Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

Testing Competing Models of the Temperature Hiatus: Assessing the effects of conditioning variables and temporal uncertainties through sample-wide break detection

Felix Pretis[a], Michael L. Mann[b], Robert K. Kaufmann[c]

[a]Programme for Economic Modelling, Oxford Martin School, University of Oxford, Oxford, United Kingdom. *Corresponding author*: felix.pretis@nuffield.ox.ac.uk

[b]Department of Geography, George Washington University, Washington, DC, USA.

[c]Department of Earth and Environment, Center for Energy & Environmental Studies, Boston University, Boston, MA, USA.

Abstract

Explaining the recent slowdown in the rise of global mean surface temperature (the hiatus in warming) has become a major focus of climate research. Efforts to identify the causes of the hiatus that compare simulations from experiments run by climate models raise several statistical issues. Specifically, it is necessary to identify whether an experiment's inability to simulate the hiatus is unique to this period or reflects a more systematic failure throughout the sample period. Furthermore, efforts to attribute the hiatus to a particular factor by including that mechanism in an experimental treatment must improve the model's performance in a statistically significant manner at the time of the hiatus. Sample-wide assessments of simulation errors can provide an accurate assessment of whether or not the control experiment uniquely fails at the hiatus, and can identify its causes using experimental treatments. We use this approach to determine if the hiatus constitutes a unique failure in simulated climate models and to re-examine the conclusion that the hiatus is uniquely linked to episodes of la Niña like cooling (Kosaka and Xie, 2013). Using statistical techniques that do not define the hiatus *a priori*, we find no evidence that the slowdown in temperature increases are *uniquely* tied to episodes of la Niña like cooling.

Keywords: Temperature Hiatus, Climate Model, ENSO, Statistical Break

Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

1 Introduction

Despite steady increases in the concentrations of major greenhouse gases, global mean surface temperature has not risen significantly over the past two decades. The seeming failure of the widely held hypothesis that increases in radiative forcing due to anthropogenic emissions of greenhouse gases will cause surface temperature to rise, has made the hiatus in warming (herein "the hiatus") a major focus of climate research. Attributing the hiatus to its driving factors is pursued with a variety of approaches, including statistical models that are parameterized using observed patterns from climate records (e.g. Kaufmann et al. 2011) and experiments (dynamic simulations) run by coupled climate models.

Using coupled climate models to identify the causes of climate phenomena, such as the hiatus raise several philosophical and statistical issues. As described by Oreskes *et al.* (1994) open models, such as general circulation models, cannot be verified or validated. These obstacles are not insurmountable because efforts to identify the cause for the hiatus are based on a simpler criterion; a comparison of model accuracy. Using this criterion, climate model experiments often attempt to identify the cause(s) of the hiatus by simulating experiments that do and do not include certain causal factors. The logic of these experiments is based on the assumption that omitting/including the relevant mechanism(s) will affect the experiment's ability to simulate the hiatus. Put simply, including the correct driving mechanism/variable will increase the experiment's ability to simulate the observed climate relative to an experiment that omits the driving mechanism/variable. Consistent with this logic, Oreskes *et al* (1994) conclude "Legitimately, all we can talk about is the relative performance of a model with respect to …other models of the same site." This approach leads to an array of explanations for the hiatus that include the effects of the El Niño-Southern Oscillation (ENSO) (e.g. Kosaka and Xie, 2013), volcanic cooling (e.g. Santer et al. 2014), and changes in ocean heat (e.g. Meehl et al., 2011).

Although the logic of these experiments is scientifically valid, model results must be analysed using statistical methodologies that are consistent with the limits of evaluating the performance of open models. Changing the system boundaries of an open model changes the model's performance. This change makes it difficult to evaluate the explanatory power of variables. For example, efforts to identify the causes for the hiatus do so via treatment experiments that 'exogenize' a variable that was endogenous in the control experiment. If exogenizing the variable enhances the treatment models' performance relative to the control model during that hiatus, this is consistent with the notion that the exogenized variable plays an important role. But such consistency may not be meaningful because conditioning on the observed values of the exogenized variable gives the model more information about the climate system. Almost by definition, using additional exogenous variables in the treatment experiment will improve the fit of a dynamic simulation in an open system throughout the sample period. Consequently, efforts to identify the cause for climate behaviour, such as the hiatus, must demonstrate that the increased accuracy of the treatment experiment during the period of interest goes beyond an increase in model accuracy throughout the entire sample period.

Sample-wide assessments of simulation errors can provide an accurate assessment of whether or not the control experiment uniquely fails at the hiatus, and can identify its causes using experimental treatments. Simply comparing the model's error during the hiatus to the rest of the sample is likely to generate misleading results. This danger is highlighted by Christiano (1992) who shows that testing for a structural change (in this case in the simulation error) at a non-independently chosen time period is likely to find a change when none is present. Depending on the nature of the time series analysed, he finds that the critical value ($p = 0.05$) for an F test for a structural break is about 10, as opposed to a value of about 3 in the F distribution. This arises because the break dates are not chosen independently from the data being tested, and the critical values are not adjusted to reflect this pre-test examination of the data. These results spawn an extensive literature that describes tests to identify a structural change when the timing of the change is not known *a priori*, some of which are discussed in section 2.

Two aspects of the hiatus necessitate sample-wide assessments of the model error. First, the hiatus is not defined *a priori* . Rather, the hiatus in warming is defined *ex post* by empirical failures of either

theoretical models that suggest temperature should rise as the radiative forcing of greenhouse gases increase, or empirical simulations generated by climate models, which seem to simulate temperature poorly over the last 15 year. Because both of these methods identify the hiatus in warming from a sample-wide pre-test examination of model performance, the critical values that are used to assess the statistical significance of changes associated with the hiatus must account the increased probability of finding a break at a pre-selected date. Secondly, the hiatus period cannot be defined *a priori* based on physical measurements alone. Before the mid-2000s the highest global surface temperature anomaly relative to 1960-1991 was observed in 1998 at 0.531 degrees (Hadcrut 4 data - Morice et al. 2012). This would imply that the hiatus started in 1998 but temperatures in 1998 were raised in part by an El Niño event (Kaufmann *et al*., 2011). This implies that the hiatus may have started earlier or later. Temperatures drop after 1998, but rise from 0.293 in 2000 to 0.539 in 2005, and this value is not exceeded until 2010 (0.547). This pattern could be used to define 2005 as the start of (and 2010 as the end of) the hiatus. But temperature falls from 2010 through 2013, which could make 2013 the end of the hiatus. Even after identifying the hiatus by examining the model errors through the sample period, there are many possibilities for the start and end of the hiatus. This reinforces the notion that the hiatus is an arbitrary period that is defined by searching the entire sample period relative to a climate model and not an exogenously given break date.

Beyond the statistical issues associated with the hiatus being an *ex-post* phenomenon, the scientific method dictates a sample-wide search. If one wants to argue that the radiative forcing fails to simulate temperature at the hiatus, the scientific method dictates that this failure must be unique and/or more severe than failures in other portions of the sample period. If a model fails repeatedly throughout the sample period, one cannot conclude that yet another failure at the hiatus has any meaning unique to the hiatus

Similarly, if one wants to demonstrate that a given variable/mechanism helps explain the hiatus, the improvement must be unique to the hiatus. If expanding the conditioning variables improves the treatment experiment's ability to simulate temperature throughout the sample period, the improved performance during the hiatus is not unique to the hiatus and so the improvement engendered by an expanded set of conditioning variables does not tell us anything about what happened during the hiatus relative to the rest of the sample period.

We present a statistical methodology to identify model departures such as the hiatus based to on the detection of structural breaks when the presence and timing of changes is not known *a priori*. The lack of an *a priori* definition for the hiatus and the pre-test examination of the data mean that the method used to compare climate model simulations must account for the non-exogenous choice of the breakpoint by searching the entire sample for breaks, and the critical values used to evaluate test statistics must account for the sample-wide search. We illustrate the importance of these philosophical and methodological issues for climate change research by revisiting the methods and results reported by Kosaka and Xie (2013), who conclude that the hiatus is *uniquely* linked to episodes of la Niña like cooling. Kosaka and Xie (herein KX) generate this conclusion by using a Mann-Kendall test to compare the accuracy of an experimental treatment that conditions the coupled climate model on observed values for sea surface temperature in the Southern Pacific (POGA-H Experiment, see KX) to a control experiment, in which sea surface temperature (SST) is simulated endogenously (HIST Experiment, see KX). KX define the hiatus as 2002-2012 either through visual inspection or prior hypotheses.

