Quantifying the lead time required for a linear trend to emerge from natural climate variability

Jingyuan Li,* David W.J. Thompson, and Elizabeth A. Barnes

Department of Atmospheric Science, Colorado State University, Fort Collins, CO

Susan Solomon

Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA

*Corresponding author address: Jingyuan Li, Department of Atmospheric Science, Colorado State University, 1371 Campus Delivery, Fort Collins, CO 80523.

E-mail: jingyuan.li@colostate.edu
ABSTRACT

We introduce a simple analytic expression for calculating the lead time required for a linear trend to emerge in a Gaussian first order autoregressive process. The expression is derived from the standard error of the regression and is tested using the NCAR Community Earth System Model Large Ensemble of climate change simulations. It is shown to provide a robust estimate of the point in time when the forced signal of climate change has emerged from the natural variability of the climate system with a predetermined level of statistical confidence. The expression provides a novel analytic tool for estimating the time of emergence of anthropogenic climate change and its associated regional climate impacts from either observed or modeled estimates of natural variability and trends.
1. Introduction

The time of emergence (TOE) is defined as the point in time when the signal of climate change emerges from the underlying noise of background natural variability (Madden and Ramanathan 1980; Santer et al. 1995; Weatherhead et al. 1998; Christensen et al. 2007; Giorgi and Bi 2009; Mahlstein et al. 2011; Deser et al. 2012b; Hawkins and Sutton 2012; Zappa et al. 2015). It is helpful for anticipating when the impacts of climate change will have significant effects across societies and ecosystems and is highly useful in risk assessments, mitigation policies, and climate adaptation planning.

As noted in the IPCC AR5 (Kirtman et al. 2013), there is “no single metric” for estimating TOE. But for the most part, TOE is estimated as the first lead time when the anthropogenic signal in climate change exceeds a predetermined factor of the amplitude of the natural variability, often presented as a signal-to-noise problem. In this case, the TOE for a time series $x(t)$ is expressed as:

$$n_{TOE} = \frac{ks_e}{b}$$  \hspace{1cm} (1)

where $n_{TOE}$ is the time of emergence (the number of time steps when the anthropogenic signal in climate change has “emerged”), $k$ is the required ratio of the forced signal to the natural variability (generally between 1 and 3), $s_e$ is the amplitude of the internal (unforced) variability, and $b$ is the linear trend per time step. Note that in this study, the term “natural variability” refers to the sum of internal climate variability due to stochastic dynamic processes and external variability due to natural forcings such as volcanic eruptions and solar variability.

Most previous studies of TOE are based on empirical estimates of the first lead time when Equation 1 (or a closely related variant) is satisfied. The differences lie in the methodologies used to determine $b$, $s_e$, and $k$. For example, Giorgi and Bi (2009), Mahlstein et al. (2011), Diffenbaugh and Scherer (2011), and Zappa et al. (2015) all define $b$ as the change in the climate state averaged...
over an ensemble of climate change simulations, where the forced signal is smoothed with differ-
ent averaging periods. Weatherhead et al. (1998) estimates $b$ using a generalized least squares
regression model. Hawkins and Sutton (2012) define $b$ as the linear projection of regional tem-
peratures onto smoothed values of simulated global-mean temperatures. Giorgi and Bi (2009),
Mahlstein et al. (2011), and Hawkins and Sutton (2012) estimate $b$ from ordinary least squares
(OLS) linear regression and prescribe a signal to noise ratio ($k$) that is an integer factor of the
natural variability. Christensen et al. (2007), Deser et al. (2012b) and Zappa et al. (2015) con-
sider various “epoch differences” and a value of $k$ derived from the t-statistic for the difference
of means. Mahlstein et al. (2012) also considers differences between epochs when assessing the
time of emergence, and apply a Kolmogorov-Smirnov test to assess whether two sample epochs
are significantly different (i.e., rather than a t-statistic).

The existing literature on TOE provides valuable insight into the point in time when anthro-
pogenic climate change will emerge from natural climate variability on regional spatial scales.
But it also has several shortcomings. The methodologies used to estimate the trend ($b$ in Equa-
tion 1), the amplitude of the natural variability ($s_e$ in Equation 1), and the predetermined signal to
noise ratio ($k$ in Equation 1) vary widely from one study to the next, which makes it difficult to
reproduce and compare different estimates of the TOE. Times of emergence defined on the basis
of an integer signal to noise ratio (e.g., Hawkins and Sutton 2012) do not correspond to a particular
level of statistical significance. Several existing methods require smoothing the data using a wide
range of methodologies. Furthermore, many of the methods are based on empirical, not analytical
techniques.

