Warming Pause?

Origins of the Global Warming ’hiatus’ Concept

Jan Galkowski
Westwood Statistical Studios
empirical_bayesian@ieee.org
Westwood, MA 02090

May 10, 2014
(revised draft 004c)
When present, changes are indicated with green changebars, such as to the left of this paragraph.
1 Who Be I

I’m Jan Galkowski. I’m a statistician and signals engineer, with an undergraduate degree in Physics and a Masters in EE & Computer Science. Professionally, I work for Akamai Technologies of Cambridge, MA, where I study time series of Internet activity and other data sources, doing data analysis primarily using spectral and Bayesian computational methods. I am not a climate scientist, but am keenly interested in the mechanics of oceans, atmosphere, and climate disruption, approaching these problems from that of a statistician and from the perspective of dynamical systems [1]. Thus, climate science is an avocation. I have 32 years experience doing quantitative analysis, primarily in industry. Why? It’s great! As the great J. W. Tukey said:

*The best thing about being a statistician is that you get to play in everyone’s backyard.*

Anyone who doubts the fun of doing so, or how statistics enables such, should read Young [2].
2 How Heat Flows and Why It Matters

Heat is most often experienced as energy density, related to temperature. While technically temperature is only meaningful for a body in thermal equilibrium, temperature is the operational definition of heat content, both in daily life and as a scientific measurement, whether at a point or averaged. For the present discussion, it is taken as given that increasing atmospheric concentrations of carbon dioxide trap and re-radiate Earth blackbody radiation to its surface, resulting in a higher mean blackbody equilibration temperature for the planet, via radiative forcing \[3, 4, 5, 6\]. The question is, how does a given Joule of energy travel? Once on Earth, does it remain in atmosphere? Warm the surface? Go into the oceans? And, especially, if it does go into the oceans, what is its residence time before released to atmosphere? These are important questions \[7, 8\]. Because of the miscibility of energy, questions of residence time are very difficult to answer. A Joule of energy can’t be tagged with a radioisotope like matter sometimes can. In practice, energy content is estimated as a constant plus the time integral of energy flux across a well-defined boundary using a baseline moment.

Variability is a key aspect of natural systems, whether biological or large scale geophysical systems such as Earth’s climate \[9\]. Variability is also a feature of statistical models used to describe behavior of natural systems, whether they be straightforward empirical models or models based upon \textit{ab initio} physical calculations. Some of the variability in models captures the variability of the natural systems which they describe, but some variability is inherent in the mechanism of the models, an artificial variability which is not present in the phenomena they describe\[10\]. No doubt, there is always some variability in natural phenomena which no model captures. This variability can be partitioned into parts, at the risk of specifying components which are not directly observable. Sometimes they can be inferred.

Models of planetary climate are both surprisingly robust and understood well enough that appreciable simplifications are possible, such as setting aside fluid dynamism, without damaging their utility \[4, Preface\]. Thus, long term or asymptotic and global predictions of what consequences arise when atmospheric carbon dioxide concentrations double or triple are known pretty well. What is less certain are the dissipation and diffusion mechanisms for this excess energy and its behavior in time \[11, 12, 13, 14\]. There is keen interest in these mechanisms because of the implications differing magnitudes have for regional climate forecasts and economies \[15, 16, 17\]. Moreover, there is a natural desire to obtain empirical confirmation of physical calculations, as difficult as that might be, and as subjective as judgments regarding quality of predictions might be \[18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36\].

Observed rates of surface temperatures in recent decades have shown a moderating slope compared with both long term statistical trends and climate model projections \[40, 41, 18, 42, 29, 21, 24, 43, 19\]. It’s the purpose of this article to present this evidence, and report the research literature’s consensus on where the heat resulting from radiative forcing is going, as well as sketch some implications of that containment.
3 On Surface Temperatures, Land and Ocean

Independently of climate change, monitoring surface temperatures globally is a useful geophysical project. They are accessible, can be measured in a number of ways, permitting calibration and cross-checking, are taken at convenient boundaries, land-and-atmosphere or ocean-and-atmosphere, and coincide with the living space about which we most care. Nevertheless, like any large observational effort in the field, such measurements need careful assessment and processing before they can be properly interpreted. The Berkeley Earth Surface Temperature (“BEST”) Project represents the most comprehensive such effort, but it was not possible without many predecessors, such as HadCRUT4, and works by Kennedy, et al and Rohde [37, 30, 32, 38, 39].

Surface temperature is a manifestation of four interacting processes. First, there is warming of the surface by the atmosphere. Second, there is lateral heating by atmospheric convection and latent heat in water vapor. Third, during daytime, there is warming of the surface by the Sun or insolation which survives reflection. Last, there is warming of the surface from below, either latent heat stored subsurface, or geologic processes. Roughly speaking, these are ordered from most important to least. These are all manifestations of energy flows, a consequence of equalization of different contributions of energy to Earth.

Physically speaking, the total energy of the Earth climate system is a constant plus the time integral of energy of non-reflected insolation less the energy of the long wave radiation or blackbody radiation which passes from Earth out to space, plus geothermal energy ultimately due to radioisotope decay within Earth’s aethenosphere and mantle, plus thermal energy generated by solid Earth and ocean tides, plus waste heat from anthropogenic combustion and power sources [44]. The amount of non-reflected insolation depends upon albedo, which itself slowly varies. The amount of long wave radiation leaving Earth for space depends upon the amount of water aloft, by amounts and types of greenhouse gases, and other factors. Our understanding of this has improved rapidly, as can be seen by contrasting Kiehl, et al in 1997 with Trenberth, et al in 2009 and the IPCC’s 2013 WG1 Report [45, 46, 47]. Steve Easterbrook has given a nice summary of radiative forcing at his blog as well as provided a succinct recap of the 2013 IPCC WG1 Report and its take on energy flows elsewhere at the Azimuth blog. I refer the reader to those references for information about energy budgets, what we know about them, and what we do not.