Conditioning on observed SSTs and testing only the *a-priori* selected hiatus period embodies both difficulties described above. Focusing only on the hiatus ignores the possibility that the control experiment fails at times other than the hiatus, and that conditioning on observed SSTs improves the fit of a dynamic simulation by an open system throughout the sample period. Second, the failure to acknowledge the pre-selection to define the hiatus alters the critical values of the test statistic used to evaluate the null hypothesis of no break. This distorts the size of the test and makes false positives more likely (Christiano, 1992, Zivot and Andrews, 1992).

Using statistical techniques that do not define the hiatus *a priori*, the results contradict the two primary arguments made by KX; (1) that the hiatus marks a period when the ability of radiative forcing to explain surface temperature declines in a statistically meaningful fashion, and that (2) sea surface temperature helps explain this decline. Instead, our results indicate that the control experiment's (HIST) (in)ability to simulate temperature during the hiatus is not different from the rest of the sample period. While these results do not reject the hypothesis that the hiatus can be explained by SSTs, they are not supported by the tests presented in KX but instead are indicative of the increased accuracy of the POGA-H experiment more generally. Instead, our results are consistent with arguments that conditioning on additional observed variables (e.g. SSTs) improves the fit of a dynamic simulation within an open system (e.g. Oreskes et al, 1994, Haavelmo 1940, Hendry and Richard, 1982).

These methods and results are described the following sections 2 to 4. Section 2 describes the methods used to test for changes in the ability of climate models to simulate global surface temperature. Section 3 describes how these methods are used to analyse the experiments simulated by KX (2013), and section 4 concludes.

2 Methodologies to assess differences in model performance: breaks in model errors

Efforts to identify the causes for climate phenomena, such as the hiatus, that compare the accuracy of experiments simulated by climate models are based on two testable hypotheses. First, the accuracy of the model's 'control' experiment declines during the climate phenomena under investigation, such as the hiatus, relative to the rest of the sample period (referred to as hypothesis A herein). This decline is based on the hypothesis that the 'control' experiment either omits the variable/mechanism that causes the phenomena of interest (e.g. the hiatus) or does a poor job of simulating the variable/mechanism that causes the phenomena.

Second, the accuracy of the 'treatment' experiment increases during the hiatus (in a statistically significant fashion) relative to the control experiment (referred to as hypothesis B herein). Hypothesis B is based on the notion that including the variable/mechanism and/or improving the treatment experiment's ability to simulate it will increase its accuracy during the period of interest (e.g. the hiatus) because the influence of this variable/mechanism increases during that period.

Both Hypothesis A and Hypothesis B can be tested using a variety of metrics for model performance. Standard measures of 'goodness of fit' should not be used to compare dynamic simulations generated by open models that use nested sets of conditioning variables, as the fit will improve when conditioning on observed values (Oreskes et al., 1994, Hendry and Richard, 1982). Instead, efforts to distinguish between the performance of control and treatment experiments can focus on breaks in the simulation errors if there is little or no feedback from the variable of interest (global surface temperature) to the conditioning variable (SST). For example, if there is little feedback from the global mean surface temperature anomaly to sea surface temperatures in the Pacific, breaks in the model errors can be informative when one model is conditioned on Pacific SSTs while the other is not. This point is illustrated by a simple simulation in the supplementary material.[2]

Testing hypothesis A and hypothesis B for breaks only at the point of the hiatus ignores the possibility that the control experiment fails across the sample period, and the possibility that the treatment experiment improves the model's explanatory power during periods other than the hiatus (and therefore breaks at the hiatus would be expected). Ignoring these possibilities distorts the size of the test statistic in a way that increases the likelihood of false-positives. These difficulties can be "be avoided if the investigator were required to present a break-date selection algorithm that selects his or her break as a function of all the observations" (Christiano, 1992).

[2] If there is significant feedback, then it is not possible to determine whether the variable explains the hiatus or simply adjusts to the variable of interest.

Formally, performance across the sample period can be assessed using each experiments' simulation error. Let T_t denote observed temperatures at time *t*, and $\hat{T}_{t,\text{Control}}$ and $\hat{T}_{t,\text{Treatment}}$ denote temperatures simulated by a climate model as run under the control and treatment experiments respectively. The experimental or model errors are given by the difference between observed and simulated temperatures:

$$\hat{\epsilon}_{t,j} = \left(T_t - \hat{T}_{t,j}\right) \text{ for } j = (\text{Control, Treatment}) \quad (1)$$

If the treatment experiment adequately simulates the mechanisms that generate the hiatus, the resultant model error $\hat{\epsilon}_{t,\text{Treatment}}$ will be stable throughout the sample period. If there is a unique decrease in the accuracy of the control experiments at the hiatus, the model error $\hat{\epsilon}_{t,\text{Control}}$ will be well-behaved and stable prior to the hiatus (denoted here as $t = H_{\text{Hiatus}}$), and non-stationary[3] thereafter. For example, if an experiment systematically fails to simulate the hiatus, this failure will cause the model error to be stable around zero (or a constant if a bias is present) prior to the hiatus, and show a trend[4] (divergence between model and observations) that starts at the beginning of the hiatus and continues through it.

Formal statistical tests are required to determine whether the model errors change (i.e. break). These tests specify a general functional form to evaluate an array of possible changes. Assuming that the hiatus creates a single break, the change can be represented by the start of a deterministic trend at $(t = H_{\text{Hiatus}})$, and the model error $\hat{\epsilon}_{t,j}$ can be written as:

$$\left(T_t - \hat{T}_{t,j}\right) = \hat{\epsilon}_{t,j} = \mu_{0,j} + \beta_{0,j}t + \beta_{1,j}1(t \geq H_{\text{Hiatus}})(t - H_{\text{Hiatus}}) + u_{t,j} \qquad (2)$$

in which $\mu_{0,j}$ denotes the intercept, $\beta_{0,j}, \beta_{1,j}$ are regression coefficients, and $u_{t,j}$ is a mean-zero additive error term. This statistical model can be augmented with additional dynamics (e.g. an autoregressive term may be necessary to represent dynamic simulations, see Hendry and Richard 1982). Our analysis concentrates on deterministic trends for two reasons. First, the hiatus primarily causes a systematic divergence of the error terms driven by control models diverging from observations. Second, tests for a simple linear trend have power for a wide range of alternatives in which the trends are non-linear or stochastic (Gonzalo and Gadea, 2015).

Hypothesis (A) implies a single breakpoint in the model error that coincides with the hiatus, $H_1 = H_{\text{Hiatus}}$ such that the size of the error increases at H_1. Hypothesis (B) implies the treatment experiment improves the control experiment's ability to fit observations during hiatus. If correct, there is a single breakpoint in the difference in absolute errors: $\left(\left|T_t - \hat{T}_{t,Control}\right| - \left|T_t - \hat{T}_{t,Treatment}\right|\right)$ such that this difference will increase significantly during the hiatus.

Because the statistical methodology should not impose restrictions on the nature and number of breaks, equation (2) can be parameterized to represent more than one break in the trend, intercept, or both as given by equations (3), (4) and (5) respectively:

$$(T_t - \hat{T}_{t,j}) = \mu_{0,j} + \beta_{0,j} \cdot t + \sum_{s=1}^{k} \beta_{s,j} \cdot DT_s + e_{t,j} \quad (3)$$

$$(T_t - \hat{T}_{t,j}) = \mu_{0,j} + \sum_{r=1}^{m} \mu_{r,j} \cdot DU_r + \beta_{0,j} \cdot t + e_{t,j} \quad (4)$$

$$\left(T_t - \hat{T}_{t,j}\right) = \mu_{0,j} + \sum_{r=1}^{m} \mu_{r,j} \cdot DU_r + \beta_{0,j} \cdot t + \sum_{s=1}^{k} \beta_{s,j} \cdot DT_s + e_{t,j} \quad (5)$$

[3] Non-stationary refers to a non-time-invariant joint distribution.