The purposes of this study are twofold. First, we introduce a simple and novel expression
for estimating the lead time required for a linear trend to emerge from natural variability at a
predetermined level of statistical confidence. The expression is developed from the standard error
of the regression, which is widely used in climate research, but as far as we know has not been
exploited for the explicit purpose of calculating TOE. Second, we will test the resulting expression
in a large ensemble of climate change simulations. The results demonstrate the robustness of the
assumptions that underlie the expression, and make clear its utility for assessing the emergence
of linear trends in climate data. The expression is developed in Section 2. The application of the
expression to climate trends is explored in Section 3, and its advantages relative to other methods
of calculating TOE are considered in Section 4. Conclusions are provided in Section 5.

2. An analytic expression for the lead time required for a linear trend to emerge from internal
variability

Consider the case of a first order autoregressive (AR(1)) time series of length \( N \) with a linear
trend \( b \) imposed upon it such that

\[
x(n_t) = b n_t + \alpha x(n_t - 1) + \beta \epsilon(n_t)
\]

where \( n_t = 1, 2, ..., N \) is the number of time steps, \( x(0) = 0 \) by assumption, and \( \epsilon \) is white noise
(independent Gaussian noise with mean of zero and variance of 1). Further, \( \alpha \) and \( \beta \) are restricted
such that \( \alpha^2 + \beta^2 = 1 \) since the \( x(n_t) \)'s have been normalized such that \( \text{Var}(x(n_t)) = 1 \), where \( \alpha \)
represents the memory in the time series \( x(n_t) \) from one time step to the next,

Here we will estimate the trend \( b \) using simple linear regression, where \( \hat{b} \) is the regression
estimator of the trend. The parameter \( \alpha \) can be estimated as the lag-one autocorrelation of the
time series \( r_1 \). The confidence interval in the total change in \( x(n_t) \) over time \( n_t \) can thus be
expressed as:

\[
CI = \hat{b} n_t \pm \epsilon
\]
where \( n_t \) is again the number of time steps and \( e \) is the uncertainty in the change in \( x(n_t) \) over time \( n_t \) given by \( \hat{b} n_t \). The margin of error \( (e) \) is given in units \( \Delta x/(n_t \Delta t) \), where \( \Delta t \) is the time step. The trend \( \hat{b} \) is given in units \( \Delta x/\Delta t \) so that \( \hat{b} n_t \) is the change over the length of the record and has the same units as \( e \).

Following Thompson et al. (2015), if detrended values of \( x(n_t) \) are well-modeled as an AR(1) process, then the margins of error on the linear trend in \( x(n_t) \) can be expressed as:

\[
e = t_c \cdot n_t \cdot s \cdot \gamma(n_t, r_1) \cdot g(n_t)
\]

where

\[
\gamma(n_t, r_1) \equiv \left( \frac{[n_t - 2]}{n_t \left( \frac{1 - r_1}{1 + r_1} \right) - 2} \right)^{1/2}
\]

and

\[
g(n_t) \equiv \sqrt{\frac{12}{n_t^2 - n_t}}
\]

In Equation 2, \( t_c \) is the t-statistic for the desired confidence level and \( s \) is the standard deviation of the residuals of the regression (i.e., of detrended values of \( x(t) \)). The expressions for \( \gamma(n_t, r_1) \) and \( g(n_t) \) account for 1) the effects of persistence on the estimate of \( s \), where \( r_1 \) is the lag-one autocorrelation of the residuals, and 2) the standard deviation of the time axis, respectively. Note that Equation 2 is simply the standard error of the regression for the case where 1) the predictor is time, and 2) detrended values of the predictand are well-modeled as an AR(1) process (e.g., Santer et al. 2000; Thompson et al. 2015).