Some ask whether or not there is a physical science basis for the “moderation” in global surface temperatures and, if there is, how that might work. It is an interesting question, for such a conclusion is predicated upon observed temperature series being calibrated and used correctly, and, further, upon insufficient precision in climate model predictions, whether simply perceived or actual. Hypothetically, it could be that the temperature models are not being used correctly and the models are correct, and which evidence we choose to believe depends upon our short-term goals. Surely, from a scientific perspective, what’s wanted is a reconciliation of both, and that is where many climate scientists invest their efforts. This is also an interesting question because it is, at its root, a statistical one, namely, how do we know which model is better [48, 9, 28, 49, 50, 51, 52]?
A first graph, Figure 1.1 depicting evidence of warming is, to me, quite remarkable. A similar graph is shown in the important series recapping the recent IPCC Report by Steve Easterbrook [53]. A great deal excess heat is going into the oceans. In fact, most of it is. This can happen in many ways, but one dramatic way is due to a phase of the El Niño Southern Oscillation (“ENSO”). Another way is storage by the Atlantic Meridional Overturning Circulation (“AMOC”) [54].

The trade winds along the Pacific equatorial region vary in strength. When they are weak, the phenomenon called El Niño is seen, affecting weather in the United States and in Asia. Evidence for El Niño includes elevated sea-surface temperatures (“SSTs”) in the eastern Pacific. This short-term climate variation brings increased rainfall to the southern United States and Peru, and drought to east Asia.
and Australia, often triggering large wildfires there. The reverse phenomenon, La Niña, is produced by strong trades, and results in cold SSTs in the eastern Pacific, and plentiful rainfall in east Asia and northern Australia. Strong trades actually pile ocean water up against Asia, and these warmer-than-average waters push surface waters there down, creating a cycle of returning cold waters back to the eastern Pacific. This process is depicted in Figures 1.2 and 1.3. At its peak, a La Niña causes waters to accumulate in the Pacific warm pool, and this results in surface heat being pushed into the deep ocean. To the degree to which heat goes into the deep ocean, it is not available in atmosphere. To the degree to which the trades do not pile waters into the Pacific warm pool and, ultimately, into the depths, that warm water is in contact with atmosphere [55]. There are suggestions warm waters at depth rise to the surface [56].

Documentation of land and ocean surface temperatures is done in variety of ways. There are several important sources, including Berkeley Earth, NASA GISS, and the Hadley Centre/Climatic Research Unit (“CRU”) data sets [37, 57, 30]. The three, referenced here as BEST, GISS, and HadCRUT4, respectively have been compared by Rohde [39]. They differ in duration and extent of coverage, but allow comparable inferences. For example, a linear regression establishing a trend using July monthly average temperatures from 1880 to 2012 for Moscow from GISS and BEST agree that Moscow’s July 2010 heat was 3.67 standard deviations from the long term trend [58]. Nevertheless, there is an important difference between BEST and GISS, on the one hand, and HadCRUT4.
Figure 1.3: Trade winds vary in strength, having consequences for pooling and flow of Pacific waters and sea surface temperatures. (Note that this will be replaced on the Web page with a GIF file.)

Figure 1.4: Strong trade winds cause the warm surface waters of the equatorial Pacific to pile up against Asia.

BEST and GISS attempt to capture and convey a single best estimate of temperatures on Earth’s surface, and attach an uncertainty measure to each number. Sometimes, because of absence of measurements or equipment failures, there are no measurements, and these are clearly marked in the series. HadCRUT4 is different. With HadCRUT4 the uncertainty in measurements is described by a hundred member ensemble of values, actually a 2592-by-1967 matrix. Rows correspond to observations from 2592 patches, 36 in latitude, and 72 in longitude, with which it represents the surface of Earth. Columns correspond to each month from January 1850 to November 2013. It is possible for any one of these cells to be coded as “missing”. This detail is important because HadCRUT4 is the basis for a paper suggesting the pause in global warming is structurally inconsistent with climate models. That paper will be discussed later.
4 Rumors of Pause

Figure 1.5 shows the global mean surface temperature anomalies relative to a standard baseline, 1950-1980. Before going on, consider that figure. Study it. What can you see in it?

Figure 1.6 shows the same graph, but now with two trendlines obtained by applying a smoothing spline, one smoothing more than another. One of the two indicates an uninterrupted uptrend. The other shows a peak and a downtrend, along with wiggles around the other trendline. Note the smoothing algorithm is the same in both cases, differing only in the setting of a smoothing parameter [59]. Which is correct? What is “correct”?

Figure 1.1 shows trend curves for ocean heat content over roughly the same period. Figure 1.7 shows a time series of anomalies for Moscow, in Russia. Do these all show the same trends? These are difficult questions, but the changes seen in Figure 1.6 could be evidence of a warming “hiatus” [60, 40, 61]. Note that, given Figure 1.6 the most which can be said about it is that there is a reduction in the rate of temperature increase. We’ll have a more careful look at this in Section 5. With that said, people have sought reasons and assessments of how important this phenomenon is. The answers have ranged from
the conclusive “Global warming has stopped” to “Perhaps the slowdown is due to ’natural variability’”, to “Perhaps it’s all due to ’natural variability’” to “There is no statistically significant change”. Let’s see what some of the perspectives are.

It is hard to find a scientific paper which advances the proposal that climate might be or might have been cooling in recent history. The earliest I can find are repeated presentations by a single geologist in the proceedings of the Geological Society of America, a conference which, like many, gives papers limited peer review [62, 62, 63, 64, 65, 66, 67, 68]. It is difficult to comment on this work since their full methods are not available for review. The content of the abstracts appear to ignore the possibility of lagged response in any physical system.

These claims were summarized by Easterling and Wehner in 2009, attributing claims of a “pause” to cherry-picking of sections of the temperature time series, such as 1998-2008, and what might be called media amplification [60]. Further, technical inconsistencies within the scientific enterprise, perfectly normal in its deployment and management of new methods of measurement, have been highlighted and abused to parlay claims of global cooling [69, 70, 71]. Based upon subsequent papers, climate science seemed to not only need to explain such variability is to be expected, but to provide a specific explanation for what could be seen as a recent moderation in the abrupt warming of the mid-late 1990s, appealing to oceanic capture, as described in Section [3, 55, 29, 40]. The reader should note that, given the overall
temperature anomaly series, such as Figure 1.6 and specific series, such as the one for Moscow in Figure 1.7 moderation in warming is not definitive. It is a statistical question, and, pretending for the moment we knew nothing of geophysics, a difficult one.

