, editors. The 22nd International Congress on Modelling and Simulation (MODSIM2017); 3-8 December 2017; Hobart: The Modelling and Simulation Society of Australia and New Zealand Inc.; 2017.

[22] Salmon WC. Scientific explanation and the causal structure of the world: Princeton University Press; 2020.

[23] Mayo DG. Error and the growth of experimental knowledge: University of Chicago Press; 1996.

[24] Jeffrey SJ, Carter JO, Moodie KB, Beswick AR. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environmental Modelling and Software. 2001;16(4): 309-30.

[25] Cochrane W, Cox G. Experimental designs., 2nd edn (John Wiley & Sons: Sydney). 1957.

[26] Minobe S. A 50–70 year climatic oscillation over the North Pacific and North America. Geophysical Research Letters. 1997;24(6):683-6. doi: 10.1029/97GL00504.

[27] Mantua NJ, Hare SR, Zhang Y, Wallace JM, Francis RC. A Pacific interdecadal climate oscillation with impacts on salmon production. Bulletin of the american Meteorological Society. 1997;78(6):1069-79.

[28] Trenberth KE, Hurrell JW. Decadal atmosphere-ocean variations in the Pacific. Climate Dynamics. 1994;9(6): 303-19.

[29] Hope P, Drosdowsky W, Nicholls N. Shifts in the synoptic systems influencing southwest Western Australia. Climate Dynamics. 2006;26 (7-8):751-64. doi: 10.1007/s00382-006-0115-y.

[30] Rajaratnam B, Romano J, Tsiang M, Diffenbaugh N. Debunking the climate hiatus. Climatic Change. 2015:1-12. doi: 10.1007/s10584-015-1495-y.

[31] Lewandowsky S, Risbey JS, Oreskes N. The "Pause" in Global Warming: Turning a Routine Fluctuation into a Problem for Science. Bulletin of the American Meteorological Society. 2015. doi: 10.1175/BAMS-D-14-00106.1.

[32] Risbey JS, Lewandowsky S, Cowtan K, Oreskes N, Rahmstorf S, Jokimäki A, et al. A fluctuation in surface temperature in historical context: reassessment and retrospective on the evidence. Environmental Research Letters. 2018;13(12):123008.

[33] Cahill N, Rahmstorf S, Parnell AC. Change points of global temperature. Environmental Research Letters. 2015; 10(8):084002.

[34] Fyfe JC, Meehl GA, England MH, Mann ME, Santer BD, Flato GM, et al. Making sense of the early-2000s warming slowdown. Nature Climate Change. 2016;6(3):224-8.

[35] Meehl GA, Hu A, Arblaster JM, Fasullo J, Trenberth KE. Externally Forced and Internally Generated Decadal Climate Variability Associated with the Interdecadal Pacific Oscillation.

Journal of Climate. 2013;26(18): 7298-310. doi: 10.1175/JCLI-D-12-00548.1.

[36] Trenberth KE. Has there been a hiatus? Science. 2015;349(6249):691-2.

[37] Foster G, Abraham J. Lack of evidence for a slowdown in global temperature. US CLIVAR. 2015:6.

[38] Vives B, Jones RN. Detection of abrupt changes in Australian decadal rainfall (1890-1989): CSIRO Atmospheric Research; 2005.

[39] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: Journal of the Econometric Society. 1980:817-38.

[40] Rodionov SN. Use of prewhitening in climate regime shift detection. Geophysical Research Letters. 2006;33 (12).

[41] Mizon GE. A simple message for autocorrelation correctors: Don't. Journal of Econometrics. 1995;69(1): 267-88.

[42] Beaulieu C, Killick R. Distinguishing trends and shifts from memory in climate data. Journal of Climate. 2018;31 (23):9519-43.

[43] Percival DB, Overland JE, Mofjeld HO. Interpretation of North Pacific variability as a short-and long-memory process. Journal of Climate. 2001;14(24):4545-59.