The trend in \( x(t) \) is statistically significant when it exceeds its margins of error. Setting \( e = \hat{b} n_t \) in Equation 2 yields:

\[
\left( \frac{T_{\text{SIG}}^3 - T_{\text{SIG}}}{12} \right) \cdot \frac{T_{\text{SIG}} \left( \frac{1 - r_1}{1 + r_1} \right) - 2}{[T_{\text{SIG}} - 2]} = \left( \frac{t_c s}{\hat{b}} \right)^2
\]

where \( T_{\text{SIG}} \) denotes the lead time when the trend in \( x(t) \) is statistically significant (in units of time steps). That is: \( T_{\text{SIG}} \) is the number of time steps required for a linear trend superposed on an
AR(1) process to be statistically significant at the desired confidence level. The value of $T_{SIG}$ can be trivially calculated given the estimated amplitude of the forced signal ($\hat{b}$), the amplitude of the natural climate variability ($s$), and the lag-one autocorrelation of the natural climate variability ($r_1$). It requires no subjective analysis choices (such as the length of the periods used in epoch differences) and no smoothing of the data.

Equation 3 can be simplified greatly given two conditions: 1) detrended values of $x(n_t)$ are not serially correlated ($r_1 \approx 0$). This condition holds for climate variability at most terrestrial locations on interannual timescales, since there is very little memory in the internal variability of land surface climate from one year to the next (see discussion in Thompson et al. (2015)). 2) The trend length is at least $\sim 10$ time steps. In this case, $T_{SIG}^3 \gg T_{SIG}$ and the two tailed t-statistic for the 95% confidence level is $t_c \sim 2$. Applying both conditions yields a simplified version of Equation 3 that is suitable for cases where the internal variability is not serially correlated from one time step to the next:

$$T_{95\%} \approx 3.6 \left( \frac{s}{\hat{b}} \right)^{2/3}$$

where $T_{95\%}$ is the lead time when the trend in $x(t)$ is statistically significant at the 95% confidence level. Equation 4 places Equation 3 in a “signal to noise” format similar but not identical to that used in many previous studies. For all our analyses we use the general expression of Equation 3.

3. Application to climate trends

In this section, we test the robustness of Equation 3 ($T_{SIG}$) for assessing the point in time when the signature of anthropogenic warming emerges from the background noise of natural climate variability (i.e., achieves statistical significance) on regional scales. We perform the assessment for land surface temperature changes at individual grid boxes. To do so, we exploit a large ensemble of climate change simulations.
In a large ensemble of climate change simulations, each individual ensemble member may be viewed as a unique realization of “model reality.” In the context of large ensembles, for a confidence level of 95%, the expression for $T_{SIG}$ should thus correspond to the lead time when 95% of all possible realizations of “model reality” exhibit trends of the same sign as the forced signal. Here we test the expression for $T_{SIG}$ using output from the NCAR Community Earth System Model Large Ensemble (CESM-LE).

Details of the simulations are provided in Hurrell et al. (2013) and Kay et al. (2015). In short, the CESM-LE consists of 40 climate change simulations run using the same model configuration with the same external forcings. Differences in climate trends from one realization to the next are due entirely to the internal variability (note that this is different from natural variability) in the model. Here, we use the original 30 CESM-LE simulations released in 2014. The runs are available from 1920 to 2100, with historical forcings used for the period 1920-2005 and RCP8.5 forcings used for the period 2006-2100. The analyses are based on seasonal-mean values of surface temperature for the Northern Hemisphere (NH) cold (October-March) and warm (April-September) season months over the 1970-2015 period. There are three reasons for choosing this time period: 1) We wish to focus on the period with the largest global warming observed to date (Bindoff et al. 2013); 2) We wish to compare results derived from the CESM-LE with results derived from observations, which are sparse before 1970 and end in 2015; and 3) Our analytic expression is based on a linear least-squares fit to the forced signal, which is approximately linear over the selected period (the linear assumption is discussed in more detail in the final section). The simulated trends in global-mean surface temperature from the CESM-LE are not linear over the full simulation period 1920-2100, i.e., the trends increase from roughly zero in the mid-20th century to roughly 0.5 K/decade in the latter part of the 21st century (Kay et al. (2015), c.f., Fig. 2). However, they are approximately linear over the comparatively short 1970-2015 period examined here.
The expression for $T_{SIG}$ is tested as follows. First, we calculate the “empirically-derived $T_{SIG}$” by empirically calculating the time step when 95% of all ensemble members exhibit trends of the same sign as that of the model forced signal. Since we have 30 ensemble members, the “empirically-derived $T_{SIG}$” is defined as the first time step when 29 out of 30 ensemble members exhibit warming in the current and all subsequent time steps.

