

Comments on “Detection of Undocumented Changepts: A Revision of the Two-Phase Regression Model”

XIAOLAN L. WANG

Climate Research Branch, Meteorological Service of Canada, Downsview, Ontario, Canada

28 January 2003 and 30 April 2003

Lund and Reeves (2002) revisited the two-phase linear regression test for changepoint detection at an undocumented time. The two-phase regression model is of the following form:

$$X_t = \begin{cases} \mu_1 + \alpha_1 t + \epsilon_t, & 1 \leq t \leq c \\ \mu_2 + \alpha_2 t + \epsilon_t, & c < t \leq n, \end{cases} \quad (1)$$

where $\{\epsilon_t\}$ is the zero-mean independent random error with a constant variance σ_ϵ^2 . This allows for both step- and trend-type changepoints. The time c is called a changepoint if $\mu_1 \neq \mu_2$ (step type) and/or $\alpha_1 \neq \alpha_2$ (trend type).

The F statistic for a changepoint at time $c \in \{2, \dots, n-1\}$ is

$$F_c = \frac{(\text{SSE}_{\text{Red}} - \text{SSE}_{\text{Full}})/2}{\text{SSE}_{\text{Full}}/(n-4)}, \quad (2)$$

where SSE_{Full} is the “full model” sum of squared errors:

$$\begin{aligned} \text{SSE}_{\text{Full}} = & \sum_{t=1}^c (X_t - \hat{\mu}_1 - \hat{\alpha}_1 t)^2 \\ & + \sum_{t=c+1}^n (X_t - \hat{\mu}_2 - \hat{\alpha}_2 t)^2 \end{aligned} \quad (3)$$

and SSE_{Red} is the “reduced model” sum of squared errors:

$$\text{SSE}_{\text{Red}} = \sum_{t=1}^n (X_t - \hat{\mu}_{\text{Red}} - \hat{\alpha}_{\text{Red}} t)^2, \quad (4)$$

where $\hat{\mu}_{\text{Red}}$ and $\hat{\alpha}_{\text{Red}}$ are estimated under the constraints $\mu_1 = \mu_2 = \mu_{\text{Red}}$ and $\alpha_1 = \alpha_2 = \alpha_{\text{Red}}$. Under the null hypothesis of no changepoint and assuming Gaussian errors ϵ_t , F_c has an F distribution of 2 numerator degrees

of freedom and $(n-4)$ denominator degrees of freedom, denoted as $F_{(2, n-4)}$. The existence of an undocumented changepoint is concluded when

$$F_{\text{max}} = \max_{1 \leq c \leq n} F_c \quad (5)$$

is too large to be attributed to chance variation. The most prominent changepoint is estimated as the argument(s) c that maximizes F_c . Lund and Reeves (2002) pointed out errors that had propagated in the related literature and presented the true percentiles of the F_{max} distribution under the null hypothesis of no changepoint, which is very useful for detecting step- and trend-type changepoints.

However, as mentioned in Lund and Reeves (2002), changes in trend slopes could be rooted in true climate change. Also, a trend-type changepoint could well be just a point between two phases of long quasi-periodic variation (e.g., multidecadal fluctuations). Thus, extra caution should be exercised when using the above changepoint detection technique, especially when a trend-type changepoint is involved. In particular, Eq. (4.2) in Lund and Reeves (2002), which shows how to adjust the time series for the detected changepoint, should be used with caution. If the ultimate goal of the changepoint detection is to form homogeneous climate data series by correcting biases, a trend-type change in climate data series should only be adjusted if there is sufficient evidence showing that it is related to a change at the observing station, such as a change in the exposure or location of the station, or in its instrumentation or observing procedures.

In addition, a simpler situation is most often encountered when correcting biases in climate observations but was not considered explicitly in Lund and Reeves (2002). That is a case when $\alpha_1 = \alpha_2 = \alpha$; that is, a two-phase regression model with a common trend α :

$$X_t = \begin{cases} \mu_1 + \alpha t + \epsilon_t, & 1 \leq t \leq c \\ \mu_2 + \alpha t + \epsilon_t, & c < t \leq n. \end{cases} \quad (6)$$

Corresponding author address: Dr. Xiaolan L. Wang, Climate Research Branch, Meteorological Service of Canada, 4905 Dufferin Street, Downsview, ON M3H 5T4, Canada.
E-mail: xiaolan.wang@ec.gc.ca

TABLE 1. The F_{\max} percentiles of model (6).

