AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE

10.1002/2014JD022805

This article is a companion to *Outten et al.* [2015] doi:10.1002/2015JD023859.

Key Points:

- NorESM suggests forcing disparities are unimportant in explaining the hiatus
- Ensembles of runs capture many salient features of the surface temperatures
- NorESM cannot be rejected as a plausible model capable of explaining hiatus

Supporting Information:

- Readme
- Figure S1
- Figure S2
- Figure S3
- Figure S4
- Figure S5
- Figure S6
- Figure S7
- Figure S8
- Figure S9
- Figure S10
- Figure S11
- Figure S12
- Figure S13
- Figure S14
- Figure S15
- Figure S16
- Figure S17
- Figure S18
- Figure S19

Correspondence to:
P. Thorne,
Peter@peter-thorne.net

Citation:
Thorne, P., S. Outten, I. Bethke, and Ø. Seland (2015), Investigating the recent apparent hiatus in surface temperature increases: 2. Comparison of model ensembles to observational estimates, *J. Geophys. Res. Atmos.*, *120*, 8597–8620, doi:10.1002/2014JD022805.

Received 6 NOV 2014
Accepted 28 FEB 2015
Accepted article online 30 JUL 2015
Published online 4 SEP 2015

Investigating the recent apparent hiatus in surface temperature increases: 2. Comparison of model ensembles to observational estimates

Peter Thorne[1,2], Stephen Outten[1], Ingo Bethke[3], and Øyvind Seland[4]

[1]Nansen Environmental and Remote Sensing Center, Bjerknes Centre for Climate Research, Bergen, Norway, [2]Department of Geography, Maynooth University, Maynooth, Ireland, [3]Uni Research Climate, Bjerknes Centre for Climate Research, Bergen, Norway, [4]Norwegian Meteorological Institute, Oslo, Norway

Abstract To assess published hypotheses surrounding the recent slowdown in surface warming (hiatus), we compare five available global observational surface temperature estimates to two 30-member ensembles from the Norwegian Earth System Model (NorESM). Model ensembles are initialized in 1980 from the transient historical runs and driven with forcings used in the CMIP5 experiments and updated forcings based upon current observational understanding, described in Part 1. The ensembles' surface temperature trends are statistically indistinguishable over 1998–2012 despite differences in the prescribed forcings. There is thus no evidence that forcing errors play a significant role in explaining the hiatus according to NorESM. The observations fall either toward the lower portion of the ensembles or, for some observational estimates and regions, outside. The exception is the Arctic where the observations fall toward the upper ensemble bounds. Observational data set choices can make a large difference to findings of consistency or otherwise. Those NorESM ensemble members that exhibit Nino3.4 Sea Surface Temperature (SST) trends similar to observed also exhibit comparable tropical and to some extent global mean trends, supporting a role for El Nino Southern Oscillation in explaining the hiatus. Several ensemble members capture the marked seasonality observed in Northern Hemisphere midlatitude trends, with cooling in the wintertime and warming in the remaining seasons. Overall, we find that we cannot falsify NorESM as being capable of explaining the observed hiatus behavior. Importantly, this is not equivalent to concluding NorESM could simultaneously capture all important facets of the hiatus. Similar experiments with further, distinct, Earth System Models are required to verify our findings.

1. Introduction

Rightly or wrongly great import is attached to the global surface temperature record as a key indicator of climate change. This is despite the near-surface air temperature changes accounting for only a tiny fraction of a percent of the total energy balance of the climate system as a whole [*Rhein et al.,* 2013, Box 3.1].

It is beyond doubt that the rate of surface warming over the most recent 15 or so years has been considerably less than that since the mid-20th century [*Hartmann et al.,* 2013]. In spite of this, the most recent decade has been the warmest decade since the start of the instrumental record in the mid-19th century by a significant margin relative to recognized and quantified uncertainties [*Hartmann et al.,* 2013; *Morice et al.,* 2012]. The summary for policymakers of the Fifth Assessment Report of Working Group 1 of the Intergovernmental Panel on Climate Change (IPCC) explicitly recognized that the most recent 15-year period starts with a strong El Nino and that starting just a year or two earlier can have a strong impact on the resulting estimated trend [*Intergovernmental Panel on Climate Change*, 2013].

Both the observational record and numerous climate models (in historical runs and future projection runs) contain decade-plus stretches of either cooling or effectively no change in global surface temperatures during periods of multidecadal warming [*Easterling and Wehner*, 2009; *Knight et al.*, 2009; *Liebmann et al.,* 2010; *Foster and Rahmstorf*, 2011; *Santer et al.,* 2011]. That the surface temperatures (or virtually any other climatic parameter) are a superposition of internal climate system variability and the response to external forcings is well known and has been so for decades [e.g., *Hasselmann*, 1979]. Internal variability tends to dominate on timescales of a decade or so, while the response to external forcings tends to be dominant on multidecadal timescales [*Hawkins and Sutton*, 2009; *Marotzke and Forster*, 2015] (Figure 1).

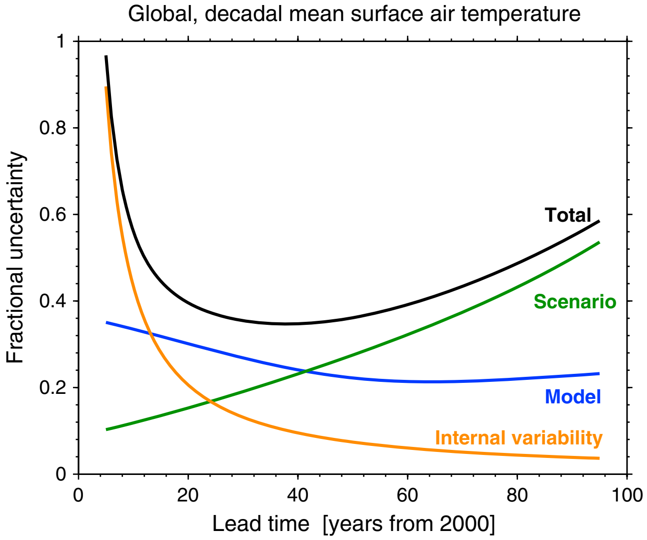

Figure 1. Modified version of Figure 3 from *Hawkins and Sutton* [2009] showing how various terms contribute to the total uncertainty in estimates of model global mean surface air temperatures with changing timescale. Model uncertainty relates to how different models respond to the same forcing. Here we use a single model, so this source is not considered (see section 5 for further discussion). Out to 20 years internal variability is greater than scenario (forcing) uncertainty. Figure courtesy of Ed Hawkins and Rowan Sutton, Reading University, UK.

Although the presence of a decadal timescale reduction in the rate of warming is not scientifically surprising even as the radiative forcing imbalance continues to increase, there has been intense public and scientific interest in the phenomenon which has been dubbed a "hiatus" (unless and until warming at the long-term rate or faster resumes, we note that this name is a misnomer, as the hiatus will only be verifiable upon such a resumption; but given its broad usage, we will adopt it here). There is interest in at least two inter-linked questions: (i) Does the hiatus call into fundamental question our understanding of climate change and its causes; and (ii) what does it imply about what we can expect on the planning horizon of 10–30 years and beyond.

The interest in the hiatus led to its highlighting in the most recent Fifth Assessment Report of Working Group 1 of IPCC [*Flato et al.,* 2013, Box 9.2]. That assessment had a limited literature available at the time of paper consideration cut-off. They concluded that

> In summary, the observed recent warming hiatus, defined as the reduction in GMST trend during 1998–2012 as compared to the trend during 1951–2012, is attributable in roughly equal measure to a cooling contribution from internal variability and a reduced trend in external forcing (expert judgment, medium confidence). The forcing trend reduction is primarily due to a negative forcing trend from both volcanic eruptions and the downward phase of the solar cycle. However, there is low confidence in quantifying the role of forcing trend in causing the hiatus, because of uncertainty in the magnitude of the volcanic forcing trend and low confidence in the aerosol forcing trend.

Since the paper cut-off date for AR5, there have been a vast number of high profile papers all contending various proximal causes (and hence resulting implications) for the hiatus. As it stands presently, it has been hypothesized in the literature that the hiatus may have arisen because of

1. Variations in El Nino Southern Oscillation/Pacific Decadal Oscillation and associated mechanisms and teleconnections [*Kosaka and Xie*, 2013; *Trenberth and Fasullo*, 2013; *Goddard,* 2014; *England et al.,* 2014; *Risbey et al.,* 2014; *Huber and Knutti*, 2014; *Trenberth et al.,* 2014; *McGregor et al.,* 2014; *Watanabe et al.,* 2014]
2. A series of small volcanic eruptions leading to persistently elevated stratospheric aerosol loadings [*Neely et al.,* 2013; *Santer et al.,* 2014; *Schmidt et al.,* 2014; *Huber and Knutti*, 2014]
3. A quiet solar cycle [*Schmidt et al.,* 2014; *Huber and Knutti*, 2014]
4. Changes in stratospheric water vapor [*Solomon et al.,* 2010]
5. Changes in rate of ocean heat content increase and its vertical distribution [*Balmaseda et al.*, 2013; *Meehl et al.,* 2011, 2013, 2014; *England et al.,* 2014; *Chen and Tung*, 2014; *Drijfhout et al.,* 2014]
6. Reductions in ozone depleting substances (ODSs) (which are also greenhouse gases) [*Estrada et al.,* 2013]
7. The efficacy of anthropogenic aerosol forcings [*Shindell*, 2014]
8. Arctic sea-ice reductions and associated circulation changes [*Petoukhov and Semenov*, 2010; *Outten and Esau*, 2012; *Honda et al.,* 2009; *Mori et al.,* 2014]
9. Residual observational uncertainties [*Cowtan and Way*, 2014]

These hypotheses are not necessarily mutually exclusive. Nor is it likely a case of there being a single cause. Which combination of causes has actually contributed has potentially significant implications both globally and regionally for the coming decades. The real world does not afford us the luxury of repeat experimentation whereby we can observe multiple real-world trajectories changing the boundary conditions systematically. Recourse therefore necessarily must be made to climate models. Models are by necessity a simplification of the real world although still massively complicated.