Second, we calculate the “analytically-derived $T_{SIG}$” at all grid boxes by solving Equation 3 for $T_{SIG}$ using: 1) the ensemble mean trends in temperature calculated over the period 1970-2015 ($\bar{b}$); 2) the standard deviations of the residuals of the regression (i.e., the variability about the long-term trends; $s$); and 3) the lag-1 autocorrelations of the residuals of the regression ($r_1$). The ensemble mean trends are assumed to reflect the forced signal in surface temperature. The standard deviation and lag-1 autocorrelation of the residuals are found by 1) detrending the seasonal-mean temperature time series in each of the ensemble members and at all grid boxes and 2) calculating the pooled standard deviations and ensemble-averaged lag-1 autocorrelations of the residual time series. The resulting values of $s$ and $r_1$ are assumed to reflect the amplitude and persistence of the model’s natural variability.

In principle, the model’s natural variability can be isolated using a variety of different methodologies. We have chosen to isolate the natural variability by removing a linear fit to the temperature time series in all ensemble members since Equation 3 is a function of the standard error of the residuals of the regression. In the Appendix, we explore two alternative methods for isolating the natural variability: 1) removing an second order polynomial (rather than linear) fit to the grid box temperature time series, thus allowing for exponential changes in temperature; and 2) removing the grid box ensemble-mean temperature time series from the grid box temperature time series in all ensemble members, thus explicitly removing the forced signal from the ensemble members and
allowing for forced variability on a range of timescales. In practice, the amplitudes of the natural variability are effectively identical for all three methods (Figure A1).

Figure 1 illustrates the analytically and empirically-derived values of $T_{SIG}$ in NH wintertime surface temperatures at three grid boxes: one from Northern Hemisphere midlatitudes (at a grid box whose node is close to London, UK); one from a region of relatively high temperature variance (located in central Siberia); and one from a region of relatively low temperature variance (close to Jakarta, Indonesia). The sloping black lines in all three panels indicate the ensemble mean trend over the 1970-2015 period at each location. As noted above, the ensemble mean trend is interpreted as the “forced signal” of climate change. The small red dots indicate the trends in surface temperature from all 30 individual ensemble members, where the trends start in 1970 and end on the date indicated on the ordinate axis. The units on all trends are K/length of the record. The dashed lines in all three panels indicate the 95% margins of error on the “forced signal,” where the margins of error are derived from Equation 2. Note that the close agreement of the 95% margins of error given by Equation 2 (dashed lines) and the large ensemble (red dots) attest to the robustness of Equation 2 for estimating the role of natural variability in climate trends (see Thompson et al. (2015)).

The values of $T_{SIG}$ are calculated at each terrestrial location by inserting the estimated forced signal and natural variability for each grid point into Equation 3. For example, in the case of London, the estimated forced signal is $\hat{b} = 0.02 \text{ K/year}$, the amplitude of the natural variability is $s = 0.6 \text{ K}$, and the winter-to-winter autocorrelation is not significantly different from zero ($r_1 \sim 0$). Inserting the above values into Equation 3 yields $T_{SIG} = 40 \text{ years}$, or 2010, which by definition is the lead time when the lower bound of the 95% confidence levels intersects zero (the intersection is marked by the vertical blue line in Figure 1). The forced signal and natural variability both vary from one location to the next in Figure 1, but in general the latter dominates the variations in $T_{SIG}$. 
For example, $T_{SIG}$ is longer over Siberia where the interannual temperature variance is much larger ($s = 2$ K), but shorter over Indonesia where the interannual temperature variance is relatively small ($s = 0.2$ K). The inverse relationship between regional temperature variance and the emergence of the forced signal has been noted extensively in previous works (e.g., Christensen et al. 2007; Mahlstein et al. 2011). The key point in Figure 1 is that the expression given in Equation 3 for $T_{SIG}$ clearly provides a simple and robust estimate of the first lead time at which 95% of the realizations of model “reality” (as given by individual ensemble members) exhibit warming.