n	F_{\max} percentiles		
	$F_{\max, 0.90}$	$F_{\max, 0.95}$	$F_{\max, 0.99}$
10	11.646	15.559	28.412
20	9.651	11.948	18.043
30	9.379	11.396	16.249
40	9.261	11.148	15.750
50	9.269	11.068	15.237
60	9.296	11.072	15.252
70	9.296	11.059	14.985
80	9.341	11.072	15.013
90	9.397	11.080	14.891
100	9.398	11.085	14.874
150	9.506	11.127	14.828
200	9.604	11.208	14.898
250	9.691	11.310	14.975
300	9.790	11.406	14.998
350	9.794	11.392	15.044
400	9.840	11.416	14.980
450	9.872	11.474	15.072
500	9.929	11.537	15.115
550	9.955	11.552	15.086
600	9.995	11.549	15.164
800	10.102	11.673	15.292
1000	10.169	11.749	15.154
2500	10.478	12.064	15.519

TABLE 2. The detection rates (in 1/200) of models (6) and (1) for time series of step-type changepoints (at 5% level of significance).

n	Detection rate (in 1/200)	
	Model (6)	Model (1)
50	37	31
70	58	49
90	76	70
100	71	61
200	111	103
300	123	120
400	126	123
500	136	132
600	133	129
800	140	135
1000	142	137
2500	152	151

In this case, the F statistic for the changepoint $c \in \{2, \dots, n - 1\}$ is

$$F_c = \frac{(\text{SSE}_{\text{Red}} - \text{SSE}_{\text{Full}})/1}{\text{SSE}_{\text{Full}}/(n - 3)}, \tag{7}$$

which is distributed as $F_{(1, n-3)}$ under the null hypothesis of no changepoint and assuming Gaussian errors ϵ_t . Here,

$$\begin{aligned} \text{SSE}_{\text{Full}} = & \sum_{t=1}^c (X_t - \hat{\mu}_1 - \hat{\alpha}t)^2 \\ & + \sum_{t=c+1}^n (X_t - \hat{\mu}_2 - \hat{\alpha}t)^2 \quad \text{and} \end{aligned} \tag{8}$$

$$\text{SSE}_{\text{Red}} = \sum_{t=1}^n (X_t - \hat{\mu}_{\text{Red}} - \hat{\alpha}_{\text{Red}}t)^2, \tag{9}$$

where $\hat{\mu}_{\text{Red}}$ and $\hat{\alpha}_{\text{Red}}$ are estimated under the constraint $\mu_1 = \mu_2 = \mu_{\text{Red}}$ (i.e., $X_t = \mu_{\text{Red}} + \alpha_{\text{Red}}t + \epsilon_t$ for $1 \leq t \leq n$).

Under the null hypothesis the F_{\max} percentiles of model (6) are given in Table 1, which was obtained by simulating 100 000 F_{\max} values for each series length n . The simulation procedure is similar to the one described in Lund and Reeves (2002).

One of the advantages of using model (6) is that it has one less parameter to estimate, and hence it has smaller sampling variability {the variance of the estimates of α is $\text{var}(\hat{\alpha}) = \sigma_\epsilon^2 / \sum_{t=1}^n [t - (n + 1)/2]^2$ in this case} and a higher power of detecting step-type changepoints (i.e., more likely to detect a changepoint when it really exists). To illustrate this, for each series length

n , we simulated 200 time series of the form $X_t = \mu + \alpha t + \epsilon_t$ with arbitrary parameters μ and α and σ_ϵ (note that under the null hypothesis the F_{\max} statistic does not depend on the particular values of μ and α and σ_ϵ). Then, we imposed a step $\Delta X = \sigma_\epsilon j/100$ at an arbitrary point c ($1 < c < n$) in the j th simulated time series ($j = 1, 2, \dots, 200$). In other words, the steps imposed in the 200 simulated time series range from $0.01\sigma_\epsilon$ to $2\sigma_\epsilon$. Then, both models (6) and (1) were applied to each of the 200 time series to detect the imposed step-type changepoint. The results clearly show that model (6) has a somewhat higher power of detection than model (1), especially when the time series is short (cf. Table 2) or the ‘‘step size to noise’’ ratio (i.e., $\Delta X/\sigma_\epsilon$) is low (not shown). Note that each of our simulations was drawn from a model with a constant slope, in which case model (1) is not suitable and hence it is not surprising that its power of detection is lower than that of model (6).