In an attempt to try to disentangle what the important factors are, herein we set out to systematically assess the potential import of most of these hypotheses in the framework of a large set of runs of the Norwegian Earth System Model (NorESM) in the model configuration that was used in the CMIP5 submitted runs and described in the accompanying Part 1 [*Outten et al.*, 2015]. To our knowledge, this is the first time that a substantial multi-member ensemble experiment for any Earth System Model has been used to systematically assess the hiatus and its potential causes. In this paper, we compare these runs to the full suite of available observational global surface temperature estimates. The two ensembles enable an initial exploration of all posited hypotheses except stratospheric water vapor (for reasons elucidated in Outten et al.). The remainder of this paper is structured as follows. Section 2 briefly introduces the observational data sets used and recaps the features of the two NorESM ensembles described in Part 1. Section 3 compares the modeled and observed surface temperatures. Section 4 tests various posited mechanistic explanations for the hiatus. The implications are discussed in section 5, while section 6 concludes.

2. Data Sets

2.1. Observational Data and Its Principal Features

We make use of five available global surface temperature data sets that combine land surface air temperature and sea surface temperature data to create a global surface temperature estimate. Rather than discuss at length each data set, for ease of reader interpretation, the principal characteristics of these products and their references are summarized in Table 1. Although there is overlap in data sources, in homogenization, and, to a lesser extent, in gridding and interpolation, these data products span a broad range of structural uncertainty [*Thorne et al.,* 2005] in global surface temperature estimation. We have taken each data set and regridded to a common 5 degree resolution basis (the coarsest source resolution), truncated to 1980 to 2012 for consistency with the modeled ensembles [*Outten et al.*, 2015] (the end date being dictated by when we could get reasonable observational estimates for a number of the perturbed forcings), and renormalized all gridbox series to a common 1981–2010 climatology for comparability.

Over the period of primary interest here—1998 to 2012—the five data products agree on several aspects of the broad scale spatial nature of the surface temperature trends in annual averages (Figure 2). As would be expected, differences and uncertainties are greatest where data are sparse or there are no data. HadCRUT4 data are uninterpolated and provide a reasonable approximation to the true underlying data coverage contributing to all products. All observational data sets agree that the Arctic has experienced significant warming during this period (to the extent sampled) while substantive parts of the Northern Hemisphere (NH) midlatitude continents have experienced cooling. Trends over the Southern Hemisphere (SH) midlatitude continents are mixed, particularly over Australia with Berkeley and the two HadCRUT-based products showing substantial gradients from a warming SW to a cooling NE and the two remaining products showing a much more muted gradient. Trends over Antarctica are particularly uncertain. Over the oceans, all products agree that much of the Pacific has cooled, consistent with the spatial signature of La Nina or the cold phase PDO, and there is good agreement on the sign if not the magnitude. All also agree that the Southern Ocean in the vicinity of the Drake Passage has experienced significant cooling. In the Atlantic, all data sets agree on a pattern of warming high latitudes, cooling midlatitudes, and warming tropics in both hemispheres. All data sets also agree that the Indian Ocean warmed during this period.

The annual cooling trend signal in the NH midlatitude continents over 1998–2012 arises in the boreal winter season (Figure S1 in the supporting information). This result was already shown for CRUTEM3 [*Cohen et al.,* 2012] and confirmed in several subsequent studies [e.g., *Kosaka and Xie*, 2013; *Trenberth et al.,* 2014], and is very robust to choice of observational data product used here. In contrast, most of the NH midlatitude landmass has warmed during boreal summer (Figure S2). The tropical Eastern Pacific trends dramatically change between the two seasons as DJF of 97/98 was affected by a strong El Niño, which had dissipated

Table 1. Principal Features of the Various Observed Products Used

Data Set	Land Source (Station Count)	Land Homogenization	SST Source	SST Homogenization	Gridding and Interpolation	Uncertainty Quantification	References and Source URL
HadCRUT4.3.0.0	CRUTEM4.3.0.0 (5682)	Primarily homogenization nationally/regionally	HadSST3.1.1.0	Bucket correction model prior to 1942, adjustments based upon call sign and metadata post-WW2	Gridding to 5 degree resolution regular grid; no infilling or interpolation	100 member ensemble that captures many sources of uncertainty in land and marine data homogenization	*Morice et al.* [2012], *Kennedy et al.* [2011a, 2011b], *Jones et al.* [2012] www.hadobs.org
MLOST v3.5.3	GHCNMv3.2.2 (7280)	Application of NCDC's pairwise homogenization algorithm	ERSSTv3b	Bucket correction prior to 1942	2 degree gridded product interpolated to +/−65 degrees with land data in high latitudes reinjected. Interpolation by EOTs limited to 2000 km	Post-facto quantification on global mean series not applicable at other scales	*Vose et al.* [2012], *Lawrimore et al.* [2011], *Smith et al.* [2008] ftp://ftp.ncdc.noaa.gov/pub/data/mlost/operational/
NASA GISTEMP	GHCNMv3.2.2 (7280)	As above	ERSTTv3b	As in MLOST	Gridding and distance weighting to 250 km from data. 2 degree resolution	None	*Hansen et al.* [2010] [note that this describes a version with different SST source] http://www.esrl.noaa.gov/psd/data/gridded/data.gistemp.html
Berkeley monthly land and ocean with air temperatures at sea ice	Berkeley Earth (39028)	Individual outliers are implicitly down-weighted. Neighbor-based test to identify breaks and each apparently homogeneous segment treated separately.	HadSST3.1.0.0	As in HadCRUT4	Kriging to as globally complete a field as possible. Kriging is cut-off at distance from any data. 1 degree resolution	Quantification on global series	*Rohde et al.* [2013] (for land, no reference for merged product) http://berkeleyearth.org/data
Cowtan and Way (krig v2.0.0)	CRUTEM4.2.0.0	As for HadCRUT4	HadSST3.1.0.0	As for HadCRUT4	Krig to globally complete field, observed data points are not allowed to be modified. 5 degree resolution	None	*Cowtan and Way* [2014] http://www-users.york.ac.uk/~kdc3/papers/coverage2013/series.html

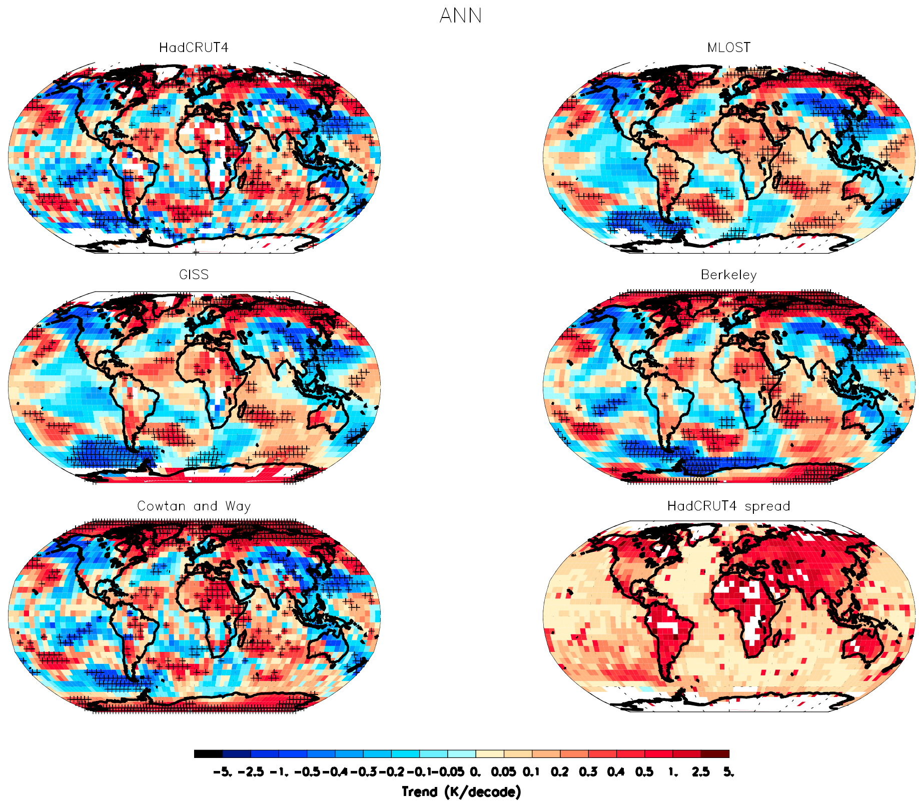

Figure 2. Annual mean time series ordinary least squares trends over 1998–2012 from the five observational estimates and the absolute spread in estimates in the HadCRUT4 100-member ensemble (bottom right). Trends in the first five panels have been assessed for significance after accounting for AR(1) residuals after *Santer et al.* [2008]. Gridbox trends that are significant are denoted by +. Gridbox trends were calculated only if >70% complete and at least one of first 2 years and last 2 years, respectively, are both present. Differences between data sets arise through input data choices, homogenization adjustments, and gridding and interpolation choices (Table 1).

by the following boreal summer which is an endpoint effect on the DJF but not the JJA trend calculation. Commensurately, there is strong cooling in the 15-year DJF trends not evident in the JJA trends. This highlights the sensitivity to start and end dates for linear trends for such short periods as is well known [*Santer et al.,* 2011, and references therein]. Figures S3 and S4 show the trend patterns for the somewhat longer period 1991–2012 and somewhat shorter period 2003–2012 in annual means, respectively. For the slightly longer period, trends are geographically smoother and vice versa.

The broad geographical features of the surface temperature evolution of concern to the analysis of the apparent hiatus including their seasonality are therefore relatively robust to different methods of analysis and interpolation. Nevertheless, there are potentially important differences, which may impact any resulting analysis and so recourse is made to the suite of observationally based estimates wherever possible and practical in sections 3 and 4.

2.2. Model Data

The present analysis makes use of two moderately sized 30-member ensembles from the CMIP5 model coupled configuration of NorESM run over 1980 to 2012. The experimental design is outlined in detail in the accompanying *Outten et al.* [2015], so only the essential details are recapped here. In particular, the reader should refer to Part 1 and references therein for details of the model configuration.