Figure 2 shows the results for a similar test at all terrestrial grid boxes during the NH winter and summer seasons. The top panels indicate the “empirically-derived” values of $T_{SIG}$ found by empirically calculating the lead time when 29 of the 30 ensemble members exhibit warming in the current and subsequent time steps. The bottom panels in Figure 2 indicate the analytically-derived values of $T_{SIG}$ from Equation 3 (very similar results are derived for Equation 4, since the lag-one autocorrelation of seasonal-mean surface temperature is not significantly different from zero at most terrestrial grid boxes). Warm colors indicate lead times of 45 years since 1970 (e.g., times of emergence less than 2015). White denotes lead times that exceed the analysis period (TOE beyond 2015), while grey denotes oceans and any missing data.

The strong similarities between the top and bottom panels in Figure 2 are important. They suggest that the lead time given by Equation 3 provides a reliable estimate of the geographical pattern of detection time - the time at which 95% of all possible renditions of model “reality” indicate trends of the same sign as the forced signal. They also support the assumptions that underlie Equation 3, e.g., that the natural variability is sufficiently Gaussian and the forced signal sufficiently linear to warrant use of the standard error of the regression. As noted in numerous previous studies (e.g., Christensen et al. 2007; Mahlstein et al. 2011; Hawkins and Sutton 2012),
the forced signal in surface temperature emerges earliest in regions where the variance is smallest, i.e., the tropics during all seasons and the extratropics during the warm season months.

The top panel in Figure 3 examines analogous results, but for the case where 1) the estimated forced signal ($\hat{b}$) is again derived from the CESM-LE but 2) the natural variability ($s$ and $r_1$) is derived from observations of surface temperature from the HadCRUT4 dataset. The HadCRUT4 data are obtained from the Climatic Research Unit at the University of East Anglia and are analyzed on a 5x5 degree resolution for the time period January 1970 until September 2015. The advantage of using observations to estimate the natural variability is that - by definition - they best reflect the variance of the “real-world.” The disadvantages are that: 1) the observed record may be too short to fully sample variability on decadal timescales; and 2) the observed record includes missing data and may include residual errors that influence estimates of the observed variability. As done for Figure 2, $\hat{b}$ is defined as the ensemble-mean trend from CESM over 1970-2015. In contrast to Figure 2, $s$ is found by 1) detrending the observed wintertime-mean surface temperature data as a function of grid box, and 2) calculating the standard deviation of the resulting time series. Note that the detrending methodology is identical to that applied to individual ensemble members (except for the pooling) in Figure 2 and is discussed in the Appendix.

Results based on the amplitude of observed natural variability are similar but not identical to those based on the natural variability displayed by the CESM-LE. Regions of strong agreement between the top panels in Figure 3 and Figures 2c,d correspond to areas where the variability in the CESM-LE closely corresponds to that in the observations. Regions where the top panels in Figure 3 and Figures 2c,d are notably different point to areas where differences between the simulated and observed natural variability lead to differences in the lead time when surface warming emerges in a statistically significant sense. For example, lead times over the western United States are at least a decade shorter when using observed rather than modeled estimates of natural variability.
The bottom panel in Figure 3 shows analogous results, but in this case both the natural variability and the forced signal (the linear trends) are estimated from observations. That is: the forced signal \( \hat{b} \) is defined as the linear trends from the observations over 1970-2015, and the natural variability \( s \) is found in an identical manner to the top panel. The observed trends are more susceptible to sampling variability than the ensemble-mean trends since they reflect only one realization of reality, particularly over regions of large temperature variance such as the Northern Hemisphere midlatitudes during winter (e.g. Deser et al. 2012a). Nevertheless, the resulting lead times are interesting in that they provide a purely observational estimate of the lead time when the observed warming emerges from the observed internal climate variability.

The differences between the upper and lower panels in Figure 3 derive from differences between trends from the CESM-LE ensemble mean and the observed trends. The CESM-LE ensemble mean trends from 1970-2015 are weaker than those derived from the observations over much of the tropical land areas, Europe and East Asia during summer (see Figure 4). Hence, the purely observational lead times in these regions are shorter than those derived from the ensemble-mean trends. Are the TOE estimates derived entirely from observations outside the range of TOE estimates derived from individual ensemble members? To test this, we calculated the TOE at all grid boxes and for all ensemble members using the individual ensemble member trends and detrended standard deviations as estimates of the forced signal and natural variability (i.e., we treated output from individual ensemble members as we treated the observations in the lower panel of Figure 3). Importantly, the observed TOE estimates given in the bottom panel of Figure 3 lie within the 95% bounds on TOE estimates derived from individual ensemble members over 95% of all land areas (Figure 5).
4. Discussion