[44] Stock JH. Unit roots, structural breaks and trends. Handbook of econometrics. 41994. p. 2739-841.

[45] Chang Y, Kaufmann RK, Kim CS, Miller JI, Park JY, Park S. Time series analysis of global temperature distributions: Identifying and estimating persistent features in temperature anomalies. 2016.

[46] Tsonis AA, Swanson K, Kravtsov S. A new dynamical mechanism for major climate shifts. Geophysical Research Letters. 2007;34(13).

[47] Fischer JW, Walter WD, Avery ML. Brownian Bridge Movement Models to Characterize Birds' Home Ranges: Modelos de Movimiento de Puente Browniano Para Caracterizar el Rango de Hogar de las Aves. The Condor. 2013; 115(2):298-305.

[48] Dickey DA, Fuller WA. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica: Journal of the Econometric Society. 1981:1057-72.

[49] Kwiatkowski D, Phillips PC, Schmidt P, Shin Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of econometrics. 1992;54 (1):159-78.

[50] Zivot E, Andrews DW. Further Evidence on the Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis. Journal of Business & Economic Statistics. 1992.

[51] Fukac M. Inflation Expectations in the Czech Interbank Market. 2005.

[52] Kejriwal M, Perron P. A sequential procedure to determine the number of breaks in trend with an integrated or stationary noise component. Journal of Time Series Analysis. 2010;31(5):305-28.

[53] Harvey DI, Leybourne SJ, Taylor AR. Testing for unit roots in the possible presence of multiple trend breaks using minimum Dickey–Fuller statistics. Journal of Econometrics. 2013; 177(2):265-84.

[54] Liddle B, Messinis G. Revisiting sulfur Kuznets curves with endogenous breaks modeling: Substantial evidence of inverted-Us/Vs for individual OECD

countries. Economic Modelling. 2015; 49:278-85. doi: http://dx.doi.org/10.1016/j.econmod.2015.04.012.

[55] Elliott G, Rothenberg TJ, Stock JH. Efficient tests for an autoregressive unit root. National Bureau of Economic Research Cambridge, Mass., USA; 1992.

[56] Byrne JP, Perman R. Unit roots and structural breaks: a survey of the literature. Paper provided by Business School-Economics, University of Glasgow in its series Working Papers with. 2006;(2006_10).

[57] Pfaff B, Zivot E, Stigler M. Unit Root and Cointegration Tests for Time Series Data. 2016.

[58] Hacker RS. The Effectiveness of Information Criteria in Determining Unit Root and Trend Status. Royal Institute of Technology, CESIS-Centre of Excellence for Science and Innovation Studies, 2010.

[59] Kočenda E, Černý A. Elements of time series econometrics: An applied approach: Charles University in Prague, Karolinum Press; 2015.

[60] Trapletti A, Hornick K, LeBaron B. Time series analysis and computational finance. 2017.

[61] Newey WK, West KD. Automatic lag selection in covariance matrix estimation. The Review of Economic Studies. 1994;61(4):631-53.

[62] Glynn J, Perera N, Verma R. Unit root tests and structural breaks: a survey with applications. Faculty of Commerce-Papers. 2007:455.

[63] Gay-Garcia C, Estrada F, Sánchez A. Global and hemispheric temperatures revisited. Climatic Change. 2009;94 (3-4):333-49.

[64] Lumsdaine RL, Papell DH. Multiple trend breaks and the unit-root hypothesis. Review of economics and Statistics. 1997;79(2):212-8.

[65] Granger CW, Morris MJ. Time series modelling and interpretation. Journal of the Royal Statistical Society Series A (General). 1976:246-57.

[66] Jones RN, Ricketts JH. The Pacific Ocean heat engine: global climate's regulator. Earth System Dynamics (for open review). 2019.

[67] Allen MR, Smith LA. Investigating the origins and significance of low-frequency modes of climate variability. Geophysical Research Letters. 1994;21 (10):883-6.