The two ensembles are named “Reference” and “Sensitivity” (henceforth always capitalized as proper nouns to distinguish from other potential uses of these words). The Reference runs were driven with the CMIP5 historical forcings until 2005 and then with the RCP8.5 scenario forcings until 2012. This is the same concatenation of forcing ancillaries as was performed in the production of the CMIP-5 archived Historical extension runs for NorESM. The Sensitivity runs consist of updated forcing ancillaries for long-lived greenhouse gases (LLGHGs) (including ODSs), aerosol emissions, and solar and volcanic forcings. The effects of stratospheric water vapor (which may be better considered a feedback or mode of variability rather than a forcing *per se*) were investigated, but due to ambiguity in slab ocean configuration response, this was not incorporated at this time. Hence, most forcings that have been hypothesized as potential explanators of the hiatus (section 1) are included. The major differences between Sensitivity and Reference ancillaries are as follows:

1. Systematically slightly higher net forcing from LLGHGs (about 0.015 W m^{-2} or 1% of total)
2. Lower solar forcing by c. 0.125 W m^{-2} at the surface in recent years
3. Slightly more latitudinally dependent Pinatubo loadings
4. Constant background volcanic aerosol with small peaks in most recent decade resulting from the series of mildly explosive eruptions yielding a difference of −0.08 W m^{-2} in recent years.
5. Reductions in the rate of decrease of global tropospheric aerosol concentrations and greater interannual variability post-2000. Aerosols include sulfate, primary organic matter, and black carbon, and are proportionately smallest for sulfate. The net effect in recent years is c. +0.14 W m^{-2} but varies markedly interannually. As discussed in *Outten et al.* [2015], this estimate is incomplete, as certain aerosol terms could not be calculated in this version of NorESM1.

The net effect of the changes is severalfold over the hiatus period. First, on average, the Sensitivity ensemble has a somewhat lower average absolute top-of-atmosphere (TOA) radiative forcing of roughly 0.03 W m^{-2} taken over the hiatus as a whole. Second, despite the lower average TOA forcing, the Sensitivity ensemble actually ends with higher TOA forcing such that the linear trend in the TOA forcing over the hiatus is slightly more positive in Sensitivity than Reference. This is largely driven by the uncertain and incomplete aerosol forcing radiative effect estimate. Lastly, the Sensitivity ensemble shows substantially greater interannual variability in the applied external forcings.

Outten et al. [2015] also assessed some of the key model physical characteristics of the two ensembles and concluded that the ensembles were grossly similar, in such aspects which may be important in explaining the hiatus as El Nino Southern Oscillation (ENSO) variability, Ocean Heat Content (OHC) uptake, and sea-ice behavior. ENSO behavior in terms of magnitude and periodicity of Nino3.4 region Sea Surface Temperatures (SSTs) was found to be comparable to that observed. However, the NorESM ensembles' Arctic sea-ice extent consistently exhibited both too small an annual cycle (underestimating the winter maximum and overestimating the summer minimum) and too muted trends in sea-ice extent reduction compared to observations.

The model fields have been regridded to the same 5 degree basis as the observations and renormalized on a gridbox basis to a 1981–2010 climatology to enable direct comparability. The regridding and renormalization has been repeated five times—once for each observational data set mask. Only the Cowtan and Way analysis is complete throughout the entire period of interest (Berkeley has a handful of missing values and the remainder many more). For some subsequent comparisons, Cowtan and Way therefore provides the sole or primary observational comparator where it is impractical to repeat the analysis five times. But in most subsequent analyses, the model has been compared five times, each time using the appropriate observational mask for comparison.

3. Comparisons of Observed and Modeled Surface Temperatures

Within this section, comparisons are undertaken between the two 30-member NorESM ensembles (section 2.2) and the available observational estimates (section 2.1). Comparisons concentrate upon trying to disentangle the potential roles of forcing uncertainty (apparent by any differences between the two model ensembles) and internal climate system variability (apparent by the spread within each model ensemble) as explanations of the observed behavior. In section 3.1, comparisons are made of modeled and observed surface temperature series; section 3.2 compares regional trends; and section 3.3 compares geographical trends.

For comparisons of large-scale averages, data have been zonally averaged and then these zonal averages cos (lat) weighted. Averages have been calculated for the globe, each hemisphere, the tropics (30°S to 30°N), the midlatitudes (30° to 60°N/S for NH and SH, respectively), and the Arctic and Antarctic regions (defined here as poleward of 60°N and S, respectively). Zonally averaging first then weighting serves to diminish the effects of unequally sampling different latitude bands, particularly prevalent in MLOST and especially HadCRUT4, or

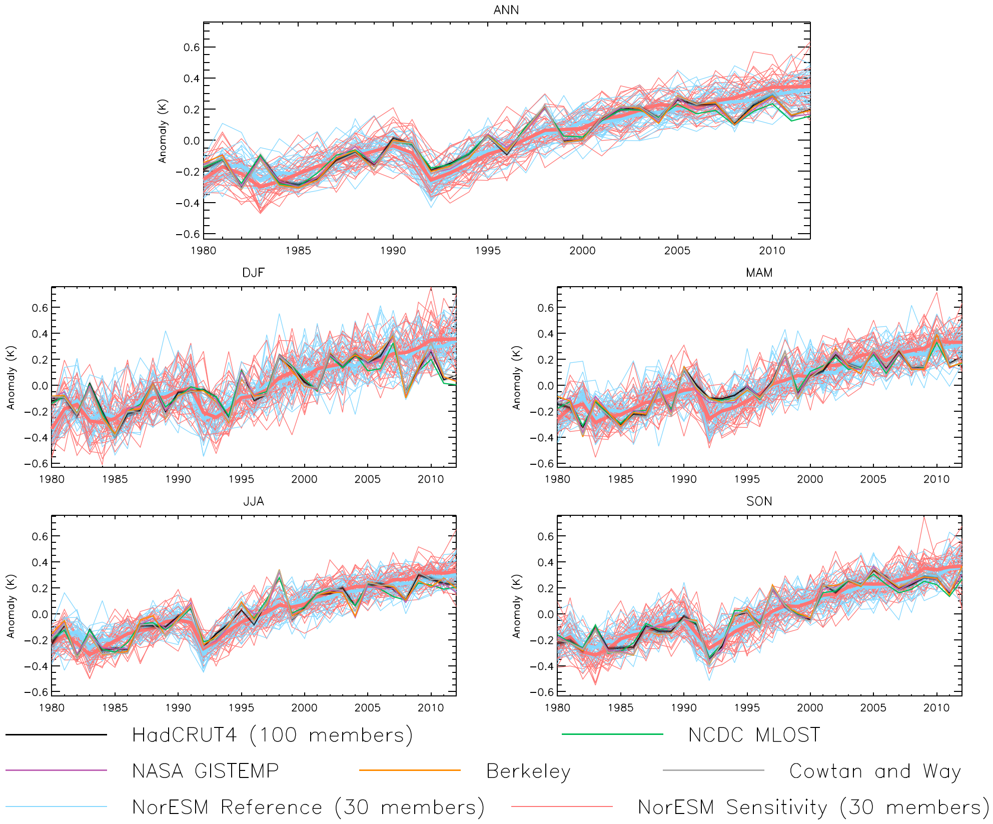

Figure 3. Annual and seasonal global mean anomaly time series relative to a common 1981–2010 climatology period from the observational estimates and the two 30-member NorESM model ensembles. HadCRUT4 and the two model ensembles are denoted by multiple traces, one for each ensemble member. In addition, ensemble means are given for the two NorESM ensembles by thicker lines of the same color. Global averages have been derived by zonally averaging and then cos(lat) weighting. Model traces are derived from the Cowtan and Way masked (spatially complete) fields. Observational traces utilize their respective masks.

NorESM runs sampled as these products. Averages are only calculated if more than 30% of the zonal band averages are sampled for the region of interest. The choice of a 30% threshold is arbitrary, but its effects will be mitigated because in each comparison, the model is subsampled to the same spatio-temporal mask as observed.

3.1. Time Series Comparisons

Globally averaged observed annual and seasonal time series agree over the period 1980–2012 to the extent that they are virtually indistinguishable from one another (Figure 3). The two NorESM-based ensembles in comparison exhibit substantial interannual spread arising from their different phasing of natural variability such as ENSO in the coupled runs. The interannual variability in the observations and individual model runs is broadly comparable, as is the seasonality, which shows greater variability in boreal winter than boreal summer. The observed estimates reside within the spread of model estimates most of the time, falling outside and below the spread toward the end of the series in both the annual average and DJF but not in the remaining seasons. The differences between the two ensemble mean series are orders of magnitude smaller than the within ensemble realization spread.

Hemispheric averages for the NH (Figure 4) and SH (Figure 5) similarly fall largely within the modeled spread. The seasonality in variance arises from the NH which exhibits markedly increased variances in all seasons compared to the SH in both the observed estimates and the models and far greater variance in boreal winter than boreal summer (note that the *y*-axis ranges are consistent between panels within Figures 4 and 5 but different between the two figures). The agreement between the two model ensembles and the observations is somewhat worse for the SH than the NH with the model tending to estimate a greater overall change and the observations end up either on the lower bound or entirely outside the modeled