In particular, model (6) can better identify a negative step (i.e., $\mu_1 > \mu_2$) that occurs in time series of a decreasing trend, or a positive step (i.e., $\mu_1 < \mu_2$) that occurs in time series of an increasing trend. Such steps could be ‘‘masked’’ and/or misrepresented as trend-type changes when using model (1). These situations were also tested in the power of detection test simulations described in the last paragraph, with results shown in Table 2. Here is a ‘‘real world’’ example: an observer change (from a part-time observer to a full-time observer) at Gander International Airport (Canada) on 1 April 1997 introduced a step-type changepoint between March and April of 1997 in the monthly cloud-cover time series: the full-time observer observes much more frequently in each hour and hence reports more cloudy conditions than the part-time observer does (cf. Fig. 1; the latter had other duties in addition to making observations). Both models (6) and (1) were applied to the time series of monthly counts of occurrences of 1/10 and 9/10 cloud-covered sky conditions (i.e., 1/10 and 9/10 of the sky covered by clouds). The results are shown in Fig. 1: When the step size to noise ratio is

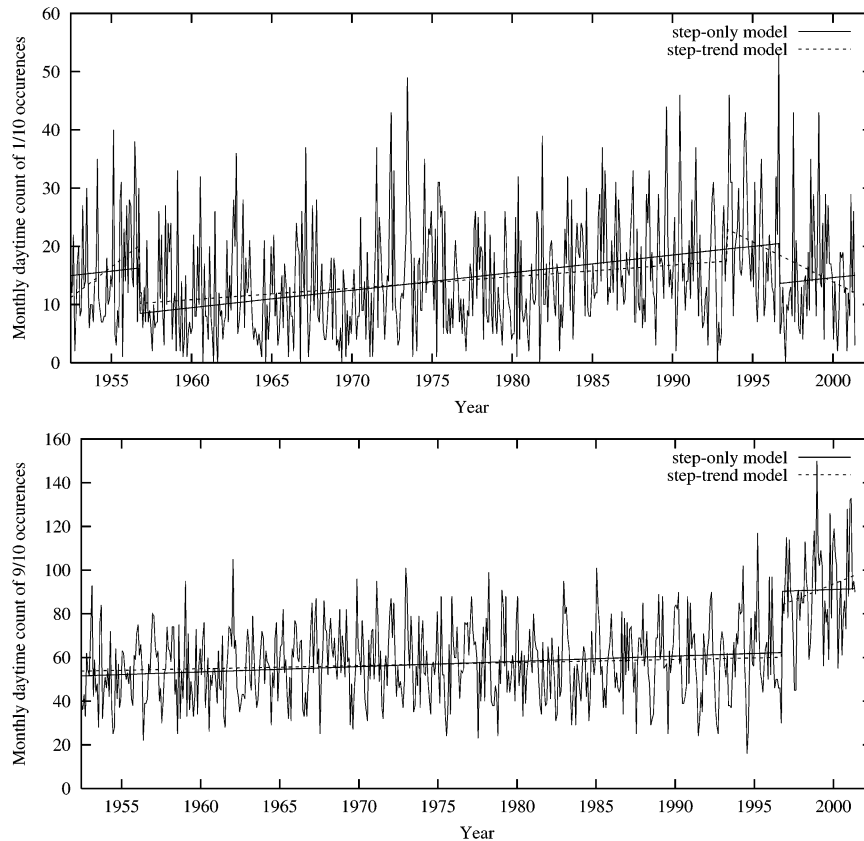


FIG. 1. Changepoints detected using the step-only model (6) and the step-trend model (1) for (top) 1/10 cloud cover and (bottom) 9/10 cloud cover.

high, such as the case of 9/10 cloud-cover (Fig. 1, bottom), both models accurately identified the changepoint, although the sizes of the step and trend estimated by the two models are different. However, when the step size to noise ratio is lower, such as the case of 1/10 cloud cover (Fig. 1, top), model (6) clearly outperforms model (1): while the changepoint identified by model (6) is identical to the actual changepoint, the one identified by model (1) is 39 time intervals (months in this case) earlier and involves a notable trend-type change, which is far from the reality.

Having detected a changepoint using model (6), one can further test whether the slopes are equal before and after the changepoint time. A standard normal z -score hypothesis test with $z = [\hat{\alpha}_1 - \hat{\alpha}_2]/[\text{var}(\hat{\alpha}_1) + \text{var}(\hat{\alpha}_2)]^{1/2}$ would suffice. But again, a trend-type change should

only be adjusted if there is sufficient evidence showing that it is not rooted in true climate change/variability.

Finally, note that both models (6) and (1) can easily accommodate reference series (see Lund and Reeves 2002 for more details).

Acknowledgments. The author thanks Dr. Francis W. Zwiers for his helpful comments on an earlier version of this manuscript. The reviewer Dr. Robert Lund is also acknowledged. The research was partly funded by the Action Plan 2000 science program of the Canadian government.

REFERENCES

- Lund, R., and J. Reeves, 2002: Detection of undocumented change-points: A revision of the two-phase regression model. *J. Climate*, **15**, 2547–2554.