The standard error of the regression is widely used in climate research. But to the best of our knowledge, it has not been explicitly used to develop an expression for the time of emergence of anthropogenic climate change. The resulting expression for $T_{SIG}$ provides a simple, novel, and complementary tool for assessing the lead time when anthropogenic climate change will emerge from natural climate variability. The methodology has some disadvantages relative to existing methods, e.g., it assumes that the natural variability is Gaussian, which is not required in existing metrics based on the Kolmogorov-Smirnov test (e.g., Mahlstein et al. 2012). However, it also has several key advantages:

1) The expression for $T_{SIG}$ given by Equation 3 (and Equation 4 for the case where the data are not serially correlated) indicates the lead time when the forced signal of the trend has emerged in a statistically significant sense. Some previous studies explicitly consider TOE in the context of statistical significance (e.g., Christensen et al. 2007; Deser et al. 2012a; Zappa et al. 2015). But others consider it in the context of specific values of the natural variability. For example, consider the case of TOE defined as the lead time when the forced signal exceeds two times the amplitude of the natural variability (e.g., one of the criteria outlined in Hawkins and Sutton (2012)). At the grid box close to London, the TOE for $k = 2$ in Equation 1 occurs at a lead time of 74 years, which is more than three decades longer than the point in time when the trend is significant (Figure 1). Similarly large differences are found throughout much of the extratropics (Figure 6).

2) The expression for $T_{SIG}$ exploits linear regression instead of epoch differences to estimate the linear trend. For example, Christensen et al. (2007), Deser et al. (2012b), and Zappa et al. (2015) all consider statistical significance when assessing the time of emergence, but consider the differences in means between epochs of various lengths rather than linear trends. The distinction
is important. Linear regression uses all of the data in a time series, while epoch differences only take data from the beginning and end of the time series. Additionally, the variance of the epoch difference estimator varies greatly depending on the length of the epoch used, and is always larger than the variance of the linear trend estimator for AR(1) time series with lag-1 autocorrelations less than about 0.85 (Barnes and Barnes 2015). Thus, for all time series with a lag-1 autocorrelation less than 0.85, we believe the linear regression estimator to be preferable to epoch differences.

3) The expression for $T_{SIG}$ can be solved analytically and requires little manipulation of the data. Hence the resulting estimate of TOE can be easily reproduced from one study to the next, and readily compared across different model configurations and forcing scenarios.

5. Conclusions

The impacts of anthropogenic climate change are felt locally. But the lead times when warming and related impacts emerge from the natural climate variability vary greatly from one location to the next. The expression derived in Equation 3 provides a simple analytic tool for estimating the lead time when regionally-dependent impacts of climate change emerge from the natural variability in a statistically significant sense.

We have focused on the application of Equation 3 to surface temperature, but the expression holds for any time series where the following three conditions are met: 1) the forced signal can be modeled as a linear trend; 2) the statistics of the natural variability (detrended values of the time series) are Gaussian; and 3) the standard deviation of the natural variability is stationary. These three assumptions derive from our use of the standard error of the regression. The bases for all three assumptions are discussed and justified in Thompson et al. (2015). The linear assumption warrants additional comment here.
In principle, the signature of anthropogenic forcing in the climate system is not necessarily linear. For example, atmospheric aerosols likely contributed to the slowdown of globally averaged warming during the middle 20th century (Bindoff et al. 2013), and the surface temperature trends of the next 50 years are expected to be notably larger than those of the past 50 years (Kirtman et al. 2013). However, in practice, the simulated response of surface temperature to greenhouse gas increases is closely linear on timescales shorter than roughly $\sim 50$ years, including the 1970-2015 period considered here [e.g., see Kay et al. (2015), c.f., Fig. 2].

The methodology outlined here is derived from statistical tools that are used widely in climate change research. It is potentially useful for climate change research for three primary reasons. One, it provides an *analytic* estimate of the lead time required for a trend to emerge, and can thus be trivially calculated given a) the amplitude and autocorrelation of the observed natural variability and b) the simulated forced signal. Two, it provides an estimate of the time required for a linear trend to emerge in a *statistically significant* sense, rather than as a (statistically arbitrary) factor of the internal variability. And three, the expression requires no treatment of the data, which renders the resulting lead times easy to compare across different model configurations, different forcing scenarios, and different estimates of the natural variability.