But there certainly is no any problem with accounting for the Earth’s energy budget overall, even if one grants the distribution of energy over its surface cannot be specifically explained [45, 46, 4]. This is not a surprise, since the equipartition theorem of physics fails to apply to a system which has not achieved thermal equilibrium.

An interesting discrepancy is presented in a pair of papers in 2013 and 2014. The first, by Fyfe, Gillet, and Zwiers, has the (somewhat provocative) title “Overestimated global warming over the past 20 years” [24, 43]. It has been followed by another paper by Fyfe and Gillet applying the same methods to argue that even with the Pacific surface temperature anomalies and accommodating the coverage bias in the HadCRUT4 dataset, as emphasized by Cowtan and Way, there are discrepancies between the surface temperature record and climate model ensemble runs [72, 23, 41]. How this pair of papers presents that challenge and its possible significance is a story of trends, of variability, and hopefully of what these investigations are saying in common.
Trends as a concept are easy. But trends as objective measures are slippery. Consider the Keeling Curve, the record of atmospheric carbon dioxide concentration first begun by Charles Keeling in the 1950s and continued in the face of great obstacles [73]. This curve is reproduced in Figure 1.8 and there presented in its original, and then decomposed into three parts, an annual sinusoidal variation, a linear trend, and a stochastic remainder. The question is, which component represents the true trend, long term or otherwise? Are linear trends superior to all others? The importance of a trend is tied up with to what use it will be put. A pair of trends, like the sinusoidal and the random residual of the Keeling, might be more important for predicting its short term movements. On the other hand, explicating the long term behavior of the system being measured might feature the large scale linear trend, with the seasonal trend and random variations being but distractions.

Consider the global surface temperature anomalies of Figure 1.5 again. What are some ways of determining trends? First, note that by “trends” what’s really meant are slopes. In the case where there are many possible points of estimating slopes, there are many slopes. When, for example, a slope is estimated from
fitting a line to all the points, there’s just a single slope. Apart from a single long term trend, such as Figure 1.9, local linear trends can be estimated from pairs of points in differing sizes of neighborhoods, as depicted in Figure 1.11. These can be averaged, if you like, to obtain an overall trend. Lest the reader think constructing lots of linear trends on varying neighborhoods is somehow crude, note it has a noble history, being used by Boscovich to estimate Earth’s ellipticity about 1750, as reported by Koenker [79, Section 1.2].

There is, in addition, a question of what to do if local intervals for fitting the little lines overlap, since these are not independent of one another. There are a number of statistical devices for making them independent. One way is to do clever kinds of random sampling from a population of linear trends. Another way is to shrink the intervals until they are infinitesimally small, and, so, necessarily independent. That’s just the point slope of a curve going through the data, or its first derivative. Numerical methods exist of estimating these, one involving a smoothing spline, as was sketched in Figure 1.6 and estimating the derivative(s) of that [74]. Such an estimate of the derivative is shown in Figure 1.12 where the instantaneous slope is plotted in orange atop the data of Figure 1.5. The value of the derivative should be read using the scale to the right of the graph. The value to the left shows, as before, temperature anomaly in degrees. The spline itself is plotted in green in that figure. It is a cubic spline and the smoothing parameter is determined by generalized cross-validation, something explained a bit more in the caption for Figure 1.12 [76].

What else might we do? We could go after a really good approximation to the data of Figure 1.5. One possibility is to use the Bayesian Rauch-Tung-Striebel (“RTS”) smoother to get a good approximation for the underlying curve and estimate the derivatives of that [77]. This is a modification of the famous Kalman filter, the workhorse of much controls engineering and signals work [77, 78, 80]. What that means and how these work is described in an accompanying inset box.

Using the RTS smoother demands prior variances of the signal are estimated. The larger the ratio of the estimate of the observations variance to the estimate of the process variance is, the smoother the RTS solution [81]. The RTS smoother result for two variance values of 0.156 and high 0.312 is shown in Figure 1.13. These are 3 and 15 times the decorrelated variance value for the series of 0.118, estimated separately.

Combining all six methods of estimating trends results in Figure 1.14 which shows the overprinted densities of slopes.

Note the spread of possibilities given by the 5 year local linear fits. The 10 year local linear fits, the spline, and the RTS smoother fits have their mode in the vicinity of the overall slope. The 10 year local linear fits slope has broader support, meaning it admits more negative slopes in the range of temperature anomalies observed. The RTS smoother results have peaks slightly below those for the spline, the 10 year local linear fits, and the overall slope. The kernel density estimator allows the possibility of probability mass below zero, even though the spline, and two RTS smoother fits never exhibit slopes below zero. This is a Bayesian-like estimator, since the prior is the real line.
Local linear fits to HadCRUT4 time series were used by Fyfe, Gillet, and Zwiers in their 2013 paper [24, 43]. We do not know the computational details of those trends, since they were not published. *Those details matter.* From these calculations, which, admittedly, are not as comprehensive as those by Fyfe, Gillet, and Zwiers, we see that robust estimators of trends in temperature during the observational record show these are always positive, even if the magnitudes vary. The RTS smoother solutions suggest slopes in recent years are near zero, providing a basis for questioning whether or not there is a warming “hiatus”.
The Rauch-Tung-Striebel smoother is an enhancement of the Kalman filter [77]. Let \( y_\kappa \) denote a set of univariate observations at equally space and successive time steps \( \kappa \). Describe these as follows:

\[
\begin{align*}
(1.1) & \quad y_\kappa = G x_\kappa + \varepsilon_\kappa \\
(1.2) & \quad x_{\kappa+1} = H x_\kappa + J_\kappa \\
(1.3) & \quad \varepsilon_\kappa \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\
(1.4) & \quad J_\kappa \sim \mathcal{N}(0, \Sigma_\eta^2)
\end{align*}
\]

The multivariate \( x_\kappa \) is called a state vector for index \( \kappa \). \( G \) and \( H \) are given, constant matrices. Equations (1.3) and (1.4) say that the noise component of observations and states are distributed as zero mean Gaussian random variables with variance \( \sigma_\varepsilon^2 \) and covariance \( \Sigma_\eta^2 \), respectively. This simple formulation in practice has great descriptive power, and is widely used in engineering and data analysis. For instance, it is possible to cast autoregressive moving average models (“ARMA”) in this form [129] Chapter 10. The key idea is that equation (1.1) describes observation at time \( \kappa \) as the result of a linear regression on coefficients \( x_\kappa \), where \( G \) is the corresponding design matrix.