[4] The divergence of model errors consistent with a deterministic trend appears to be a phenomenon across coupled climate models over the time period. For an abrupt failure of a model during the hiatus, a step-shift change may be more appropriate. Section 3 tests for both trend and step-shift changes.

in which $\mu_{0,j}$ denotes the regression intercept and $e_{t,j}$ the regression error term. Indices j = (Control, Treatment) denote the control and treatment experiments. Breaks in the trend are denoted by $DT_s = 1(t > H_s)(t - H_s)$ for up to k breaks where the trend function is joined at the breaks, and for up to m breaks in the intercept are denoted by $DU_r = 1(t > H_r)$. The general models (5, breaks in trend and intercept) and (3, breaks in trend) nest the single-break hiatus (2) as a special case, but do not impose (any) break date to coincide with the hiatus. Model (4, breaks in intercept) is included for completeness if a step-shift happens to be a better approximation to the observed hiatus. By searching the entire sample and using critical values that reflect this sample-wide search, equations (3) – (5) allow the possibility that there may be no break (k=0) or the accuracy may change repeatedly during the sample period, and so there may be up to *k* breaks in the trend[5] at times $t=H_1, \dots, H_k$, and up to *m* breaks in the intercept at $t=H_1, \dots, H_m$.

A large literature describes several methods to detect breakpoints using a sample-wide search. Here we consider three[6] methods with different properties and upper limits on the maximum number of breaks. First, we use a least-squares approach based on Bai (1997) and Bai and Perron (1998) to detect changes in regression coefficients. Perron and Zhu (2005) show that a single break in a trend (*k=1*) can be estimated consistently using this method – the procedure is referred to as least-squares (LS) throughout the remainder of the paper.[7] In the LS approach, the optimal estimator of the coefficients is given by the least squares estimator as a function of the partition time where each partition marks a breakpoint. Then the sum of squares of the residuals is minimized with respect to the partition (break) time to obtain estimates of the break date (see Perron 2006 for a comprehensive overview of the literature on least-squares based break detection).

Second, we use the Perron and Yabu (PY, 2009) test for a single break (*k=1, m=1*) in a linear trend (and/or intercept if specified) which relaxes the assumption that the error term is stationary. If the error term $u_{t,j}$ is incorrectly assumed to be stationary, this will alter the critical values used to determine the presence of a change in a trend function. To avoid such errors, Perron and Yabu (2009) describe a method to detect a change in a linear trend when the noise component $u_{t,j}$ is either stationary or unit-root non-stationary. The test uses information from every possible break partition, and has nearly the same limit distribution regardless of whether the error term is unit-root non-stationary or not. Estrada et al. (2013) use the Perron-Yabu test to argue that that the hiatus is caused in part by a reduction in CFCs (chlorofluorocarbons) following the Montreal protocol.

Third, we use an Indicator Saturation (IS) approach (see Santos et al. 2008 and Johansen and Nielsen 2013, for impulses, Doornik et al 2013 and Castle et al. 2014 for step-shifts, and Pretis et al. 2014 for designed functions e.g. trends) that saturates the model with a full set of break functions and identifies those that are statistically meaningful. The IS approach uses a general-to-specific model selection algorithm through a block search that drops all but significant breaks.[8] This allows the procedure to detect breaks at the beginning or the end of the sample without specifying a minimum break length. As such, the IS approach alleviates weaknesses in both of the previous methods. Both LS and PY require trimming (removing some portion of the sample at the beginning and end) which leads to a

[5] Multiple breaks in a deterministic trend make it possible to assess whether model failure similar to the hiatus have occurred prior to the hiatus. The presence of breaks is not imposed *a-priori*, if multiple breaks are detected, this leads to a piece-wise linear representation where later breaks can off-set earlier ones.

[6] This list is non-exhaustive and additional methods are available (see e.g. Perron 2006 for an overview).

[7] Multiple breaks (*m*) in the intercept alone can be estimated using Bai and Perron (1998) and Bai and Perron (2003). Given that the focus here lies on breaks in the trend, we restrict the analysis to at most one breakpoint in the intercept when using the LS approach. Multiple breaks are covered here using the IS methodology.

[8] False positives are easily controlled in the IS approach. For a sample of T observations, IS for a break in either the trend or intercept implies that T variables are selected over. At the chosen level of significance α one can expect to spuriously retain αT break indicators. When testing for both breaks in the trend and step shifts, we include a full set of step indicators together with our full set of trend functions. To account for the higher number of observations when step shifts and trend changes are allowed, the significance level can be tightened further.

minimum break length. Trimming also makes it impossible to detect breaks near the start or end of the sample[9]. This is important because the hiatus occurs towards the end of the sample period. Equally important, the methods described by Perron and Zhu (2005) and Perron and Yabu (2009) can identify only a single break in the trend. The IS procedure can identify multiple breaks without imposing a ceiling on the number of breaks ($k \leq T, m \leq T$).

All approaches determine the statistical probability of the break using critical values that are adjusted for a sample-wide search (Christiano 1992). The reduction in size (reduction in false positives) through a sample-wide search causes a slight reduction in power. To alleviate concerns about the reduced power of statistical methods that use a sample-wide search, we also apply a "theory-embedding" approach (see Hendry and Johansen 2014) that forces a break at the time of the hiatus (in the year 1998) and simultaneously searches for additional breaks. If the only 'true' break occurs at the hiatus, then no other break will be detected, and the forced break will be statistically significant (as evaluated by a simple t-test).

3 Application: Testing the KX Hypothesis that links the hiatus to La Niña Episodes

3.1 Overview

In this section, we use the three procedures described in section 2 – LS (for up to one break), PY (for up to one break), and IS (for multiple breaks) to assess the experiments run by KX. These procedures are implemented using *Matlab* for LS and *Gauss* for PY.[10] IS is run in *PcGive* (Doornik and Hendry, 2009) using *Ox* (Doornik 2009), step-shift breaks in the intercept using IS also can be run using the *isat* function in the *R* package *gets* (Pretis et al. 2014).

KX run two experiments (ten simulations each) using a coupled climate model (Delworth et al. 2006). The control experiment (HIST) conditions on observed radiative forcing, while the treatment experiment (POGA-H) conditions on observed radiative forcing and observed sea surface temperature anomalies in the equatorial portion of the eastern Pacific. Based on the experiments' performance during the interval of 2002-2012, KX argue that "expanding departures from observed temperatures for the recent decade" generated by the HIST experiment indicate that radiative forcing cannot account for the current hiatus in warming. Instead, the hiatus is "tied specifically to a La Niña like decadal cooling" based on the increased accuracy of the POGA-H experiment.

These conclusions are based on two untested hypotheses:

> A) "Expanding departures from observed temperatures for the recent decade" imply that the HIST experiment's ability to simulate global temperature breaks down concurrent with the hiatus in a way that differs from break-downs during other portions of the sample period.
>
> B) If the hiatus "is tied specifically to a La Niña like decadal cooling", including SST anomalies from the equatorial eastern Pacific will improve the POGA-H experiments' ability to simulate the hiatus relative to the HIST experiment during the hiatus compared to improvements during other portions of the sample period.

3.2 Data

We obtain climate model data on global surface temperature for the HIST and POGA-H experiments from KX. Observed values for global mean surface temperature anomalies relative to 1960-1991 are those used by KX from the HadCrut 4 dataset (Morice et al. 2012). The sample consists of T=64 annual observations from 1949 to 2012 for the HIST experiment and T=63 observations from 1950-2012 for the POGA-H experiment.

3.3 Break Detection

[9] Trimming also limits the maximum number of breaks.
[10] Corresponding code is available on the website of P. Perron: http://people.bu.edu/perron

We evaluate Hypothesis A by testing whether the HIST experiment's ability to simulate global temperature changes during the sample period and if a change occurs, whether the date of change coincides with the beginning of the hiatus, which we define as a broad interval from 1996-2005. This range is consistent with the many possible start dates that are described in the introduction. The existence and timing of changes in accuracy are identified by testing the model errors (throughout the sample period) for a break using the break detection methodologies outlined in section 2: LS[11], PY, and IS. We search for a deterministic break in the trend function (eq. 3), as well as a step-shift in the mean of the error (eq. 4) or both (eq. 5). Of these, equation (3) is the most consistent with the statement that the hiatus is marked by "expanding [model] departures from observed temperatures (KX)." Our analysis concentrates on the presence of breaks rather than their magnitudes (as captured by the regression coefficients) because the presence of a break implies a significant coefficient, and the magnitude of the change provides limited additional insight into the causal mechanism.