*Acknowledgments.* We thank Benjamin D. Santer and two anonymous reviewers for their insightful and constructive comments on the manuscript. We also thank Daniel Cooley for helpful discussions on the statistical methodology. JL and DWJT are supported by the NSF Climate Dynamics program. EAB is supported, in part, by the Climate and Large-Scale Dynamics Program of the National Science Foundation under grants AGS-1419818 and AGS-1545675. SS is supported, in part, by the Climate and Large-Scale Dynamics Program of the National Science Foundation under grant AGS-1419667.
Figure A1 compares three different methods for isolating the natural/internal variability in individual ensemble members: 1) removing a linear fit to the temperature time series in each of the ensemble members (top panel); 2) removing a second order polynomial (rather than linear) fit to the grid box temperature time series in all ensemble members; and 3) removing the grid box ensemble-mean temperature time series from the grid box temperature time series in all ensemble members.

The three methods have various advantages and disadvantages. The advantages of Method 1 are that a) the residuals of the linear fit correspond directly to the residuals of the regression that form the basis for $s$ in Equation 3; and b) a similar method can be applied to observations in the absence of climate model output. The disadvantages are that a) the anthropogenic signal is not necessarily best modeled as a linear trend; b) the linear fits include a component of the natural variability (i.e., internal variability includes a stochastic trend component); and c) the linear fit does not account for externally forced variability due to, say, volcanic eruptions. Method 2 is similar to Method 1, but has the additional advantage that it allows for exponential changes in temperature. However, the residuals of the second order polynomial fit do not -strictly speaking - correspond to the residuals of the regression that form the basis for $s$ in Equation 3. The residuals of Method 3 also do not form the basis for $\sigma$ in Equation 3, but removing the ensemble mean time series arguably reflects the most robust method for removing the variability due to all forms of external forcing in the CESM-LE, including anthropogenic forcings (e.g., due to increasing greenhouse gases) and external natural forcings (e.g., due to volcanoes).

In practice, the amplitudes of the “natural” variability (i.e., the standard deviations of the residuals of the fits) are effectively identical for all three methods (Figure A1).
References


LIST OF FIGURES

Fig. 1. Trend amplitudes for modeled surface temperature at the grid point collocated with (a) London, UK, (b) central Siberia, (c) Jakarta, Indonesia, using CESM-LE output. The red dots indicate actual trends from all 30 ensemble members and the dashed lines indicate the predicted ranges of trends found by applying Eq. 2 to the statistics of the model internal variability. The blue vertical line indicates the lead time when the forced trend is statistically significant as per Eq. 3. See text for details. 22

Fig. 2. Using CESM-LE output to test Equation 3. The top panels show the “empirically derived” lead times when the trends emerge from natural variability, calculated as the time step when 29 out of 30 ensemble members exhibit positive trends in the current and all subsequent time steps. The bottom panels show the “analytically derived” lead times ($T_{SIG}$) derived by applying Equation 3 to the observed natural variability. Winter corresponds to the Oct-March means; summer to the April-Sept means. Note that all lead times beyond the limit of the analysis period (45 years or 2015) are white. Grey denotes oceans and any regions of missing data. 23

Fig. 3. As in Figure 2, bottom, but for lead times calculated by applying Equation 3 to the (top) ensemble mean trends from CESM-LE and the observed natural variability and (bottom) linear trends from the observations and the observed natural variability. The observed natural variability is estimated in both panels as the standard deviation of the detrended data. The observations are used over the period 1970-2015. White denotes lead times beyond the limit of the analysis period (larger than 45 years), while grey denotes oceans and any regions of missing data. See text for details. 24

Fig. 4. Differences between the 1970-2015 ensemble-averaged trends from CESM-LE and trends from HadCRUT4 observations for (a) the winter season (October-March), and (b) the summer season (April-September). The CESM-LE trends were used in calculating the lead times in the top panels of Figure 3, while the observed trends were used in calculating the lead times in the bottom panels of Figure 3. The predominance of cool colors for both seasons indicate that observed trends from 1970-2015 were larger than the simulated ensemble mean trends over the same period. Grey denotes oceans and any regions of missing data. 25