Then, the coefficients themselves change with time, using a Markov-like development, a linear regression of the upcoming set of coefficients, \( x_{\kappa+1} \), in terms of the current coefficients, \( x_\kappa \), where \( H \) is the design matrix.

For the purposes here, an exceptionally simple version of this is used, something called a local level model and occasionally a Gaussian random walk with noise model (see [80, Chapter 2] and [78, Section 12.3.1]). In that instance, \( G \) and \( H \) are not only scalars, they are unity, resulting in the simpler

\[
\begin{align*}
(1.5) & \quad y_\kappa = x_\kappa + \varepsilon_\kappa \\
(1.6) & \quad x_{\kappa+1} = x_\kappa + \eta_\kappa \\
(1.7) & \quad \varepsilon_\kappa \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\
(1.8) & \quad \eta_\kappa \sim \mathcal{N}(0, \sigma_\eta^2)
\end{align*}
\]

with scalar variance \( \sigma_\eta^2 \).

In either case, the Kalman filter is a way of calculating \( x_\kappa \), given \( y_1, y_2, \ldots, y_n \), values for \( G \) and \( H \), and estimates for \( \sigma_\varepsilon^2 \) and \( \sigma_\eta^2 \). Choices for \( G \) and \( H \) are considered a model for the data. Choices for \( \sigma_\varepsilon^2 \) and \( \sigma_\eta^2 \) are based upon experience with \( Y_\kappa \) and the model.

In practice, and within limits, the bigger the ratio

\[
\frac{\sigma_\varepsilon^2}{\sigma_\eta^2}
\]

the smoother the solution for \( x_\kappa \) over successive \( \kappa \).

Now, the Rauch-Tung-Striebel extension of the Kalman filter amounts to (a) interpreting it in a Bayesian context, and (b) using that interpretation and Bayes Rule to retrospectively update \( x_{\kappa-1}, x_{\kappa-2}, \ldots, x_1 \) with the benefit of information through \( y_\kappa \) and the current state \( x_\kappa \). Details won’t be provided here, but are described in depth in many texts [78, 80, 77].
Figure 1.9: Global surface temperature anomalies relative to a 1950-1980 baseline, with long term linear trend atop.
Figure 1.10: Global surface temperature anomalies relative to a 1950-1980 baseline, with randomly placed trends from local linear having 5 year support atop.
Figure 1.11: Global surface temperature anomalies relative to a 1950-1980 baseline, with randomly placed trends from local linear having 10 year support atop.
Figure 1.12: Global surface temperature anomalies relative to a 1950-1980 baseline, with instantaneous numerical estimates of derivatives in orange atop, with scale for the derivative to the right of the chart. Note how the value of the first derivative never drops below zero although its magnitude decreases as time approaches 2012. Support for the smoothing spline used to calculate the derivatives is obtained using generalized cross validation [74,75]. Such cross validation is used to help reduce the possibility that a smoothing parameter is chosen to overfit a particular data set, so the analyst could expect that the spline would apply to as yet uncollected data more than otherwise. Generalized cross validation is a particular clever way of doing that, although it is abstract.
Figure 1.13: Global surface temperature anomalies relative to a 1950-1980 baseline, with fits using the Rauch-Tung-Striebel smoother placed atop, in green and dark green. The former uses a prior variance of 3 times that of the Figure 1.5 data corrected for serial correlation. The latter uses a prior variance of 15 times that of the Figure 1.5 data corrected for serial correlation. The instantaneous numerical estimates of the first derivative derived from the two solutions are shown in orange and brown, respectively, with their scale of values on the right hand side of the chart. Note the two solutions are essentially identical. If compared to the smoothing spline estimate of Figure 1.12, the derivative has roughly the same shape, but is shifted lower in overall slope, and the drift up and below a mean value is less. The process variance is, in both cases, $\frac{1}{50}$ of the observational variance.
Figure 1.14: Empirical probability density functions for slopes of temperatures versus years, from each of 6 methods. Empirical probability densities are obtained using kernel density estimation and are preferred to histograms by statisticians because the latter can distort the density due to bin size and boundary effects. Lines correspond to local linear fits with 5 years separation (dark green trace), the local linear fits with 10 years separation (green trace), the smoothing spline (blue trace), the RTS smoother with variance 3 times the corrected estimate for the data as the prior variance (orange trace, mostly hidden by brown trace), and the RTS smoother with 15 times the corrected estimate for the data (brown trace). The slope value for a linear fit to all the points is also shown (the vertical black line).
6 Internal Decadal Variability

The recent IPCC AR5 WG1 Report sets out the context [47, Box TS.3]:

Hiatus periods of 10 to 15 years can arise as a manifestation of internal decadal climate variability, which sometimes enhances and sometimes counteracts the long-term externally forced trend. Internal variability thus diminishes the relevance of trends over periods as short as 10 to 15 years for long-term climate change (Box 2.2, Section 2.4.3). Furthermore, the timing of internal decadal climate variability is not expected to be matched by the CMIP5 historical simulations, owing to the predictability horizon of at most 10 to 20 years (Section 11.2.2; CMIP5 historical simulations are typically started around nominally 1850 from a control run). However, climate models exhibit individual decades of GMST trend hiatus even during a prolonged phase of energy uptake of the climate system (e.g., Figure 9.8; Easterling and Wehner, 2009; Knight et al., 2009), in which case the energy budget would be balanced by increasing subsurface-ocean heat uptake (Meehl et al., 2011, 2013a; Guemas et al., 2013).