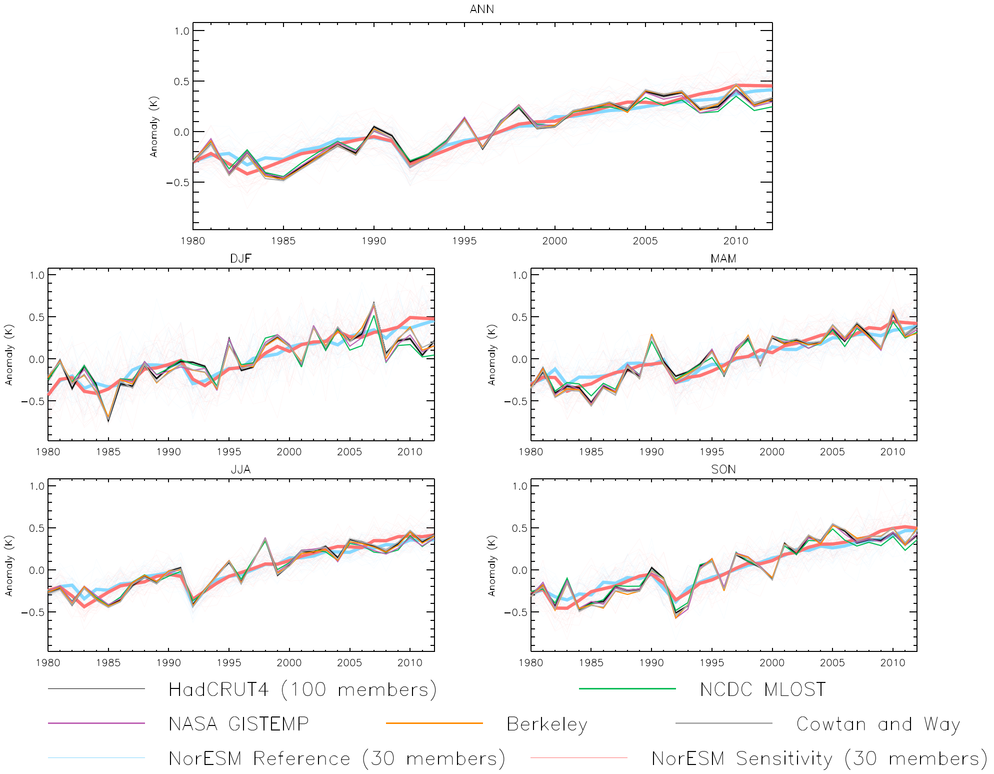

Figure 4. As Figure 3 but for Northern Hemisphere averages. Note that for presentational purposes for each region, the *y*-axis ranges are derived dynamically by the range of time series values. Although each panel is consistent within Figure 4, the *y*-axis range is distinct from those used in Figure 3.

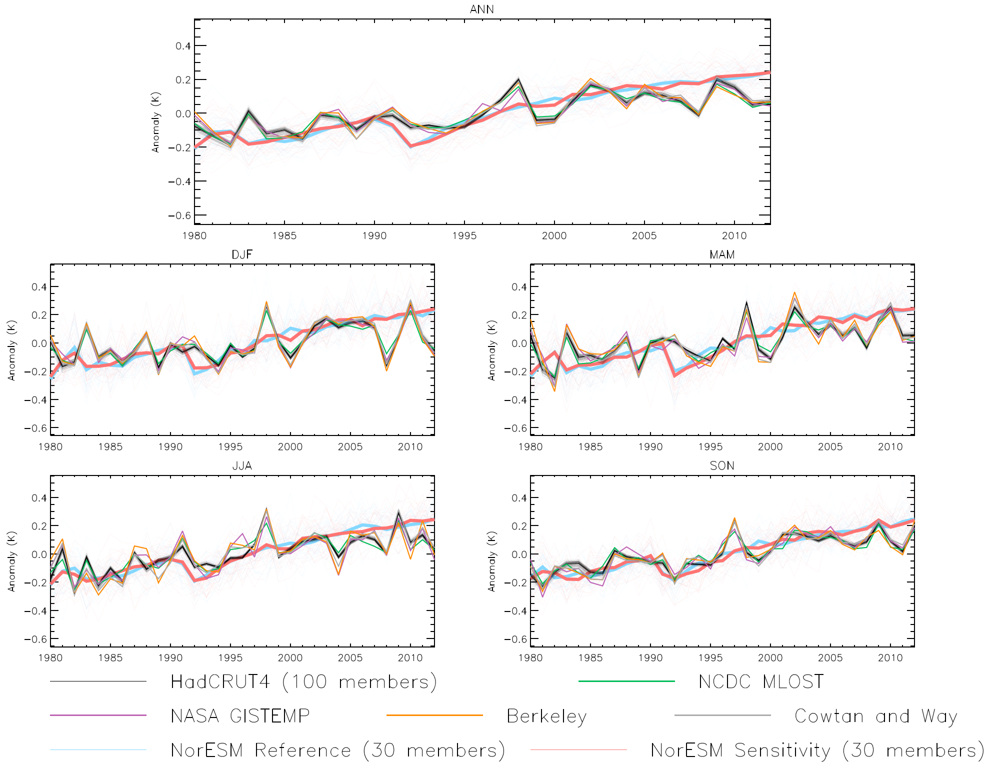

Figure 5. As Figure 4 but for the Southern Hemisphere.

range both annually and for austral summer. As at the global mean scale, the two ensemble mean series are similar to one another and much smaller than intra-ensemble spread.

Time series for the tropics, midlatitudes, and polar regions are given in Figures S5 through S9. The tropics as defined here account for half the area in the area weighted global mean which are very similar to Figure 3 (Figure S5). For the NH midlatitude regions, the annual time series are broadly similar between the two model ensembles and observations (Figure S6). This masks interesting seasonality apparent in the observations. Boreal winter temperatures in the region rose rapidly through the end of the 20th century but have since dropped substantially. Boreal summer temperatures meanwhile have risen, and this rise has continued unabated even while the winters have cooled [*Cohen et al.,* 2012]. It is not clear, given the plots, whether any individual model run captures this marked seasonality. This aspect is returned to in section 4.3. The SH midlatitudes show somewhat poorer agreement between model ensembles and observations with some evidence of slightly greater interannual variability in observations than the model ensembles (Figure S7). However, the long-term time series are in broad concurrence in this region. Finally, the Arctic and Antarctic regions (Figures S8 and S9) exhibit marked seasonality in variance with greatly lower variance in their respective summers than winters. They also exhibit far higher overall variance than other regions making any meaningful assessment of the time series harder. The Arctic is warming most rapidly of all regions considered, whereas the Antarctic region is either not changing at all or slightly cooling. For both polar regions, the observations reside within the model spread annually and for all seasons.

In summary, the time series plots show that the NorESM model captures interannual variability and its seasonality globally, hemispherically, and regionally reasonably well. There is a propensity for the observations to fall either toward the lower bound or outside the range of model estimates toward the end of the series globally, and this is driven largely by the tropics and NH midlatitudes and is particularly marked in boreal winter. Neither choice of observational data set nor choice of which NorESM model ensemble to consider has a demonstrable impact when considering time series behavior consistency.

3.2. Large Region Averages Trends

Time series traces (section 3.1) can only provide limited and indirect evidence of long-term behavior congruence between model estimates and observations. Therefore, next we consider the evidence from ordinary least squares regression fits to the time series over the period of the hiatus (1998–2012). Use of the observational data set masks leads to some changes in the model boxplots between panels shown in Figures 6 and 8 and S10 through S16. Such differences between the five model estimates can only arise due to sampling. Conversely, differences between observational estimates may arise due to any combination of input data selection or quality control and/or homogenization applied and interpolation procedures (section 2.1; Table 1).

Global mean trends on an annual and boreal winter and boreal summer basis for the period of the hiatus are at best marginally consistent between the model ensembles and the various observational estimates (Figure 6). The Sensitivity ensemble with perturbed forcings described in Part 1 [*Outten et al.*, 2015] has a tendency to slightly greater warming than the Reference ensemble performed with the forcings used in the CMIP5 submission. Differences between the ensembles are not significant at the 5% level according to a Student's *t*-test (Table 2, first row). The impact of the different forcings used in the two ensembles is a secondary effect compared to the impacts of modeled internal climate variability on trends over this period returned by NorESM in concordance with e.g. *Hawkins and Sutton* [2009] (Figure 1).

On an annual mean global mean basis only the HadCRUT4, Berkeley, and Cowtan and Way observational estimates fall within the distribution of either NorESM ensemble's trends and then only within the low tail of the ensembles. Seasonally, a single NorESM ensemble member from each ensemble falls below the observed cooling apparent in most estimates in boreal winter, whereas in boreal summer, there is somewhat better concordance between the model ensembles and observations.

Results for all regions and comparators are summarized in Table 3 and available as plots in Figures S10 through S16. There is a marked difference based upon choice of observational data set and area under consideration as to the strength of evidence for a fundamental discrepancy between the observational estimates and the model ensembles. Taking an intuitive, semi-qualitative approach, agreement with model runs is graded in Table 3 depending in decreasing order upon whether it is (i) within the interquartile range (IQR) of the ensemble (best), (ii) within the ensemble at all, or (iii) outside the ensemble (worst).

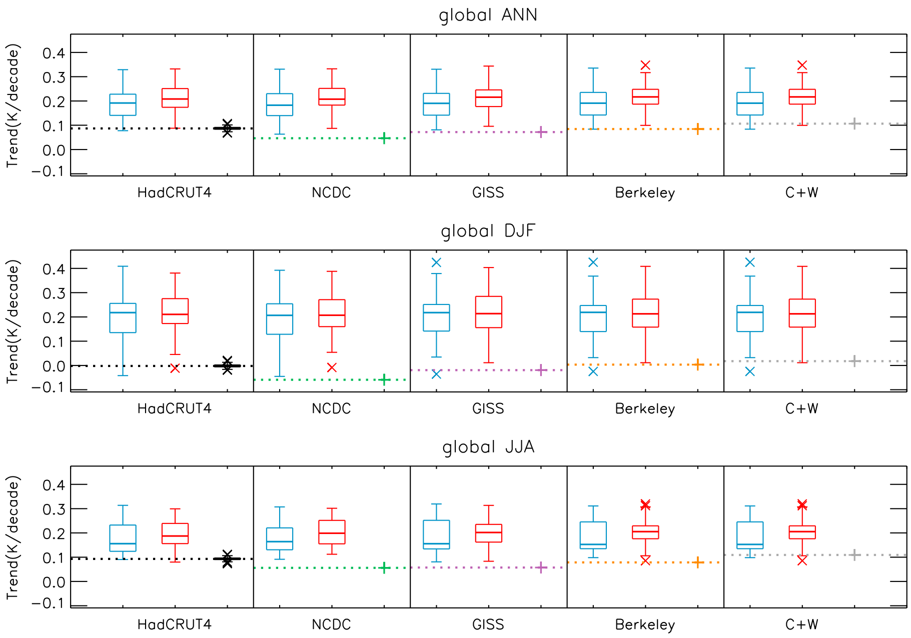

Figure 6. Global mean trend boxplots for the period 1998–2012. Trends have been calculated using ordinary least squares regression on annual (top), boreal winter (middle), and boreal summer (bottom) time series. The two 30-member NorESM ensembles are denoted by blue (reference) and red (sensitivity) boxplots. They are recalculated five times—once for each observational mask. HadCRUT4 ensemble is denoted by a black boxplot in the leftmost column. Remaining observational estimates are denoted solely by a single estimate. This estimate (the median for HadCRUT4) is extrapolated across the relevant NorESM boxplots as a dashed line to aid interpretation as to the consistency of the observational estimate with the two model ensembles. Boxplots here and in all similar plots are presented as *Tukey* [1977] boxplots with outlying values beyond 1.5 times the interquartile range denoted by crosses.