All procedures are applied to the ensemble mean of the ten control simulations (HIST) and the ensemble mean of the ten treatment simulations (POGA-H). The results for individual simulations are reported in the supplementary material. To facilitate a direct comparison with KX, the supplementary material also reports non-parametric Mann-Kendall tests estimated using a rolling window to illustrate the distortion caused by not conducting a sample-wide search with adjusted critical values.[12]

3.4 Testing Hypothesis A: Breakdown at the Hiatus and Uniqueness of Breaks

The ability of the HIST experiment to simulate temperature changes throughout the sample period – see Table 1 and Figure 1. Specifically, the HIST experiment under-predicts temperature during the early portion of the sample period and over-predicts temperature during the latter portion of the sample, much of which coincides with the hiatus. As such, the HIST experiment fails during the hiatus, but also fails before the hiatus. These sample-wide failures are inconsistent with the claim that the hiatus marks a unique failure by HIST experiment.

This visual conclusion is supported by the results generated by the statistical methodology. Results generated by the LS procedure indicate that none of the breaks in the HIST ensemble mean coincide with the hiatus (Figure 1) but rather occur during the 1970s.

If uncertainty about the time series properties of the error term is allowed, the PY test does not identify a significant break consistent with the hiatus. While exhibiting lower power relative to an optimal test in which the time series properties of the model error are known (PY 2009), this result indicates the possible presence of a unit root in the model error does not affect our conclusion that the HIST simulations do not fail in a systematic way uniquely at the start of the hiatus.

LS and PY are designed to find only a single breakpoint in a trend. As such, their use implicitly assumes that a failure caused by the hiatus is unique and there are no additional breaks associated with 'other changes' during the sample period. We relax this assumption and allow for multiple regimes (observations between breaks of any length) using the IS procedure. The IS-type algorithm detects a break during the hiatus, however it is not unique - the vast majority of breaks identified occur during the 1970s (Figure 1).

[11] For the least squares procedure the model with or without breakpoint is determined using the Bayesian Information Criterion (BIC). The trimming factor is set to 0.1, which corresponds to a minimum break length of 6 years.

[12] The Mann-Kendall (Mann 1945, Kendall 1976) test is a non-parametric test for the presence of a monotonic trend in the series, where under the null-hypothesis no monotonic trend is present.

***Figure 1**: Break Dates in HIST ensemble mean model error. Breaks in trends are shown as dotted vertical lines, breaks in trend and the mean as solid vertical lines, breaks in the mean alone as long-dash vertical lines. Green vertical lines indicate the Least-Squares (LS) results, orange vertical lines for Perron and Yabu (2009), and purple vertical lines for IS. The start of the hiatus interval (1996-2005) is shaded grey. Note that no break is detected when using the Perron-Yabu (2009) test and therefore no break is shown in orange.*

***Table 1**: Hypthesis A: Breaks in the HIST Experiment Ensemble Mean Model Error. Break Dates in Trend and both Trend and Mean of HIST Model Error using Least-Squares, Perron and Yabu (2009), for a maximum of one break, and IS for multiple breaks. Panel (i) reports the results of a sample-wide break search with no forced break at the hiatus. Panel (ii) reports the additionally detected breaks using IS in the "theory-embedded" approach with a forced hiatus break in the trend (in 1998).*

i) Method	**Breaks in Trend**	**Breaks in Constant**	**Breaks in Trend & Const.**
Least-Squares	1956	1960	1991
Perron-Yabu	-	-	-
IS	1970, 1975, 1993	1970, 1998	1973, 1975, 1993
ii) Forced break in 1998			
IS: forced break in 1998	1970, 1972	1960, 1969	1970, 1972
1998 forced break t- and p-values:	t=-4.53, p<0.001	t=-4.80, p<0.001	t=-4.53, p<0.001

Results for Least-Squares: breakpoint selected using the BIC, - denotes no break chosen. Results for Perron-Yabu Procedure: Significance of break indicated by: ** significant at 1%, * significant at 5%, - no significant break. Results for IS: The test procedure for a change in the trend is run at 0.5% significance level to avoid over-fitting. Two breaks within a short time period (2-3 years) of each other likely identify outlying observations. For a sample of T observations, at the chosen level of significance one can expect to spuriously retain 0.005T break indicators. In the present case, the main sample contains 64 observations, using a IS type test we expect around 0.3 breaks to be spuriously retained. When testing for both breaks in the trend and step shifts, we include a full set of step indicators together with our full set of trend functions. To account for the higher number of observations when step shifts and trend changes are allowed, the significance level is tightened to 0.1%. Results are obtained using the PcGive software package (Doornik and Hendry, 2009).

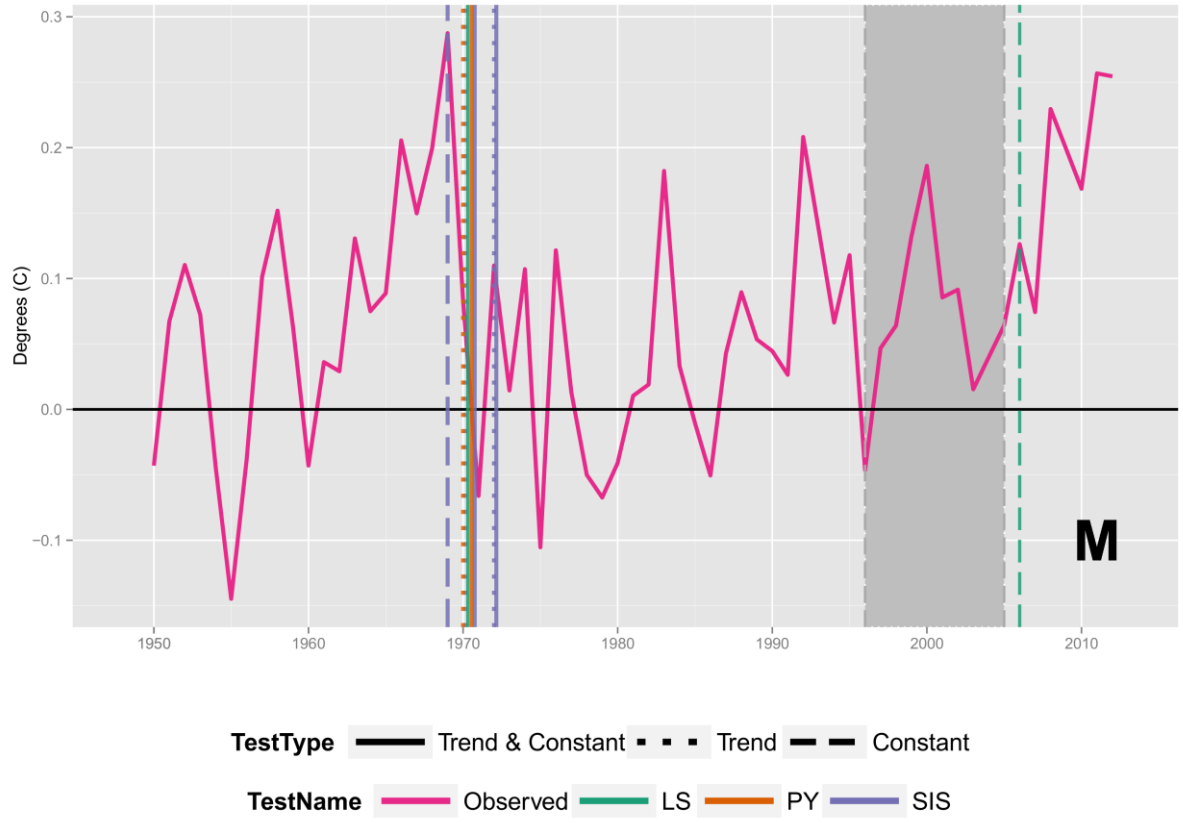
Tests based on the embedding approach indicate that imposing a break in trend in 1998 is statistically significant, but this break is not unique because a sample-wide search confirms that additional breaks are detected in the 1970s. This result reinforces those described earlier; the accuracy of the HIST experiment does not decline *uniquely* during the hiatus.