Owing to sampling limitations, it is uncertain whether an increase in the rate of subsurface-ocean heat uptake occurred during the past 15 years (Section 3.2.4). However, it is very likely that the climate system, including the ocean below 700 m depth, has continued to accumulate energy over the period 1998-2010 (Section 3.2.4, Box 3.1) [82]. Consistent with this energy accumulation, global mean sea level has continued to rise during 1998-2012, at a rate only slightly and insignificantly lower than during 1993-2012 (Section 3.7). The consistency between observed heat-content and sea level changes yields high confidence in the assessment of continued ocean energy accumulation, which is in turn consistent with the positive radiative imbalance of the climate system (Section 8.5.1; Section 13.3, Box 13.1). By contrast, there is limited evidence that the hiatus in GMST trend has been accompanied by a slower rate of increase in ocean heat content over the depth range 0 to 700 m, when comparing the period 2003-2010 against 1971-2010. There is low agreement on this slowdown, since three of five analyses show a slowdown in the rate of increase while the other two show the increase continuing unabated (Section 3.2.3, Figure 3.2).

During the 15-year period beginning in 1998, the ensemble of HadCRUT4 GMST trends lies below almost all model-simulated trends (Box 9.2 Figure 1a), whereas during the 15-year period ending in 1998, it lies above 93 out of 114 modelled trends (Box 9.2 Figure 1b; HadCRUT4 ensemble-mean trend 0.26 °C per decade, CMIP5 ensemble-mean trend 0.16 °C per decade). Over the 62-year period 1951-2012, observed and CMIP5 ensemble-mean trends agree to within 0.02 °C per decade (Box 9.2 Figure 1c; CMIP5 ensemble-mean trend 0.13 °C per decade). There is hence very high confidence that the CMIP5 models show long-term GMST trends consistent with observations, despite the disagreement over the most recent 15-year period. Due to internal climate variability, in any given 15-year period the observed
GMST trend sometimes lies near one end of a model ensemble (Box 9.2, Figure 1a, b; Easterling and Wehner, 2009), an effect that is pronounced in Box 9.2, Figure 1a, because GMST was influenced by a very strong El Niño event in 1998.

The contributions of Fyfe, Gillet, and Zwiers (“FGZ”) are to (a) pin down this behavior for a 20 year period using the HadCRUT4 data, and, to my mind, more importantly, (b) to develop techniques for evaluating runs of ensembles of climate models like the CMIP5 suite without commissioning specific runs for the purpose [24, 41]. This, if it were to prove out, would be an important experimental advance, since climate models demand expensive and extensive hardware, and the number of people who know how to program and run them is very limited, possibly a more limiting practical constraint than the hardware [83, 84].

This is the beginning of a great story, I think, one which both advances an understanding of how our experience of climate is playing out, and how climate science is advancing. FGZ took a perfectly reasonable approach and followed it to its logical conclusion, deriving an inconsistency. There’s insight to be won resolving it.

FGZ try to explicitly model trends due to internal variability [43]. They begin with two equations:

\[
M_{ij}(t) = u^m(t) + E_{int_{ij}}(t) + E_{mod_i}(t), \quad i = 1, \ldots, N^m, \quad j = 1, \ldots, N_i
\]

\[
O_k(t) = u^o(t) + E_{int^o}(t) + E_{samp_k}(t), \quad k = 1, \ldots, N^o
\]

\(i\) is the model membership index. \(j\) is the index of the \(i^{th}\) model’s \(j^{th}\) ensemble. \(k\) runs over the bootstrap samples taken from HadCRUT4 observations. Here, \(M_{ij}(t)\) and \(O_k(t)\) are trends calculated using models or observations, respectively. \(u^m(t)\) and \(u^o(t)\) denote the “true, unknown, deterministic trends due to external forcing” common to models and observations, respectively [43]. \(E_{int_{ij}}(t)\) and \(E_{int^o}(t)\) are the perturbations to trends due to internal variability of models and observations. \(E_{mod_i}(t)\) denotes error in climate model trends for model \(i\). \(E_{samp_k}(t)\) denotes the sampling error in the \(k^{th}\) sample. Notably FGZ assume \(E_{mod_i}(t)\) are exchangeable with each other as well, at least for the same time \(t\) [85, 86, 87, 88, 89]. Note that while the internal variability of climate models \(E_{int_{ij}}(t)\) varies from model to model, run to run, and time to time, the ‘internal variability of observations’, namely \(E_{int^o}(t)\), is assumed to only vary with time.

The technical innovation FGZ use is to employ bootstrap resampling on the observations ensemble of HadCRUT4 and an ensemble of runs of 38 CMIP5 climate models to perform what is essentially a two-sample comparison [90, 91]. In doing so, they explicitly assume, in the framework above, exchangeability of models [85].

So, what is a bootstrap? In its simplest form, a bootstrap is a nonparametric, often robust, frequentist technique for sampling the distribution of a function of a set of population parameters, generally irrespective of the nature or complexity of that function, or the number of parameters. Since estimates of the variance of that function are themselves functions of population parameters, assuming the variance exists,
the bootstrap can also be used to estimate the precision of the first set of samples, where “precision” is
the reciprocal of variance. More about the bootstrap is described in an inset.

In the case in question here, with FGZ, the bootstrap is being used to determine if the distribution of
surface temperature trends as calculated from observations and the distribution of surface temperature
trends as calculated from climate models for the same period have in fact similar means. This is done
by examining differences of paired trends, one coming from an observation sample, one coming from
a model sample, and assessing the degree of discrepancy based upon the variances of the observations
trends distribution and of the models trends distribution.