Given the still relatively modest ensemble sizes, no attempt is made to assess formal statistical consistency of the ensembles with the observations.

Regionally, the worst agreement is for the Northern Hemisphere midlatitudes both annually and in boreal winter when only one occurrence of qualitative consistency occurs across all comparisons (Berkeley with Reference in DJF). The tropical region also lacks consistency in winter trends, but this suffers in the observations starting with the endpoint of the large 1998 El Nino (section 3.1). There is consistency in summer for both ensembles and annually for Sensitivity. Also, only marginally consistent are the Southern Hemisphere taken as a whole and the global mean discussed previously (Figure 6). For the tropics, SH midlatitudes, and both polar regions, there is reasonable occurrence of the observations falling within the ensemble range if not within the more restricted criteria of the IQR. In all regions except the Arctic, the overall tendency is for the observational estimates to fall either within the lower portion of the ensembles or outside and below the ensembles altogether. Within the Arctic, however, there is a tendency for the observations to fall within the upper portion of the model ensemble spread.

Table 2. Two-Tailed Significance Assessments for Test of Differences in Trends Over 1998–2012 Between the Two 30-Member NorESM Model Ensembles Using a Student's *t*-Test[a]

Region	Annual	DJF	JJA
Global	0.19	0.57	*0.08*
NH	0.09	0.56	**0.02**
SH	0.91	0.78	0.80
Arctic	0.45	0.90	*0.05*
NH midlats	0.41	0.97	0.37
Tropics	0.18	0.21	0.22
SH midlats	0.77	0.78	0.84
Antarctic	0.52	0.62	0.83

[a] Cases significant at the 10% level are italicized. Cases significant at the 5% level are bolded. There are no cases where the two ensembles differ from each other at the 1% significance level. Tests have been applied to trends for the globe and seven subdomains and on annual, boreal winter and boreal summer seasons. Cases of significant differences are limited to the boreal summer season, and the three regions overlap.

Table 3. Summary of Trend Similarity for Global, Hemispheric and Five Additional Zonal Band Averages Annually and in Boreal Winter and Summer[a]

		HadCRUT		NCDC		GISS		Berkeley		C+W	
Region	Season	Ref	Sen	Ref	Sen	Ref	Sen	Ref	Sen	Ref	Sen
Global	ANN	🟩	🟩	🟦	🟦	🟦	🟦	🟩	🟦	🟩	🟩
	DJF	🟩	🟩	🟦	🟦	🟩	🟦	🟩	🟦	🟩	🟩
	JJA	🟩	🟩	🟦	🟦	🟦	🟦	🟩	🟦	🟩	🟩
NH	ANN	🟩	🟩	🟩	🟦	🟩	🟩	🟩	🟩	🟩	🟩
	DJF	🟩	🟩	🟦	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	JJA	🟨	🟩	🟩	🟦	🟩	🟦	🟨	🟩	🟨	🟩
SH	ANN	🟦	🟩	🟦	🟩	🟦	🟩	🟦	🟩	🟦	🟩
	DJF	🟩	🟦	🟦	🟦	🟩	🟦	🟩	🟦	🟩	🟦
	JJA	🟦	🟩	🟦	🟦	🟦	🟩	🟦	🟩	🟦	🟩
Tropical	ANN	🟦	🟩	🟦	🟩	🟦	🟩	🟦	🟩	🟦	🟩
	DJF	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
	JJA	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
NH mid-lats	ANN	🟦	🟩	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
	DJF	🟦	🟦	🟦	🟦	🟦	🟦	🟩	🟦	🟦	🟦
	JJA	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩	🟨	🟩
SH mid-lats	ANN	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	DJF	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	JJA	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Arctic	ANN	🟨	🟨	🟨	🟨	🟨	🟨	🟧	🟨	🟧	🟨
	DJF	🟧	🟧	🟨	🟨	🟧	🟧	🟧	🟧	🟧	🟧
	JJA	🟧	🟨	🟨	🟩	🟨	🟩	🟧	🟨	🟧	🟨
Antarctic	ANN	🟨	🟨	🟩	🟩	🟨	🟨	🟩	🟨	🟨	🟨
	DJF	🟨	🟨	🟩	🟨	🟨	🟨	🟨	🟩	🟨	🟨
	JJA	🟨	🟨	🟩	🟩	🟩	🟨	🟩	🟩	🟨	🟨

[a] Sourced from the regional boxplot figures. Each observational data set has two columns, one each for Reference and Sensitivity ensembles. Cell shading is as follows: blue—observations outside and below the model ensemble; green—observations within the model ensemble range but below the lower quartile; yellow—observations within the model ensemble IQR; and orange—observations within the model ensemble range but above the upper quartile. There are no occurrences where the observations are outside and above the model ensemble range.

Regional boxplots for *Cowtan and Way* [2014] (summarized in the last two columns of Table 3) are presented in Figure 7. Cowtan and Way has almost the most frequent occurrence of falling within the NorESM ensemble (Table 3), so this presents one of the most optimistic of five potential such comparisons. Robustness to observational data set choice can be ascertained from consideration of Figures S10 through S16.

The most immediately striking feature in Figure 7 is the marked gradation in ensemble spreads in the two NorESM experiments between the regions. The smallest spreads are in the tropics, SH midlatitudes, and the two hemispheric averages (Southern Hemisphere most markedly). By far, the largest spread is for the Arctic region where the total spread in the ensembles is up to 2.5 K dec^{-1} over this period. Another striking feature is the seasonality in observed trends in the NH extra-tropics and in particular the boreal winter trend asymmetry between the Arctic and midlatitudes. Because these largely cancel when area weighted, there is a more muted seasonality in the Northern Hemisphere averages. The marked contrast between the adjacent Arctic and NH midlatitudes and its seasonal structure points toward a potential role for wintertime circulation anomalies which may be associated with sea ice (sections 4.3 and 4.4). In the extra-tropics, the ensemble spread is larger in winter than in summer in both hemispheres, but this is more marked for the Northern Hemisphere with its greater landmass and enclosed polar sea.

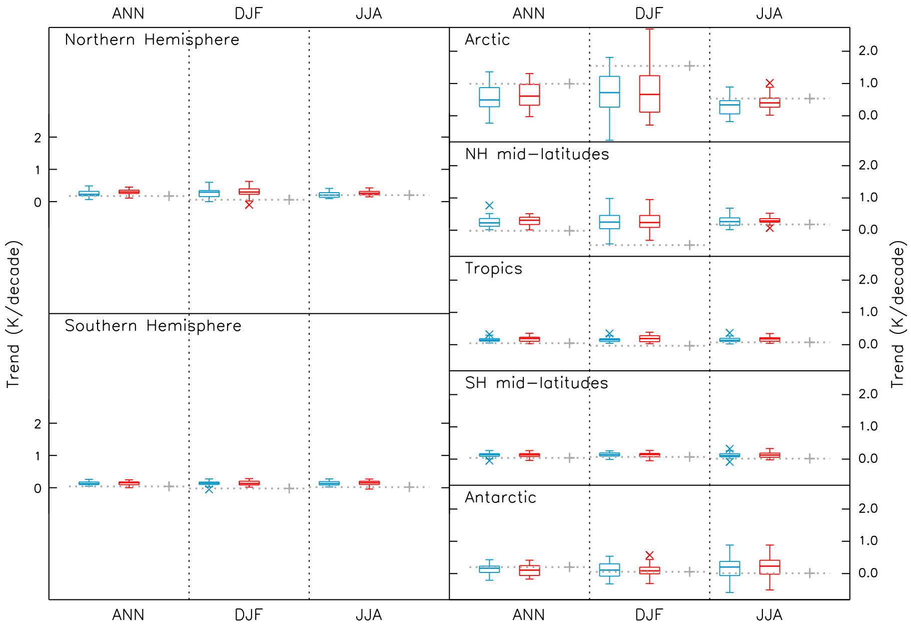

Figure 7. As Figure 6 but for Cowtan and Way data set comparison to the two ensembles for all subglobal regions considered. To aid comparability, all regions are presented on a common *y*-axis.

It is also interesting to note the interplay between a semi-qualitative assessment of consistency assessed by overlap with the moderately large ensembles and the absolute differences between the model median estimates and the observations. For regions of small ensemble spread, there are several cases where no overlap between the ensembles and the Cowtan and Way observational estimate exists (Table 3), yet an inspection of Figure 7 suggests several cases where the observations are far further from the ensemble median yet fall within the ensemble spread. This raises the question of how to formally assess consistency in a manner that retains appropriate physical interpretative value. Clearly, information on distance from the central tendency of the ensemble must be a useful piece of information to physical interpretation of any results.

It is evident from Figures 6 and 7 and Table 3 that the two model ensembles' trends globally and regionally substantively overlap. To formally assess the significance of any differences a Student's *t*-test has been applied to all regions and seasons in Figures 6 and 7 using the full spatial mask of Cowtan and Way (Table 2). There are no cases of significant differences either annually or in boreal winter between the Reference and Sensitivity ensembles. In boreal summer, the global and Arctic ensemble trends differ at 10%, and the NH averages at 5% with the Sensitivity ensemble exhibiting greater warming rates on average. These three regions overlap, so this should not be interpreted as being three independent lines of evidence. Given that we would expect, by chance, to find spurious significance in some cases and that only one case is found at 5% significance out of 24 total tests, we can reject the hypothesis that forcing errors to the extent explored herein have played a major part in the hiatus at regional and global scales according to NorESM.