Together, these results suggest that there is little evidence that the ability of observed radiative forcing (HIST experiments) to simulate temperature fails uniquely at the hiatus. The HIST experiment fails (breaks are detected) throughout the sample period, particularly during the 1960s and 1970s. As such, we reject hypothesis A.

3.5 Testing Hypothesis B: Improvement in Relative Performance and Uniqueness of Improvement

We evaluate hypothesis B by using the same statistical methodology (LS, PY, IS) to assess whether the ability of the POGA-H experiment to simulate surface temperature improves compared to the HIST experiment during the hiatus relative to the rest of the sample.

***Figure 2**: Break Dates in the ensemble mean difference in absolute errors. Breaks in trends are shown as dotted vertical lines, breaks in trend and the mean as solid vertical lines, breaks in the mean alone as long-dash vertical lines. Green vertical lines indicate Least-Squares (LS) results, orange vertical lines for Perron and Yabu (2009), and purple vertical lines for IS. The start of the hiatus interval (1996-2005) is shaded grey.*



Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

***Table 2**: Hypthesis B: Breaks in the difference of absolute error term* $(|Observed_t - HIST_t| - |Observed_t - POGAH_t|)$ *Break dates in the difference in the ensemble mean absolute errors of the HIST and POGA-H experiments using Least-Squares, Perron and Yabu (2009) and IS.*

Method	Breaks in Trend	Breaks in Constant	Breaks in Trend & Const.
Least-Squares	-	2006	1970
Perron-Yabu	1970*	-	1970*
IS	1970, 1972	1969	1970, 1972

Significance levels used for PY and IS break detection are identical to those reported in Table 1.

A visual examination of Figure 2 indicates that the difference between the errors from the HIST and POGA-H experiments generally are positive. This suggests that the POGA-H experiment is more accurate than the POGA-H experiment. For the 1950-2012 period, the R^2 for the POGA-H experiment of 0.93 is greater than the 0.82 for the control experiment. This 0.11 difference grows to 0.25 during the pre-hiatus subsample 1950-1998; 0.85 and 0.60 for the POGA-H and HIST experiments respectively. Together, these results suggest that conditioning on SST's increases the model's accuracy throughout the sample period.

Consistent with this result, the accuracy of the POGA-H experiment increases (breaks are detected) relative to the control experiment well before the hiatus (Figure 2 and Table 2). Their timing indicates that including SST's in the POGA-H experiment does not increase the climate model's accuracy *uniquely* during the hiatus relative to the rest of the sample period. This conclusion is consistent across the methods: breaks occur during the 1960s and 1970s and not the hiatus period.

These results indicate that adding the observed values of SST from the eastern Pacific improves the POGA-H experiment's ability to simulate temperature throughout the sample, not just during the hiatus. As such, this conclusion does not support the claim that the increased accuracy of the POGA-H experiment ties the hiatus *specifically* to a La Niña like cooling. Instead, conditioning on observed forcing and observed SST improves the model throughout the sample period.

Changes in the accuracy of the HIST experiments throughout the sample period rejects the notion that the hiatus marks a unique breakdown in the relation between anthropogenic forcing and global temperature. Similarly, the enhanced accuracy of the POGA-H experiment relative to the HIST experiment throughout the sample indicates that conditioning an open model on additional information increases the model's ability to simulate observations. While this does not rule out that SSTs contribute to the hiatus, it is not informative– the improved fit is likely a result of the conditioning on additional observed values in a dynamic simulation. Together, these results indicate that the hiatus is not statistically different from the rest of the 1950-2012 temperature record, when temperature can be modelled by anthropogenic forcings and natural variability such as solar insolation.

4 Conclusions

Testing experiments run by climate models to identify causes of climatic phenomena, such as the hiatus in observed temperature requires that analysts be cognizant of the limits of analysing simulations generated by open models and use these limits to choose their statistical methodology. By definition, conditioning open models on additional information will increase a model's ability to simulate observed values. As such, the ability of open models to identify determinants of shifts in the historic record is challenging. Analysts taking this approach cannot limit their analysis of model performance to the period of interest (e.g. the hiatus); they must search for a change in performance over the entire sample period. Furthermore, because the timing of climate phenomena often are not known *a priori*, analysts must search over the entire sample to avoid size distortions caused by choosing subsamples *a-priori*. Such efforts can be supported using a variety of methods, the choice of

which depends on the time series properties of the data. We illustrate these principles by revisiting the conclusion reported by KX, that the hiatus is tied specifically to a La Nina like cooling. The support for this conclusion disappears when we consider conditioning on observed values and a sample-wide search for breaks. More broadly we hope to highlight statistical tools available to assess models for systematic change at any point in the sample, which can be informative when considering simulated models with varied mechanisms or conditioning on observations.

Acknowledgements

Financial support from the Open Society Foundations and the Oxford Martin School is gratefully acknowledged. We are thankful to David F. Hendry, Max Roser, and Andrew Martinez for helpful comments on an earlier version. We thank Yu Kosaka and Shang-Ping Xie for providing model temperature data.

Literature Cited

1. Bai, J., & Perron, P. 1998. "Estimating and testing linear models with multiple structural changes." *Econometrica*, 47-78.
2. Bai, J. and P. Perron, 2003. "Computation and analysis of multiple structural change models", *Journal of Applied Econometrics*, 18(1):1-22.
3. Banerjee, Anindya, Robin L. Lumsdaine, and James H. Stock. 1992. Recursive and sequential tests of the unit-root and trend-break hypotheses: theory and international evidence. *Journal of Business & Economic Statistics* 10, no. 3: 271-287.
4. Busetti, Fabio, and A. M. Taylor. 2004. Tests of stationarity against a change in persistence. *Journal of Econometrics* 123, no. 1: 33-66.
5. Christiano, L.J, 1992. Searching for a break in GNP, Journal of Business & Economic Statistics 10(3):237-250.
6. Delworth, T. L. et al. 2006. "GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics." *Journal of Climate*. 19, 643–674.
7. Doornik, J. A. 2009. "An Object-Oriented Matrix Programming Language Ox 6".
8. Doornik, J.A., Hendry, DF and Pretis, F, 2013. "Step-Indicator Saturation", *University of Oxford Department of Economics Discussion Paper*, 584.
9. Castle, J.L, Doornik, J.A. Hendry, DF and Pretis, F, 2014. "Detecting Location Shifts During Model Selection by Step-Indicator Saturation". *University of Oxford Department of Economics Discussion Paper. Under Review.*
10. Doornik, JA, Hendry, DF, 2009. *Empirical Econometric Modelling using PcGive: Volume I*. London: Timberlake Consultants Press.
11. Estrada F, Perron, P, and Martínez-López, B. 2013. "Statistically derived contribtuions of diverse human influences to twentieth-century temperature changes", *Nature Geoscience*, doi: 10.1038/ngeo1999.
12. Gonzalo, J, Gadea, L. 2015. "Trend or no Trend in Distributional Characteristics: Existence of Global Warming". *Universidad Carlos III de Madrid Working Paper*.
13. Haavelmo, T. 1940. "The inadequacy of testing dynamic theory by comparing theoretical solutions and observed cycles." *Econometrica*, 312-321.
14. Hassler, Uwe, and Jan Scheithauer, 2011. Detecting changes from short to long memory. *Statistical Papers* 52, no. 4: 847-870.
15. Hendry, D.F. and S. Johansen 2014. "Model discovery and Trygve Haavelmo's legacy". *Econometric Theory*, doi:10.1017/S0266466614000218.
16. Hendry, D. F., & Richard, J. F. 1982. "On the formulation of empirical models in dynamic econometrics." *Journal of Econometrics*, *20*(1), 3-33.
17. Johansen, S., & Nielsen, B. 2013. "Outlier detection in regression using an iterated one-step approximation to the Huber-skip estimator." *Econometrics*, *1*(1), 53-70.
18. Kaufmann, RK, H. Kauppi, ML Mann, and J.H Stock, 2011. Reconciling anthropogenic climate change with observed temperature 1998-2008, *Proceedings National Academy of Sciences* 108(29):11790-11793 doi/10.073/pnas.1102467108.
19. Kendall, M. G. 1976. "Rank Correlation Methods". 4th ed. Griffin.
20. Kim, Jae-Young, 2000, Detection of change in persistence of a linear time series. *Journal of Econometrics* 95, no. 1: 97-116.