The equations (1.10) and (1.11) can be re-written:

\[
M_{ij}(t) - \text{Eint}_{ij}(t) = u^m_i(t) + \text{Emod}_i(t), i = 1, \ldots, N^m_i, j = 1, \ldots, N_i
\]

(1.12)

\[
O_k(t) - \text{Eint}^o(t) = u^o(t) + \text{Esamp}_k(t), k = 1, \ldots, N^o
\]

(1.13)

moving the trends in internal variability to the left, calculated side. Both \( \text{Eint}_{ij}(t) \) and \( \text{Eint}^o(t) \) are not
directly observable. Without some additional assumptions, which are not explicitly given in the FGZ
paper, such as

\[
\text{Eint}_{ij}(t) \sim N(0, \Sigma_{\text{model int}})
\]

(1.14)

\[
\text{Eint}^o(t) \sim N(0, \Sigma_{\text{obs int}})
\]

(1.15)

we can’t really be sure we’re seeing \( O_k(t) \) or \( O_k(t) - \text{Eint}^o(t) \), or at least \( O_k(t) \) less the mean of \( \text{Eint}^o(t) \).
The same applies to \( M_{ij}(t) \) and \( \text{Eint}_{ij}(t) \). Here equations [1.14] and [1.15] describe internal variabilities as
being multivariate but zero mean Gaussian random variables. \( \Sigma_{\text{model int}} \) and \( \Sigma_{\text{obs int}} \) are covariances
among models and among observations. FGZ essentially say these are diagonal with their statement “An
implicit assumption is that sampling uncertainty in [observation trends] is independent of uncertainty
due to internal variability and also independent of uncertainty in [model trends]” [43]. They might not
be so, but it is reasonable to suppose their diagonals are strong, and that there is a row-column exchange
operator on these covariances which can produce banded matrices.
7 On Reconciliation

The centerpiece of the FGZ result is their Figure 1, reproduced here as Figure 1.15. Their conclusion, that climate models do not properly capture surface temperature observations for the given periods, is based upon the significant separation of the red density from the grey density, even measuring that separation using pooled variances. But, surely, a remarkable feature of these graphs is not only the separation of the means of the two densities, but the marked difference between the variance of the two densities. Why are climate models so less precise than HadCRUT4 observations? Moreover, why do climate models disagree with one another so dramatically? We cannot tell without getting into CMIP5 details, but the same result could be obtained if the climate models came in three Gaussian populations, each with a variance 1.5x that of the observations, but mixed together. We could also obtain the same result if, for some reason, the variance of the HadCRUT4 was markedly understated.

That brings us back to the comments about HadCRUT4 made at the end of Section 3. HadCRUT4 is noted for “drop outs” in observations, where either the quality of an observation on a patch of Earth was poor or the observation was missing altogether for a certain month in history. It also has incomplete coverage [23]. Whether or not values for patches are imputed in some way, perhaps using spatial kriging, or supports to calculate trends are adjusted to avoid these holes, is an important question [23, 96]. As seen in Section 5, what trends you get depends a lot on how they are done. FGZ did linear trends. These
are nice because means of trends have simple relationships with the trends themselves. On the other hand, confining trend estimation to local linear trends binds these estimates to being only supported by pairs of actual samples, however sparse they may be. This has the unfortunate effect of producing a broadly spaced set of trends which, when averaged, appear to be a single, tight distribution, close to the vertical black line of Figure [1.14, and erasing all the detail available by estimating the density of trends with a robust function of the time series’ first derivative. FGZ might be improved by using such, repairing this drawback and also making it more robust against HadCRUT4’s inescapable data drops. As mentioned before, however, we really cannot know, because details of their calculations are not available.

In fact, that was indicated by a recent paper from Cowtan and Way, arguing that the limited coverage of HadCRUT4 might explain the discrepancy Fyfe, Gillet, and Zwiers found [23][24, 43]. In return Fyfe and Gillet argue that even admitting the corrections for polar regions which Cowtan and Way indicate, the CMIP5 models fall short in accounting for global mean surface temperatures [41]. What could be wrong?

While this has not yet been verified, I suspect there may be a shortcoming in the way which the ensemble of 100 members of HadCRUT4 studies are being used which might explain [30]. In the context of ensemble forecasts depicting future states of the atmosphere, Wilks notes [92, Section 7.7.1]

Accordingly, the dispersion of a forecast ensemble can at best only approximate the [probability density function] of forecast uncertainty . . . . In particular, a forecast ensemble may reflect errors both in statistical location (most or all ensemble members being well away from the actual state of the atmosphere, but relatively nearer to each other) and dispersion (either under- or overrepresenting the forecast uncertainty). Often, operational ensemble forecasts are found to exhibit too little dispersion . . . , which leads to overconfidence in probability assessment if ensemble relative frequencies are interpreted as estimating probabilities.

Wilks and, in fact, the IPCC reference Buizza, Toth, and others on this point [92][47][93][94]. It could be that the answer to why the variance of the observational data in the Fyfe, Gillet, and Zwiers graph depicted in Figure [1.15 is so small is that ensemble spread does not properly reflect the true probability density function of the joint distribution of temperatures across Earth. These might be “relatively nearer to each other” than the true dispersion which climate models are accommodating.

If Earth’s climate is thought of as a dynamical system, and taking note of the suggestion of Kharin that “There is basically one observational record in climate research”, we can do the following thought experiment [33]. Suppose the total state of the Earth’s climate system can be captured at one moment in time, no matter how, and the climate can be reinitialized to that state at our whim, again no matter how. What happens if this is done several times, and then the climate is permitted to develop for, say, exactly 100 years on each “run”? What are the resulting states? Suppose the dynamical “inputs” from the Sun, as a function of time, are held identical during that 100 years, as are dynamical inputs from volcanic forcings, as are human emissions of greenhouse gases. Are the resulting states copies of one another? No.
The bootstrap is a general name for a resampling technique, most commonly associated with what is more properly called the frequentist bootstrap. Given a sample of observations, \( \hat{Y} = \{y_1, y_2, \ldots, y_n\} \), the bootstrap principle says that in a wide class of statistics and for certain minimum sizes of \( n \), the sampling density of a statistic \( h(Y) \) from a population of all \( Y \), where \( \hat{Y} \) is a single observation, can be approximated by the following procedure. Sample \( \hat{Y} \) \( M \) times with replacement to obtain \( M \) samples each of size \( n \) called \( \hat{Y}_k, k = 1, \ldots, M \). For each \( \hat{Y}_k \), calculate \( h(\hat{Y}_k) \) so as to obtain \( \hat{H} = h_1, h_2, \ldots, h_M \). The set \( \hat{H} \) so obtained is an approximation of the sampling density of \( h(Y) \) from a population of all \( Y \). Note that because \( \hat{Y} \) is sampled, only elements of that original set of observations will ever show up in any \( \hat{Y}_k \). This is true even if \( Y \) is drawn from an interval of the real numbers. This is where a Bayesian bootstrap might be more suitable.