Modeled and observed trend consistency may also be impacted by trend length. To assess this, Figure 8 repeats the global analysis but for the somewhat longer and shorter periods of 1991–2012 and 2003–2012, respectively (cf. Figures 1, S3, and S4 for observed spatial trends). Consistency between the NorESM ensembles and the observations is even more marginal for 1991–2012 than it is for the hiatus period. This is primarily because there is a commensurate decrease in the ensemble spread. Indeed, the observational estimates move closer to rather than further from the ensemble median behavior by considering this slightly longer period, but this reduction in absolute distance is less than the reduction in ensemble spread [*Santer et al.*, 2011]. Conversely, the observations always fall within both model ensemble spreads over 2003–2012 despite the observational trend estimates now falling even further from the ensemble medians than is the case over the hiatus period.

In summary, the use of more recent observationally based forcings as described in *Outten et al.* [2015] is found to be inconsequential compared to natural variability in NorESM (to the extent sampled in the ensembles) and to yield no significant change in the trend over the period for almost all domains and

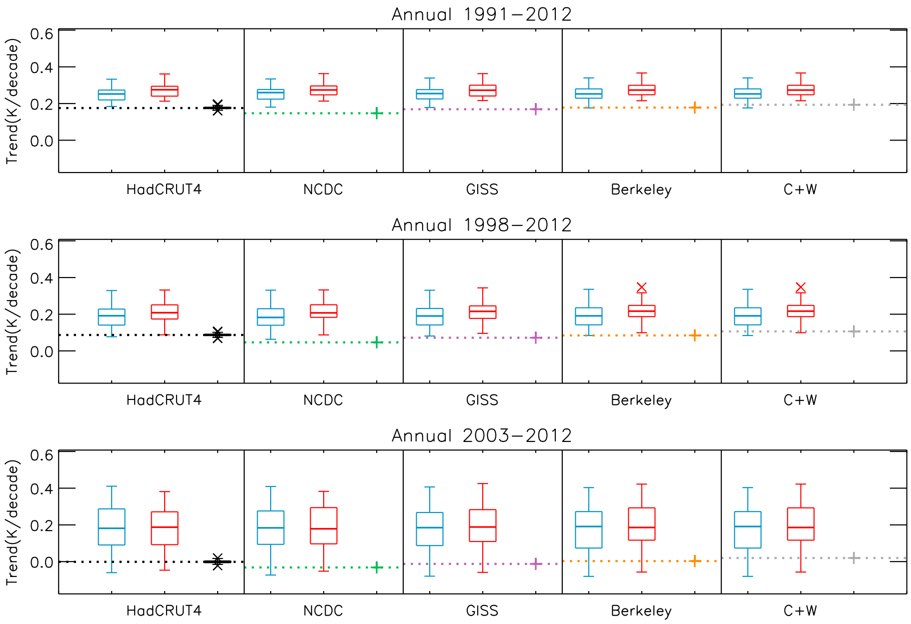

Figure 8. As Figure 6 but for different periods of record: 1991–2012 (top panel); 1998–2012 (middle panel, exactly equivalent to the top panel of Figure 6 but with stretched *y*-axis range); and 2003–2012 (bottom panel).

seasons. Only a subset of the NorESM ensemble runs considered herein are in agreement with some of the observational data sets during the period of the hiatus. Hence, the model runs considered here are at best marginally consistent with observations globally and in many regions, most markedly the NH midlatitudes where there is minimal overlap annually or in boreal winter. Conversely, there is reasonable agreement between observations and the ensembles in boreal summer in this region. Further analyses in terms of internal variability as simulated by the model are given in section 4, and section 5 includes broader discussion of necessary caveats pertaining to experimental design limitations.

3.3. Spatial Trends

In Figure 2, it was shown that the observations exhibit similar large-scale trend patterns over the hiatus (Figures S3 and S4 showed the same for boreal winter and summer). Given this, comparisons have been made between the NorESM ensemble spatial trend patterns and the observational trend estimates arising from Cowtan and Way (Figures 9, S17, and S18). The observed trend is the result of both the forced response and a single manifestation of internal real-world climate system variability. If NorESM is an adequate model, then this will be able to be more closely approximated by individual ensemble members than the ensemble means. Comparisons are hence made between individual ensemble members and the observations. Two simple similarity metrics have been used to rank similarity to the Cowtan and Way trend field-pattern correlation and Root Mean Squared Differences (RMSD). The figures show the first two ensemble members in each ensemble that rank consistently within the top five correlations and lowest five RMSDs on an annual mean trend basis. To permit an assessment of sensitivity to interpolation choices, Berkeley is shown which also uses a kriging interpolation scheme but does not retain the anomalies exactly where observations are reported. Berkeley also has other methodological distinctions from Cowtan and Way (Table 1).

The model runs shown in Figure 9, while exhibiting differences to the observations, do have some features that are similar (as would be expected given their correlation and low RMSD). Perhaps most importantly, the spatial scales are broadly similar, and some of the e.g. wave-like pattern in the extra-tropics are reasonably captured. Several runs appear capable of capturing cooling over Asia or North America, although none of the runs considered here capture both. Some ensemble members also capture reasonable signatures of cooling trends in the tropical Pacific over the hiatus. Conversely, there are some model runs that exhibit negative trend pattern correlations with Cowtan and Way over this period and much higher RMSDs (not shown).

On a seasonal basis (Figures S17 and S18), the four subselected ensemble members are capable of capturing the NH midlatitude continental cooling trends with the 13th member of the Reference ensemble and to a

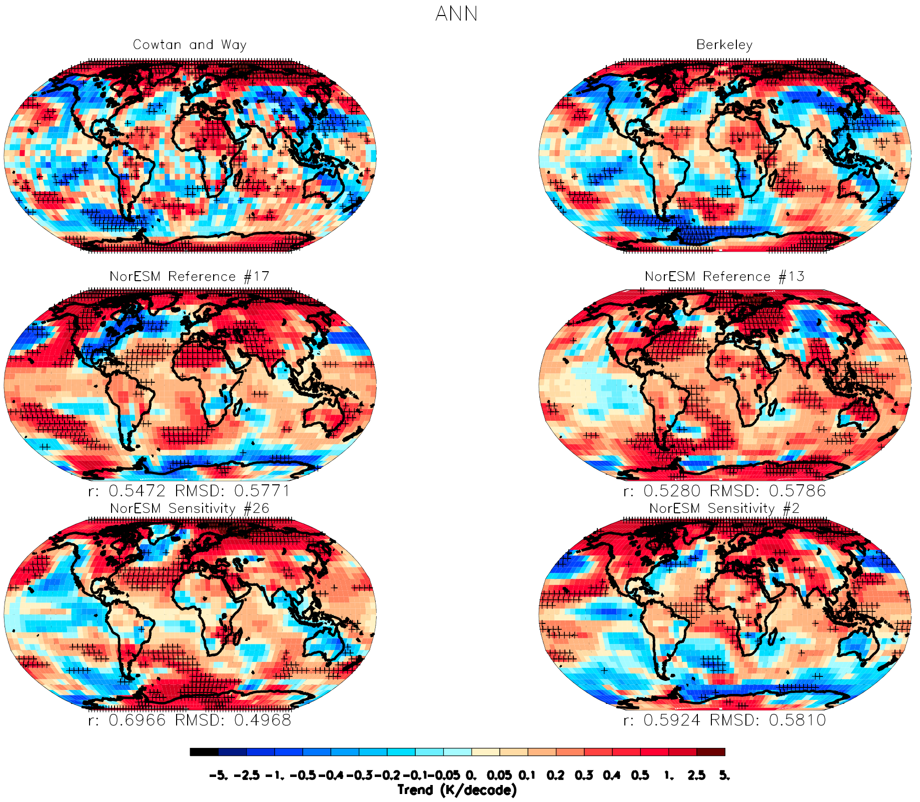

Figure 9. Comparison of linear trends over 1998–2012 for Cowtan and Way and Berkeley (top row), and for each ensemble, the two members which have correlations and RMSD within the top five of agreement (field correlations and RMSD values are shown below each model panel). See Figure 2 for technical details of trend calculation and significance assessment.

lesser extent the second ensemble member of Sensitivity capturing simultaneous cooling in both continents' interiors. The same Reference run also does a reasonable job of capturing the summertime trends over these regions. More optimistic assessments of seasonal concordance could be attained by selecting based upon seasonal trend correlation and RMSD (Figure 10).

Further, there is significant sensitivity of spatial trends to start and end dates for such short periods both in the observations (Figure 11) and the model ensembles (not shown). The sensitivity is greatest in midlatitudes to high latitudes as well as the tropical Pacific, reflecting the endpoint effects of large interannual variability in these regions. To ensure a conservative assessment and avoid conflating the effects of the differences in forcing factors and natural variability, a fixed time window agreement has been used here to assess model-observation similarity. If the spatial pattern response is almost entirely driven by internal variability, this artificially reduces the comparison population available to find good matches. An approach of looking for similarities in any 15-year model period regardless of its model start year would almost certainly yield many more cases of high trend pattern similarity.

In summary, there is a broad range of spatial trend signatures across the NorESM ensembles with a broad range of correlations with the observed patterns seen in both ensembles. Some model runs give patterns with spatial scales that look reasonably coherent and can even reproduce several of the warming/cooling hotspots in some regions over the time span of the hiatus and aspects of their seasonality. There is substantial sensitivity to choice of start and end dates, especially in the extra-tropics, that complicates clean analysis of spatial trends over periods as short as the hiatus.