21. Kosaka Y. and S.P. Xie, 2013. Recent global-warming hiatus tied to equatorial Pacific surface cooling, *Nature* 501:403-407.
22. Mann, H. B. (1945). Nonparametric tests against trend. Econometrica: Journal of the Econometric Society, 245-259.
23. Meehl, Gerald A., Julie M. Arblaster, John T. Fasullo, Aixue Hu, and Kevin E. Trenberth., 2011, Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods. *Nature Climate Change* 1, no. 7: 360-364.
24. Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. 2012. "Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set." *Journal of Geophysical Research: Atmospheres (1984–2012)*, *117*(D8).
25. Oreskes, N., Shrader-Frechette, K., & Belitz, K. 1994. "Verification, validation, and confirmation of numerical models in the earth sciences." *Science*, *263*(5147), 641-646.
26. Perron, P. and T. Yabu, 2009, Testing for Shifts in Trend with an Integrated or Stationary Noise Component, *Journal of Business and Economic Statistics*, 27, 369-396.
27. Perron, Pierre, 2006, Dealing with structural breaks. *Palgrave handbook of Econometrics* 1: 278-352.
28. Perron, P., & Zhu, X. 2005. Structural breaks with deterministic and stochastic trends. *Journal of Econometrics*, *129*(1), 65-119.
29. Pretis, F., Schneider, L. & Smerdon, J.E. 2014. Detecting breaks by designed functions applied to volcanic impacts on hemispheric surface temperatures". *University of Oxford Department of Economics Discussion Paper*
30. Santer, Benjamin D., Céline Bonfils, Jeffrey F. Painter, Mark D. Zelinka, Carl Mears, Susan Solomon, Gavin A. Schmidt et al., 2014, Volcanic contribution to decadal changes in tropospheric temperature. *Nature Geoscience* 7, no. 3: 185-189.
31. Santos, C., Hendry, D. F., & Johansen, S. 2008. "Automatic selection of indicators in a fully saturated regression." *Computational Statistics*, *23*(2), 317-335.
32. Pretis, F, Reade, J. and Sucarrat, G.(2014). "*gets*. An R package for testing for general to specific model selection and indicator saturation."
33. Zivot, Eric, and Donald W. K. Andrews., 1992. "Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis." *Journal of Business & Economic Statistics* 20, no. 1: 25-44.

Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

Supplementary Material

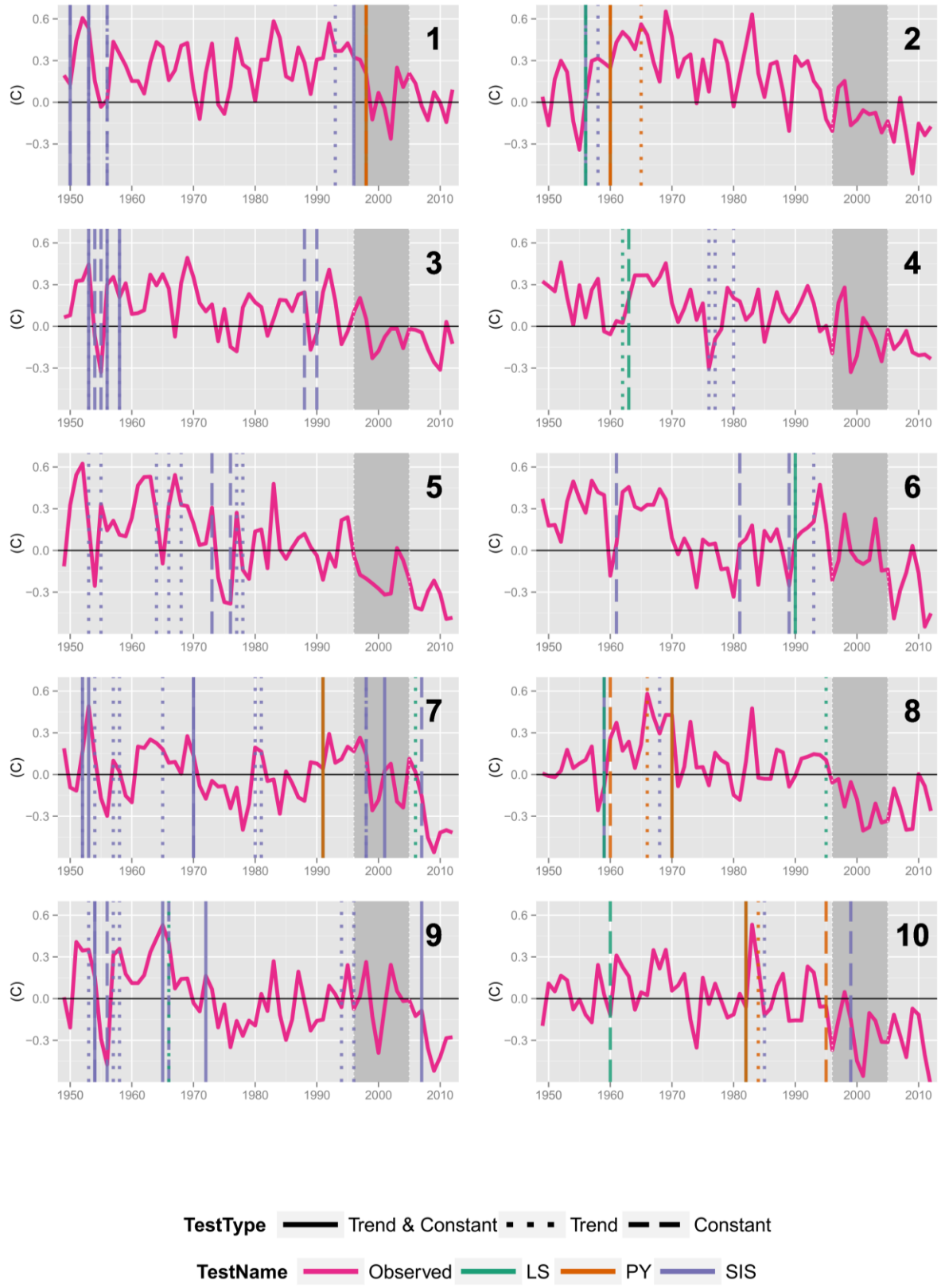
S1: Assessing breaks in individual members of each experiment.

The analysis in the main body of the text concentrates on the ensemble means of treatment and control simulations. As an exploratory analysis we also provide here the break detection results when considering each model run individually. While this is non-standard, the reason that supports an individual analysis is that the model outcomes are driven by varied initial conditions in systems that are non-stationary (and thus exhibit long-memory). Starting values can therefore affect the outcome dramatically and averaging across the different runs may therefore not be appropriate.

For hypothesis A break dates are reported for each model run from 1 to 10. For hypothesis B, each pairing of HIST and POGA-H is assessed through the difference in absolute error terms, this leads to a total of 10x10=100 pairs.