In a Bayesian bootstrap, the set of possibilities to be sampled are specified using a prior distribution on \( Y \) [91, Section 10.5]. A specific observation of \( Y \), like \( \hat{Y} \), is use to update the probability density on \( Y \), and then values from \( Y \) are drawn in proportion to this updated probability. Thus, values in \( Y \) never in \( \hat{Y} \) might be drawn. Both bootstraps will, under similar conditions, preserve the sampling distribution of \( Y \).

Stochastic variability in the operation of climate means these end states will be each somewhat different than one another. Then of what use is the “one observation record”? Well, it is arguably better than no observational record.

Setting aside the problems of using local linear trends, FGZ’s bootstrap approach to the HadCRUT4 ensemble is an attempt to imitate these various runs of Earth’s climate. The trouble is, the frequentist bootstrap can only replicate values of observations actually seen. (See inset.) In this case, these replications are those of the HadCRUT4 ensembles. It will never produce values in-between and, as the parameters of temperature anomalies are in general continuous measures, allowing for in-between values seems a reasonable thing to do.

Even if it does, however, in doing so, no algorithm can account for a dispersion which is not reflected in the variability of the ensemble. If the dispersion of HadCRUT4 is too small, it could be corrected using ensemble MOS methods [92, Section 7.7.1]. In any case, underdispersion could explain the remarkable difference in variances of populations seen in Figure 1.15. But I think there’s another way.

Consider equations (1.10) and (1.11) again. Recall, here, \( i \) denotes the \( i \)th model and \( j \) denotes the \( j \)th run of model \( i \). Instead of \( k \), however, a bootstrap resampling of the HadCRUT4 ensembles, let \( \ell \) run over all the 100 ensemble members provided, let \( j \) run over the 2592 patches on Earth’s surface, and let \( \kappa \) run over the 1967 monthly time steps. Reformulate equations (1.10) and (1.11), instead, as

\[
M_\kappa = u_\kappa + \sum_{i=1}^{N^m} x_i (\text{Emod}_{i\kappa} + \text{Eint}_{i\kappa})
\]

\[
O_\kappa = u_\kappa + x_0 \text{Eint}_0^\kappa + \sum_{j=1}^{2592} x_j \text{Esamp}_{j\kappa}
\]

Now, \( u_\kappa \) is a common trend at time tick \( \kappa \) and \( \text{Emod}_{i\kappa} \) and \( \text{Eint}_{i\kappa} \) are deflections from from that trend due to modeling error in the \( i \)th model and due to internal variability represented in the \( i \)th model, respectively, at time tick \( \kappa \). Similarly, \( \text{Eint}_0^\kappa \) denotes a deflections from the common trend baseline \( u \) due
to internal variability as seen by the HadCRUT4 observational data at time tick $\kappa$, and $E_{\text{sample}}_{jk}$ denotes the deflection from the common baseline due to sampling error in the $j^{th}$ patch at time tick $\kappa$. $x_i$ are indicator variables. This is the setup for an analysis of variance or ANOVA \[113\]. Sections 14.1.6, 18.1].

In equation (1.16), successive model runs $j$ for model $i$ are used to estimate $E_{\text{model}}_{i\kappa}$ and $E_{\text{int}}_{i\kappa}$ for every $\kappa$. In equation (1.17), different ensemble members $\ell$ are used to estimate $E_{\text{int}}^\ell_{\kappa}$ and $E_{\text{sample}}_{j\kappa}$ for every $\kappa$.

Coupling the two gives a common estimate of $u_{k\ell}$. There’s considerable flexibility in how model runs or ensemble members are used for this purpose, including incorporating information about relationships among models in the first case, perhaps relative to a Bayesian model average, or spatial dependencies among $j$ patches\[128\]. It’s entirely plausible there’s an underlying trend $u_k$ common to both estimates of temperature change, and this formulation allocates deviations from that to interpretable and comparable components.

More work needs to be done to assess the proper virtues of the FGZ technique, even without modification. A device like that Rohde used to compare BEST temperature observations with HadCRUT4 and GISS, one of supplying the FGZ procedure with synthetic data, would be perhaps the most informative regarding its character \[39\]. Alternatively, if an ensemble MOS method were devised and applied to HadCRUT4, it might better reflect a true spread of possibilities. Because a dataset like HadCRUT4 records just one of many possible observational records the Earth might have exhibited, it would be useful to have a means of elaborating what those other possibilities were, given the single observational trace.

Regarding climate models, while they will inevitably disagree from a properly elaborated set of observations in the particulars of their statistics, in my opinion, the goal should be to strive to match the distributions of solutions these two instruments of study on their first few moments by improving both. While, statistical equivalence is all that’s sought, we’re not there yet. Assessing parametric uncertainty of observations hand-in-hand with the model builders seems to be a sensible route \[95\]. Indeed, this is important. In review of the Cowtan and Way result, one based upon kriging, Kintisch summarizes the situation as reproduced in Table 1.1, a reproduction of his table on page 348 of the reference \[23, 96, 97\]:

Note that these estimates of trends, once divided by 10 years/decade to convert to a per year change in temperature, all fall well within the slope estimates depicted in the summary Figure 1.14. Note, too, how low the HadCRUT trend is.