3.4. Summary of Model-Observation Surface Temperature Comparisons

It is clear that distinctions between the forcings as prescribed in CMIP5 and more up to date set of forcings documented in *Outten et al.* [2015] do not contribute to a flattening of surface temperature trends over the

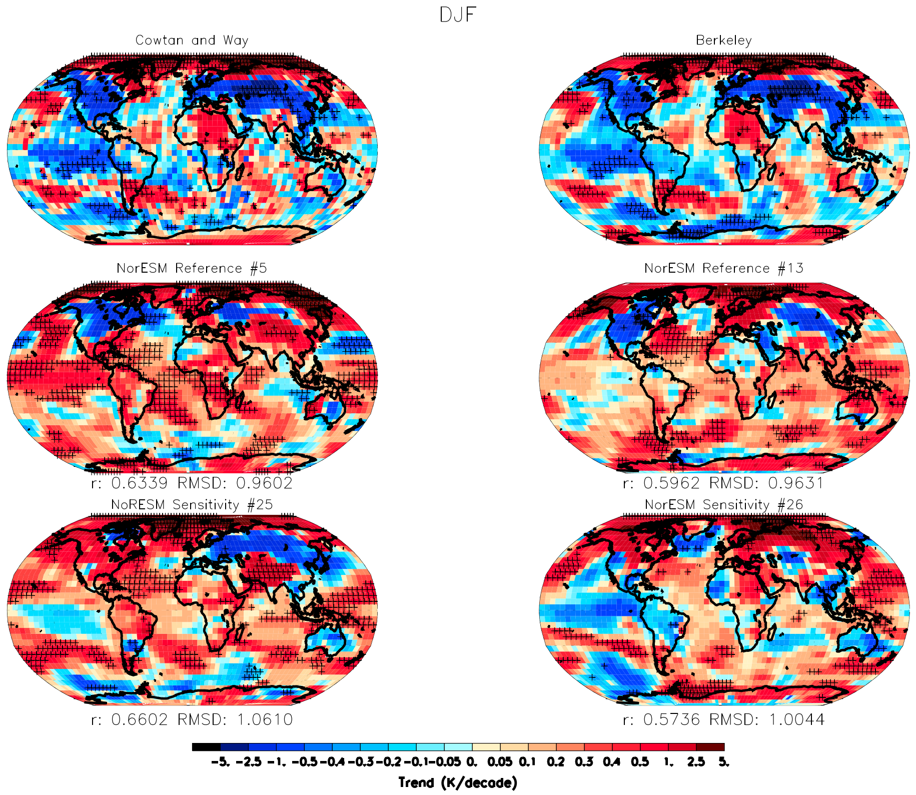

Figure 10. As Figure 9 but with the four model ensemble members being selected based upon DJF season global trend pattern correlation and RMSD.

hiatus period according to NorESM. This includes the solar and volcanic forcing discrepancies highlighted in *Flato et al.* [2013, Box 9.2] and subsequently [*Santer et al.,* 2014; *Schmidt et al.,* 2014] as potential contributors. The two ensembles are statistically indistinguishable over the hiatus period. According to the 60 NorESM runs completed, it is primarily internal climate system variability, as diagnosed by the model, which could account for the observed behavior over the hiatus period. Overall, the model runs capture many, but not all, facets of large-scale variability and features apparent in the observations with reasonable fidelity. Apparent discrepancies are most marked in the NH midlatitudes where there is a very marked seasonality in observed trends with annual and winter trends falling outside and below the model ensembles despite their relatively broad spread. This appears to be dominated by the observed wintertime cooling trends over the continental interiors of Eurasia and North America since the early 21st century.

4. Testing Posited Mechanistic Explanations

As well as an analysis of the sensitivity to forcing uncertainties, the two 30-member ensembles permit an assessment of internal climate system variability mechanisms. Previous analyses have considered multimember ensembles, which conflate differences in model formulation [e.g., *Risbey et al.,* 2014], or the use of Earth System Models of Intermediate Complexity [e.g., *Huber and Knutti*, 2014], which may not capture important mechanisms. Here, the two 30-member ensembles are used to consider various posited internal climate system mechanisms that could be important in explaining the hiatus behavior.

4.1. El Nino Southern Oscillation/Interdecadal Pacific Oscillation

Simple similarity metrics of RMSD, correlation, and linear trend were considered between observed Nino3.4 region SST behavior (estimated from HadCRUT4) and the model ensembles. It should be stressed that Nino3.4 region trends on decadal timescales are typically indicative of broader pan-Pacific changes associated with the Interdecadal Pacific Oscillation (IPO). For traceability to the *Risbey et al.* [2014] CMIP-5 multi-model ensemble analysis, we concentrate here on Nino3.4 region trends. But we note that our results may speak more broadly to pan-Pacific changes [*Meehl et al.,* 2011, 2013].

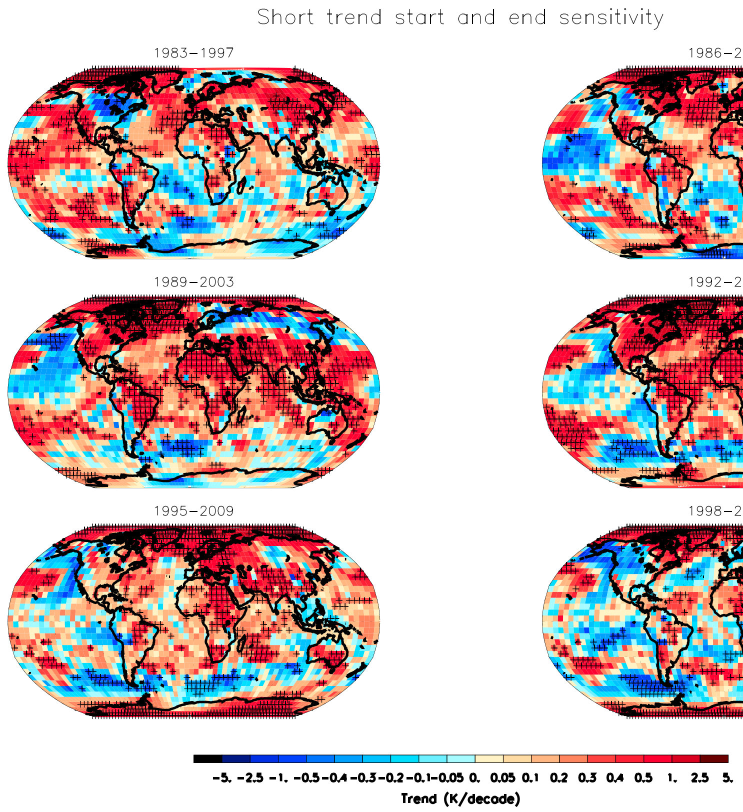

Figure 11. Fifteen-year linear trend estimates and their significance for Cowtan and Way for a 3-year moving window set of overlapping periods. The final panel is directly equivalent to the upper left panel in Figure 9. For technical details, see Figure 2.

There was no conditioning impact on long-term trend behavior found for either time series Root Mean Squared Differences (RMSD) or correlation globally, in the tropical region, or elsewhere (not shown). Consistent with *Risbey et al.* [2014], we find some correlation of regional trends to STT trends in the Nino3.4 region (Figure 12). The correlation is positive—higher Nino3.4 region trends result in greater warming. The impact is most pronounced in the tropics where Nino3.4 region trend similarity explains over half the variance in the modeled tropical mean trends. The slope is much less than 1:1 reflecting that ENSO/IPO is in part a

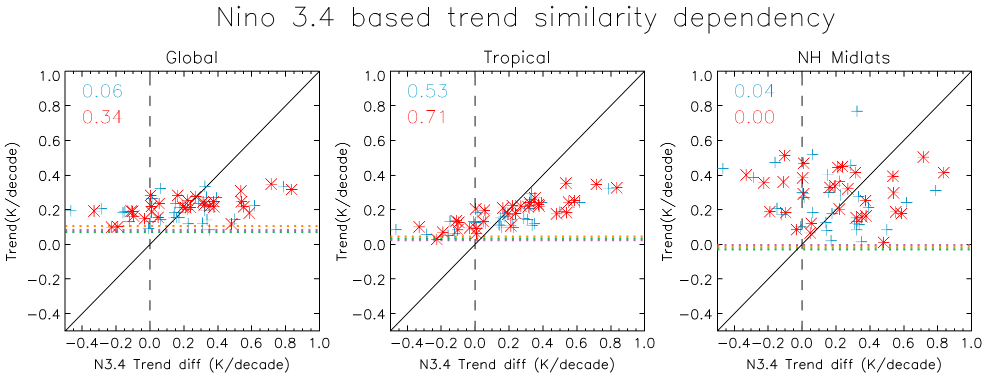

Figure 12. Scatterplot of the effects of similarity of modeled versus observed Nino3.4 temperature trends (observations from HadCRUT4, *x* axis) to modeled area average trends (*y* axis) for global, tropical, and NH midlatitude regions on an annual basis over the hiatus period. The Nino3.4 trend difference is defined as model minus observations (positive values denote model warming trend relative to observations in Nino3.4). Reference members are plotted in blue crosses and Sensitivity in red stars. The r^2 is denoted in each panel for each ensemble. Also shown are the 1:1 line (solid), exact match of model to observed Nino3.4 trend (long dashed) and the three quasi-complete observational trend estimates for each region (all short dashed): GISS (green), Berkeley (orange), and Cowtan and Way (gray).

redistribution of SST anomalies within the tropics and Pacific basin. Globally, model runs with trends in Nino3.4 region SSTs closer to observed, negative, Nino3.4 trends tend to show more mooted global warming in somewhat better concordance with observed global trends. However, this is far from always the case, and the R^2 is commensurately lower—Nino3.4 region trend similarity explains 6% and 34% of the variations in global trends in the Reference and Sensitivity NorESM ensembles, respectively. The Sensitivity ensemble has proportionately more members with strong Nino3.4 trends (presumably by chance), and this may, in part, explain the tendency for Sensitivity to warm more (section 3.2) and point to the need for even more than 30-member ensembles (cf. *Mori et al.* [2014] who found the need for 100-member Atmospheric Model Intercomparison Project runs to characterize midlatitude to high-latitude forced responses to sea ice).

In the extra-tropical regions, there is no explanatory power of Nino3.4 trend similarity in determining the resulting NorESM trends. In particular, we find no correlation between NorESM Nino3.4 trend similarity to observations and resulting NH midlatitude zonal average trend behavior. This is in contrast to *Kosaka and Xie* [2013] who, running a different climate model with prescribed Eastern equatorial Pacific SSTs, claimed to robustly reproduce the response over North America. However, their analysis got an anomaly of the wrong sign over Eurasia, which showed pronounced wintertime warming (not cooling as observed; see Figure S1) in their experiment. Our result is also in contrast to *Trenberth et al.* [2014] who showed a wave-train effect leading to teleconnections. NorESM in the configuration used is a low-top model, and such models are not capable of realizing some of the underlying dynamical mechanisms of tropical-extra-tropical teleconnections, in particular in the North Atlantic/Eurasia sector as well as a high-top model [*Ineson and Scaife*, 2009; *Scaife et al.*, 2012]. Better representation of these additional mechanisms could modify the responses found here.