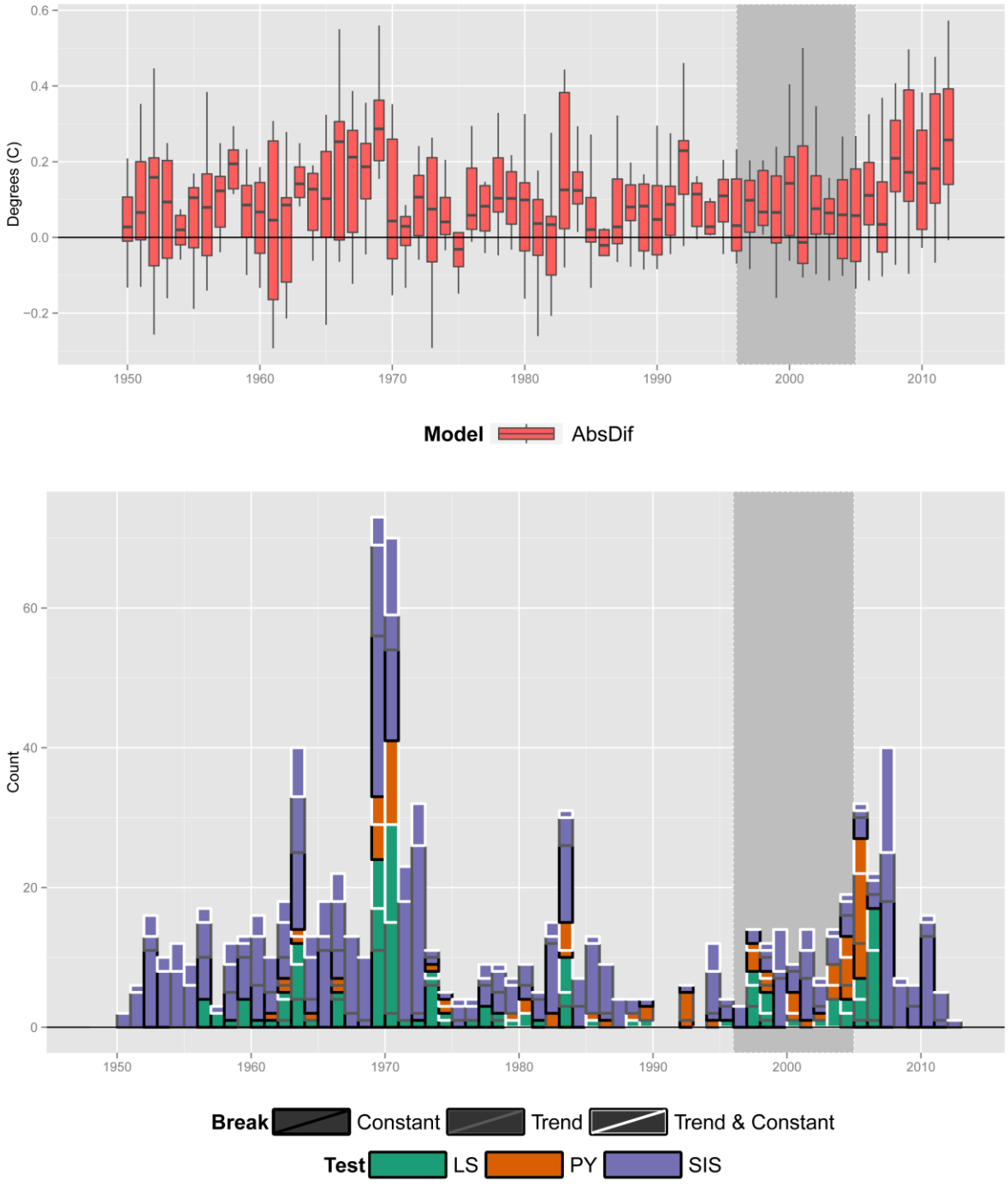
Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

Figure S1: *Break Dates in HIST model errors from experiments 1-10. Tests on the individual HIST experiment simulations (panels 1-10). Breaks in trends are shown as dotted vertical lines, breaks in trend and the mean as solid vertical lines, breaks in the mean alone as long-dash vertical lines. Green vertical lines indicate the Least-Squares results, orange vertical lines for Perron and Yabu, and purple vertical lines for IS. The start of the hiatus interval (1996-2005) is shaded grey. If the model errors exhibit no break, then no break lines are shown.*



Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

Figure S2: *Histogram of Break Dates in the Difference in Absolute Errors for each of the 100 pairings of the 10 HIST and 10 POGA-H experiments. Top panel shows the difference in absolute errors. Bottom panel shows the count of statistically significant changes in the difference in absolute errors when tested for changes in the trend (grey border), constant (black border), and trend and constant (white border) for all three break detection methods: Least-squares (green), Perron and Yabu (orange) and IS (purple). The start of the hiatus interval (1996-2005) is shaded grey.*



Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

Hypothesis A: Breaks in the HIST Experiment Model Errors 1-10

Table S1: Break Dates in Trend and both Trend and Mean of HIST Model Error using Least-Squares for a maximum of one break.

Experiment	**Breaks in Trend**	**Breaks in Constant**	**Breaks in Trend and Constant**
HIST 1	1998	1998	1998
HIST 2	1956	1956	1960
HIST 3	-	-	-
HIST 4	1962	1963	-
HIST 5	-	-	-
HIST 6	-	-	1990
HIST 7	2006	-	1991
HIST 8	1995	1959	1970
HIST 9	1966	-	-
HIST 10	-	1960	1982

Table S2: Perron-Yabu test for a breaks in HIST model errors for a maximum of one break.

Experiment	**Breaks in Trend**	**Breaks in Constant**	**Breaks in Constant and Trend**
Hist1	-	1998**	1998*
Hist2	1965**	1960**	1960**
Hist3	-	-	-
Hist4	-	-	-
Hist5	-	-	-
Hist6	-	-	-
Hist7	-	-	1991**
Hist8	1966**	1960**	1970**
Hist9	-	-	-
Hist10	1984*	1995*	1982*

Significance of break indicated by: ** significant at 1%, * significant at 5%, - no significant break.

Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

Table S3: Break Dates in Trend and both Trend and Mean of HIST Model Errors using IS.

Experiment	**Breaks in Trend**	**Breaks in Constant**	**Breaks in Trend and Constant**
HIST 1	1953, 1956, 1993	1950, 1953, 1956, 1998	1950, 1953, 1996
HIST 2	1956, 1958	1956	1956
HIST 3	1953, 1956, 1958	1954, 1955, 1988, 1990	1953, 1956, 1958
HIST 4	1976, 1977, 1980	-	-
HIST 5	1953, 1955, 1964, 1966, 1968, 1977, 1978	1973, 1976	-
HIST 6	1990, 1993	1961, 1981, 1989	-
HIST 7	1952, 1954, 1957, 1958, 1965, 1980, 1981, 1998	1970, 1998, 2007	1952, 1953, 1970, 2001
HIST 8	1968	1959	1959
HIST 9	1953, 1957, 1958, 1994, 1996	1954, 1956, 1966	1965, 1972, 1954, 2007
HIST 10	1985	1999	-

The test procedure for a change in the trend is run at 0.5% significance level to avoid over-fitting. Two breaks within a short time period (2-3 years) of each other likely identify outlying observations. For a sample of T observations, at the chosen level of significance one can expect to spuriously retain 0.005T break indicators. In the present case, the main sample contains 64 observations, using a SIS type test we expect around 0.3 breaks to be spuriously retained. When testing for both breaks in the trend and step shifts, we include a full set of step indicators together with our full set of trend functions. To account for the higher number of observations when step shifts and trend changes are allowed, the significance level is tightened to 0.1%. Results are obtained using the PcGive software package (Doornik and Hendry, 2009).

Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

Hypthesis B: Breaks in the difference of absolute error term $(|Observed_t - HIST_t| - |Observed_t - POGAH_t|)$ *for every pair of HIST and POGA-H experiments.*

Table S4: Break dates falling within and outside the Hiatus in the difference in the absolute errors of the HIST and POGA-H experiments using least-squares based on Least-squares, Perron and Yabu, and IS for all 100 pairings of HIST and POGA-H experiments.

Frequency	**Breaks in Trend**			**Breaks in Constant**			**Breaks in Trend and Constant**		
Method	*LS*	*PY*	*IS*	*LS*	*PY*	*IS*	*LS*	*PY*	*IS*
No Break	45	67	24	44	76	22	39	57	44
Break During Hiatus	9	20	11	5	11	23	16	16	12
Break Outside Hiatus	46	13	65	51	13	55	45	27	44

Significance levels used for SIS break detection are identical to those reported in Table 3. Significance of the breaks using Perron and Yabu (2009) is assessed at 5%, denoted by * for breaks in the ensemble mean.

Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

S2: Simulation example assessing conditioning in dynamic simulations

We briefly consider an artificial example of using a dynamic simulation to differentiate between whether a variable *Z* causes a hiatus in a variable *Y*. This example is based on the results on dynamic simulations in Hendry and Richard (1982). The example illustrates that breaks can be informative in the absence of feedbacks.

Consider a simple data generating process (DGP) as:

$$y_t = \alpha z_t + \beta y_{t-1} + \theta_y \iota_t + \epsilon_t$$

$$z_t = \lambda z_{t-1} + \gamma y_{t-1} + \theta_z \iota_t + v_t$$

where (ϵ_t, v_t) are iid normal, mean zero error terms with variances $(\sigma_\epsilon^2, \sigma_v^2)$ and covariance of zero. The term ι_t denotes a "hiatus" break in the form of an on-setting trend, such that $\iota_t = 0$ for $t < H_{Hiatus}$, and $\iota_t = (t - H_{Hiatus})$ for $t \geq H_{Hiatus}$onwards.