If the FGZ technique, or any other, can contribute to this process, it is most welcome.

<table>
<thead>
<tr>
<th>Source</th>
<th>Warming ($^\circ$ C/decade)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate models</td>
<td>0.102-0.412</td>
</tr>
<tr>
<td>NASA data set</td>
<td>0.080</td>
</tr>
<tr>
<td>HadCRUT data set</td>
<td>0.046</td>
</tr>
<tr>
<td>Cowtan/Way</td>
<td>0.119</td>
</tr>
</tbody>
</table>

Table 1.1: Getting warmer. New method brings measured temperatures closer to projections. Added in quotation: “Climate models” refers to the CMIP5 series. “NASA data set” is GISS. “HadCRUT data set” is HadCRUT4. “Cowtan/Way” is from their paper \[23\]. Note values are per decade, not per year.
As an example Lee reports how the GLOMAP model of aerosols was systematically improved using such careful statistical consideration [99]. It seems likely to be a more rewarding way than “black box” treatments. Incidentally, Dr Lindsay Lee’s article was runner-up in the *Significance*/Young Statisticians Section writers’ competition. It’s great to see bright young minds charging in to solve these problems.
8 Summary

Various geophysical datasets recording global surface temperature anomalies suggest a slowdown in anomalous global warming from historical baselines. Warming is increasing, but not as fast, and much of the media attention to this is reacting to the second time derivative of temperature, which is negative, not the first time derivative, its rate of increase. Explanations vary. In one important respect, 20 or 30 years is an insufficiently long time to assess the state of the climate system. In another, while the global surface temperature increase is slowing, oceanic temperatures continue to soar, at many depths. Warming might even decrease. None of these seem to pose a challenge to the geophysics of climate, which has substantial support both from experimental science and ab initio calculations. An interesting discrepancy is noted by Fyfe, Gillet, and Zwiers, although their calculation could be improved both by using a more robust estimator for trends, and by trying to integrate out anomalous temperatures due to internal variability in their models, because much of it is not separately observable. Nevertheless, Fyfe, Gillet, and Zwiers may have done the field a great service, making explicit a discrepancy which enables students of datasets like the important HadCRUT4 to discover an important limitation, that their dispersion across ensembles does not properly reflect the set of Earth futures which one might wish they did and, in their failure for users who think of the ensemble as representing such futures, give them a dispersion which is significantly smaller than what we might know.

In summary, working out these details is the process of science at its best, and many disciplines, not least mathematics, statistics, and signal processing, have much to contribute to the methods and interpretations of these series data. It is possible too much is being asked of a limited data set, and we have not yet observed enough of climate system response to tell anything definitive [100]. But the urgency to act responsibly given scientific predictions remains.
Bibliography

[1] I have taken courses in geology from Binghamton University, but the rest of my knowledge of climate science is from reading the technical literature, principally publications from the American Geophysical Union and the American Meteorological Society, and self-teaching, from textbooks like Pierrehumbert [4]. I seek to find ways where my different perspective on things can help advance and explain the climate science enterprise. I also apply my skills to working local environmental problems, ranging from inferring people’s use of energy in local municipalities, as well as studying things like trends in solid waste production at the same scales using Bayesian inversions. I am fortunate that techniques used in my professional work and those in these problems overlap so much. I am a member of the American Statistical Association, the American Geophysical Union, the American Meteorological Association, the International Society for Bayesian Analysis, as well as the IEEE and its signal processing society.


[10] The nomenclature can be confusing. With respect to observations, variability arising due to choice of method is sometimes called structural uncertainty [30, 101].


[41] J. C. Fyfe, N. P. Gillett, “Recent observed and simulated warming”, *Nature Climate Change*, 4, March 2014, 150-151, [http://dx.doi.org/10.1038/nclimate2111](http://dx.doi.org/10.1038/nclimate2111).


[44] (There are tiny amounts of heating due to impinging ionizing radiation from space, and changes in Earth’s magnetic field.


[58] 3.667 (GISS) versus 3.670 (BEST).

[59] The smoothing parameter is a constant which weights a penalty term proportional to the second directional derivative of the curve. The effect is that if a candidate spline is chosen which is very bumpy, this candidate is penalized and will only be chosen if the data demands it. There is more said about choice of such parameters in the caption of Figure 1.12.


[61] The term hiatus has a formal meaning in climate science, as described by the IPCC itself [47, Box TS.3].


[81] Here, the process variance was taken here to be $\frac{1}{50}$ of the observations variance.

[82] “In this Report, the following terms have been used to indicate the assessed likelihood of an outcome or a result: Virtually certain 99–100% probability, Very likely 90–100%, Likely 66–100%, About as likely as not 33–66%, Unlikely 0-33%, Very unlikely 0-10%, Exceptionally unlikely 0-1%. Additional terms (Extremely likely: 95–100%, More likely than not > 50–100%, and Extremely unlikely 0-5%) may also be used when appropriate. Assessed likelihood is typeset in italics, e.g., verylikely (see Section 1.4 and Box TS.1 for more details).”


[84] “‘It’s great there’s a new initiative,” says modeler Inez Fung of DOE’s Lawrence Berkeley National Laboratory and the University of California, Berkeley. ’But all the modeling efforts are very short-handed. More brains working on one set of code would be better than working separately’” [83].
[85] *Exchangeability* is a weaker assumption than *independence*. Random variables are *exchangeable* if their joint distribution only depends upon the set of variables, and not their order [86, 87, 88]. Note the caution in Coolen [89].


[89] F. P. A. Coolen, “On nonparametric predictive inference and objective Bayesianism”, *Journal of Logic, Language and Information*, 15, 2006, 21-47, [http://dx.doi.org/10.1007/s10849-005-9005-7](http://dx.doi.org/10.1007/s10849-005-9005-7) (“Generally, though, both for frequentist and Bayesian approaches, statisticians are often happy to assume exchangeability at the prior stage. Once data are used in combination with model assumptions, exchangeability no longer holds post-data due to the influence of modelling assumptions, which effectively are based on mostly subjective input added to the information from the data.”).