Fig. 5. Grid boxes where TOE calculated from HadCRUT4 surface temperature observations fall outside the 95% bounds on TOE calculated for individual ensemble member trends and standard deviations. Only 5% of the observed TOE estimates lie outside the bounds given by the individual ensemble members. 26

Fig. 6. Comparison between lead times calculated using Equation 3 (top) and Equation 1 where $k = 2$ (bottom). In both cases, the forced signal is given as the ensemble mean temperature trends over 1970-2015, and the natural variability as the detrended observed interannual standard deviation. Grey denotes oceans and any regions of missing data. Note the top panel is reproduced from Figure 3a. 27

Fig. A1. Comparisons of the standard deviations calculated from the CESM-LE over the period 1970-2015 using three different methods to remove the long-term forced signal. Panels (a) and (b) show the pooled standard deviations after removing the linear trend from all grid point time series in all ensemble members; panels (c) and (d) show pooled standard deviations after removing a 2nd order polynomial fit from all grid point time series in all ensemble members; and panels (e) and (f) show pooled standard deviations after removing the ensemble mean time series from all grid point time series in all ensemble members. Grey denotes oceans and any regions of missing data. 28
FIG. 1. Trend amplitudes for modeled surface temperature at the grid point collocated with (a) London, UK, (b) central Siberia, (c) Jakarta, Indonesia, using CESM-LE output. The red dots indicate actual trends from all 30 ensemble members and the dashed lines indicate the predicted ranges of trends found by applying Eq. 2 to the statistics of the model internal variability. The blue vertical line indicates the lead time when the forced trend is statistically significant as per Eq. 3. See text for details.
FIG. 2. Using CESM-LE output to test Equation 3. The top panels show the “empirically derived” lead times when the trends emerge from natural variability, calculated as the time step when 29 out of 30 ensemble members exhibit positive trends in the current and all subsequent time steps. The bottom panels show the “analytically derived” lead times ($T_{SIG}$) derived by applying Equation 3 to the model natural variability. Winter corresponds to the Oct-March means; summer to the April-Sept means. Note that all lead times beyond the limit of the analysis period (45 years or 2015) are white. Grey denotes oceans and any regions of missing data.
Fig. 3. As in Figure 2, bottom, but for lead times calculated by applying Equation 3 to the (top) ensemble mean trends from CESM-LE and the observed natural variability and (bottom) linear trends from the observations and the observed natural variability. The observed natural variability is estimated in both panels as the standard deviation of the detrended data. The observations are used over the period 1970-2015. White denotes lead times beyond the limit of the analysis period (larger than 45 years), while grey denotes oceans and any regions of missing data. See text for details.
Fig. 4. Differences between the 1970-2015 ensemble-averaged trends from CESM-LE and trends from HadCRUT4 observations for (a) the winter season (October-March), and (b) the summer season (April-September). The CESM-LE trends were used in calculating the lead times in the top panels of Figure 3, while the observed trends were used in calculating the lead times in the bottom panels of Figure 3. The predominance of cool colors for both seasons indicate that observed trends from 1970-2015 were larger than the simulated ensemble mean trends over the same period. Grey denotes oceans and any regions of missing data.
FIG. 5. Grid boxes where TOE calculated from HadCRUT4 surface temperature observations fall outside the 95% bounds on TOE calculated for individual ensemble member trends and standard deviations. Only 5% of the observed TOE estimates lie outside the bounds given by the individual ensemble members.
Fig. 6. Comparison between lead times calculated using Equation 3 (top) and Equation 1 where $k = 2$ (bottom). In both cases, the forced signal is given as the ensemble mean temperature trends over 1970-2015, and the natural variability as the detrended observed interannual standard deviation. Grey denotes oceans and any regions of missing data. Note the top panel is reproduced from Figure 3a.
Fig. A1. Comparisons of the standard deviations calculated from the CESM-LE over the period 1970-2015 using three different methods to remove the long-term forced signal. Panels (a) and (b) show the pooled standard deviations after removing the linear trend from all grid point time series in all ensemble members; panels (c) and (d) show pooled standard deviations after removing a 2nd order polynomial fit from all grid point time series in all ensemble members; and panels (e) and (f) show pooled standard deviations after removing the ensemble mean time series from all grid point time series in all ensemble members. Grey denotes oceans and any regions of missing data.