Assume there is a single hiatus at $t = H_{Hiatus}$ and consider two cases:

1) z_t causes the hiatus in y_t : $\theta_z \neq 0$ and $\theta_y = 0$

2) z_t does not cause the hiatus in y_t: $\theta_z = 0$ and $\theta_y \neq 0$

There are two ways to compute the simulation path of our variable of interest y_t. We can condition on z_t taking the observations as given (a dynamic simulation in an open model), or alternatively simulate both z_t and y_t jointly (a dynamic simulation in a closed model).

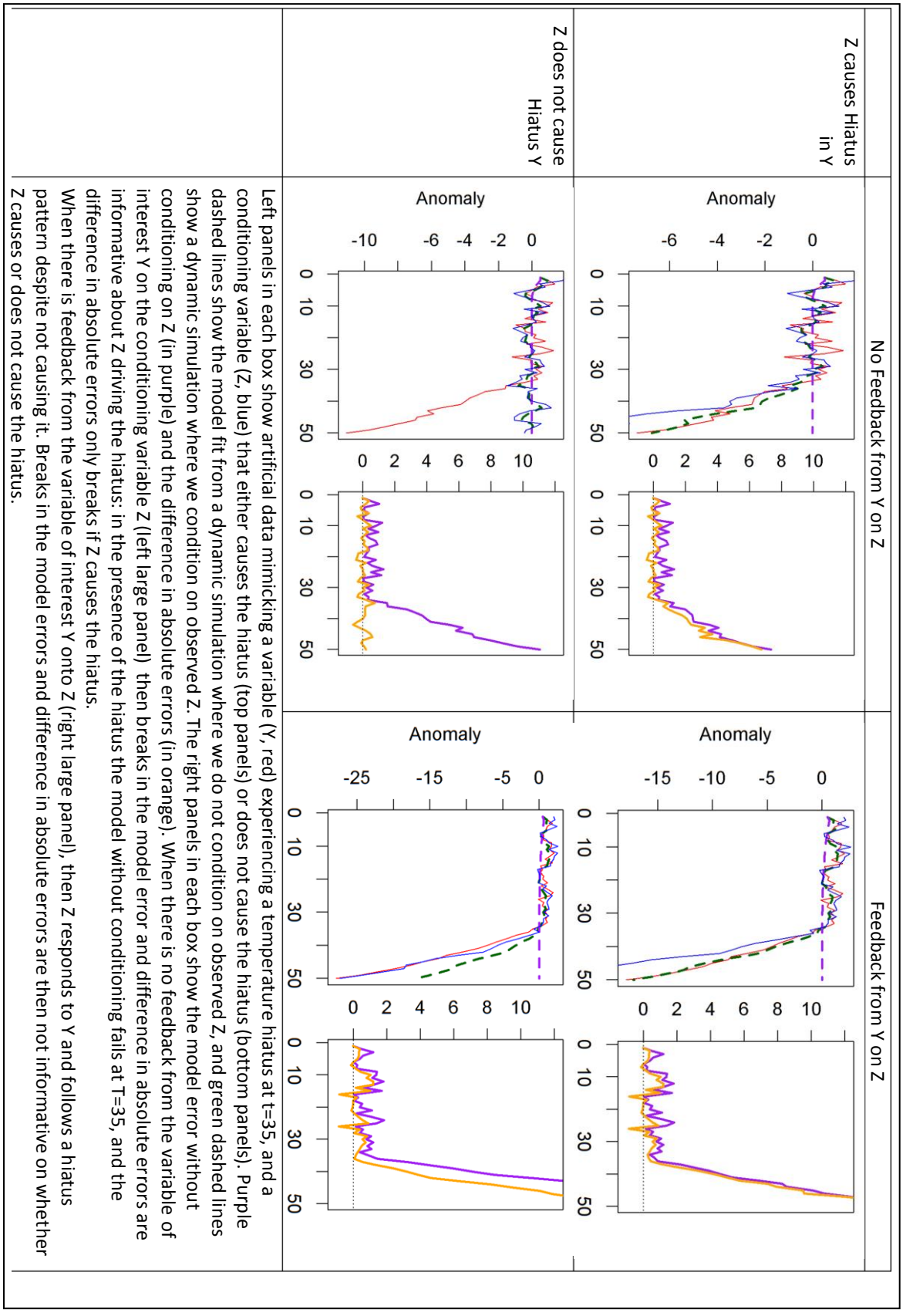
In a simple example in Figure S3 we illustrate dynamic simulations for y_t in a closed and open system for four different scenarios. These four scenarios are: whether z_t causes the hiatus in y_t or not; and whether or not there are feedbacks from y_{t-1} onto z_t: feedbacks ($\gamma \neq 0$) vs. no feedbacks ($\gamma = 0$) (i.e. z_t is strongly exogenous). The dynamic simulations are conducted under the assumption that the true parameter values $(\alpha, \beta, \lambda, \gamma)$ are known, while the hiatus itself is not.

First, conditioning on observed future values improves the fit - the dynamic simulation in an open model yields an overall closer fit than a dynamic simulation in a closed model, this intuitive result is also highlighted in Hendry and Richard (1982). Therefore comparing the fit alone is not indicative of the model structure. Second, when there are no feedbacks ($\gamma = 0$) of y_{t-1} onto z_t, then by looking at the model error of the closed simulation, and the difference in absolute errors from the open and closed simulation, one can distinguish between whether z_t causes the hiatus in y_t or not. If z_t causes the hiatus in y_t, then the model error in the closed system breaks at the time of the hiatus, as does the difference in absolute error terms. If z_t does not cause the hiatus in y_t, then the model error in the closed system breaks at the hiatus, however, the difference in absolute error terms does not.

However, if there are feedbacks from of y_{t-1} onto z_t, then one cannot differentiate between z_t causing the hiatus in y_t or whether the hiatus is driven by some other factor outside of the model. Because there are feedbacks, a hiatus driven by some other factor affects z_t through y_{t-1}. Then conditioning on z_t improves the dynamic simulation fit for y_t during the hiatus even though z_t does not cause it. This is a result of conditioning on observed future values.

Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

Figure S3: Dynamic Simulation Example using artificial data where a variable of interest *Y* (red) experiences a hiatus at *t=35*, either caused by *Z* (blue) in the top panels or not caused by *Z* (bottom panels).



Published as: Pretis, Mann, and Kaufmann, 2015, *Climatic Change* 131:4. The final publication is available at Springer via http://link.springer.com/article/10.1007%2Fs10584-015-1391-5

S3: Non-Parametric Mann-Kenndall Tests

To facilitate a comparison to KX who use the Mann-Kendall test at the time of the hiatus, we also report Mann-Kendall test results. Pre-selecting a break-point around the time of the hiatus suggests an on-setting trend, however, this is occurs in absence of a sample-wide search or adjustment of the critical values. Equally a different time period could be pre-selected as we illustrate using a 10-year rolling-window. The Mann-Kendall test then finds similar changes during the 1960s, 1970s, and early 1990s (Figure S4). This approach is shown here as an illustration only given that this is not a correct sample-wide search and suffers from the multiple testing problem. The same results also apply to using the Mann-Kendall test on the difference in absolute error terms (Figure S5).

Figure S4: *Ten year rolling-window Mann-Kendall Test for a trend in HIST experiment ensemble mean error. Top panel shows HIST ensemble mean error (observed-ensemble mean). Bottom panel shows the p-value (orange) of the 10-year Mann-Kendall test. P-values below 0.05 (dashed black) reject the null hypothesis of no change in trend over the next 10 year window (horizontal red bars indicate the interval tested). The start of the hiatus interval (1996-2005) is shaded grey.*

Figure S5: *Ten year rolling-window Mann-Kendall Test for a trend in ensemble mean difference in absolute errors. Top panel shows HIST ensemble mean error (observed-ensemble mean). Bottom panel shows the p-value (orange) of the 10-year Mann-Kendall test. P-values below 0.05 (dashed black) reject the null hypothesis of no changing trend over the next 10 year window (horizontal red bars indicate the interval tested). The start of the hiatus interval (1996-2005) is shown as grey-shaded.*