In summary, selecting NorESM runs with similarity in Nino3.4 region trends over the hiatus serves to improve concordance of tropical mean trends and can explain over 50% of the tropical mean NorESM trend behavior. Conversely, there is no evidence that in NorESM, conditioning on Nino3.4 region trend similarity helps to explain temperature trends in extra-tropical regions. As a result, Nino3.4 region trend conditioning can only help explain part of the global response in NorESM. Those model runs which, like the observations, have Nino3.4 anomalies with El Nino type conditions early and La Nina late, reflecting perhaps broader Pacific basin IPO behavior, have a tendency to produce slightly mooted global trends in somewhat better concordance with the observed global mean trends.

4.2. Ocean Heat Content Changes

It has been hypothesized that the hiatus in surface warming may have resulted from an anomalously high rate of deep ocean heat uptake [*Meehl et al.,* 2011, 2013, 2014; *Balmaseda et al*., 2013; *Drijfhout et al.,* 2014]. *England et al.* [2014] concluded that about half the heat trapped in the system during the hiatus could have been taken up by the Pacific subtropical cells alone. *Meehl et al.* [2011] noted three ocean mixing processes that contributed the subtropical cells in the Pacific, Antarctic Bottom Water formation, and Atlantic Meridional Overturning Circulation in the Atlantic. *Meehl et al.* [2011, 2013] further noted that hiatus-like periods tended to occur when deep ocean heat uptake was high and that deep and shallow OHC tended to be anticorrelated.

Therefore, herein we consider changes in global ocean heat content changes in the uppermost 325 m and the remainder of the oceans and their respective covariability with the surface temperature changes within the two NorESM ensembles (Figure 13). As has been shown in *Outten et al.* [2015], the two measures tend to be anticorrelated such that periods of anomalously high near-surface uptake are associated with anomalously low deeper ocean uptake as documented for the broader CMIP-5 ensemble by e.g. *Meehl et al.* [2013].

In NorESM, there is positive correlation between the rate of shallow ocean heat content uptake and rate of surface warming as would be expected. More interestingly, there is, as hypothesized, an anticorrelation between deep ocean heat uptake and surface temperatures. Surprisingly, this effect in both ensembles is higher (explaining 5% to 9% more of the variance in surface temperatures) than that between surface temperatures and near-surface heat content changes. For both shallow and deep ocean correlations with surface temperatures, this is likely a somewhat conservative estimate as we have used a single, globally invariant, depth to divide between near-surface and deep oceans. Using a different, optimized, depth or a

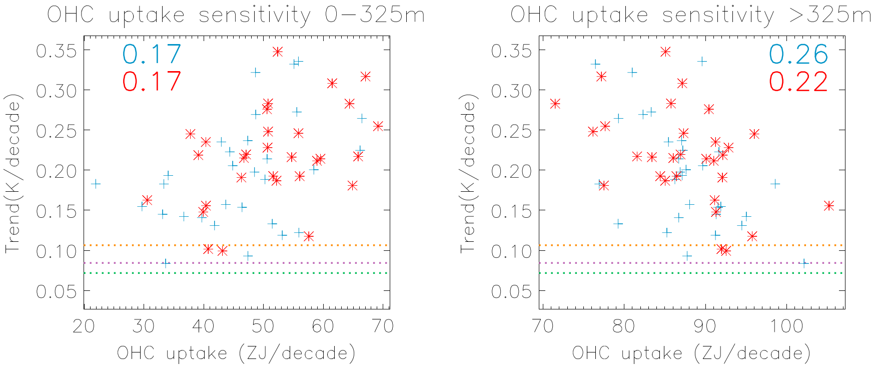

Figure 13. Scatterplots of rate of global OHC uptake over the upper 325 m of the ocean column (left-hand plot) and the remainder of the ocean column (right-hand plot) against rate of global mean surface temperature change over the period of the hiatus for the two 30-member NorESM ensembles. Symbols are as in Figure 12, and r^2 values are shown inline in the upper left and right corners, respectively. The three quasi-complete global surface observational estimates are also shown as described in Figure 12 by dashed lines. Analyses of the complete column (not shown) exhibit no correlation with the surface temperatures (r^2 0.02 and 0.01 for the two respective ensembles).

depth that varied on a physical basis would likely yield somewhat higher both positive (for upper portion) and negative (for deeper) correlations given that a non-optimal depth will tend to yield cancelation. Indeed, treating the whole column as a single entity, there is no correlation with surface temperature changes for either ensemble (not shown).

So, according to the modeled NorESM representation of the physics and exchange processes governing OHC changes, anomalously high rates of deep ocean OHC increase are a potential contributor to the observed

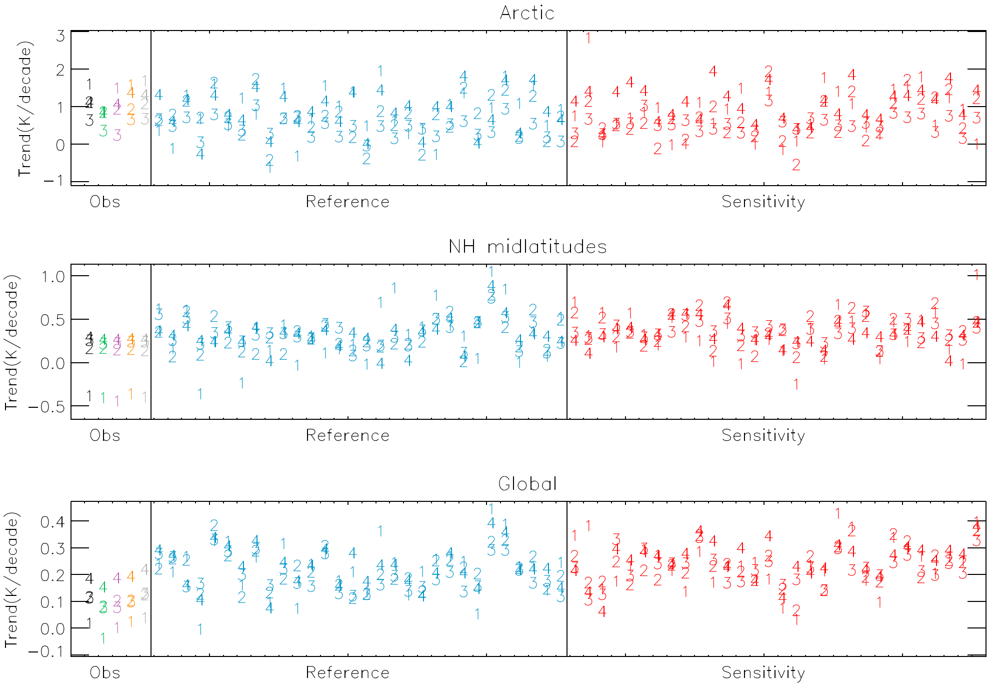

Figure 14. Summary of seasonality in surface temperature trends over 1998–2012 for the Arctic, Northern Hemisphere midlatitudes and the global mean. Seasons are 1—DJF, 2—MAM, 3—JJA, and 4—SON. The first five entries consist of HadCRUT4 (median), NCDC MLOST, GISTEMP, Berkeley, and Cowtan and Way. Their spread gives an indication of the uncertainty in real-world trend seasonality. The remaining entries summarize the seasonality of trends apparent in each run of the two 30-member NorESM ensembles.

hiatus. All of the ensemble members that are close to the observed surface temperature changes in Figure 13 have somewhat higher to much higher rates of deep ocean heat content uptake than the ensembles taken as a whole. If, as hypothesized, OHC change is an important mechanism in the real-world hiatus behavior, then NorESM could capture such behavior.

4.3. Seasonality

As evidenced by Figure 7, much of the seasonality in observed global mean trends during the period 1998–2012 is driven by the NH extra-tropics. The boxplots in Figure 7 cannot be used to infer whether individual NorESM members are capable of capturing such seasonality. Figure 14 therefore plots seasonality in trends for the five observational estimates and all 60 NorESM runs in the two NH extra-tropical zones and the global mean series. In all three regional averages, the observed spread in trend seasonality is broadly consistent with the range across the ensembles. Hence, NorESM is capable of producing seasonality in 15-year trends comparable to that observed. Furthermore, in each region, there are several ensemble members that capture the relative ordering and magnitudes of distinction between different seasons. For example, in the Northern Hemisphere midlatitudes, there are six members of Reference and three of Sensitivity that simulate winter cooling relative to the three remaining seasons, as is observed.

It is less evident whether NorESM is capable of capturing the zonal asymmetry in trends in the NH extra-tropics. Specifically in the observations, the Arctic warms rapidly during DJF whereas the NH midlatitudes cool rapidly. While individual ensemble members capture these aspects, it is not obvious that NorESM is capable of capturing both facets within the same model run. This may be linked to Arctic sea ice as discussed in the next section.

4.4. Arctic Sea Ice

It has been hypothesized that reductions in Arctic sea ice have led to a dynamic response that favors cold-air outbreaks over NH midlatitudes during boreal winter [*Petoukhov and Semenov*, 2010; *Outten and Esau*, 2012; *Honda et al.,* 2009; *Mori et al.,* 2014]. As outlined in *Outten et al.* [2015], the two NorESM ensembles have some deficiencies in both the annual cycle and secular trends in sea-ice extent showing a muted annual cycle and muted reductions compared to observational estimates. Further, *Iversen et al*. [2013] have documented deficiencies in NorESM with respect to its ability to develop and maintain blocking patterns over the Atlantic sector in all seasons and over Europe in boreal winter and spring. Nevertheless, to test whether the mechanism is evident within the two NorESM ensembles, Figure 15 compares trends in boreal winter Arctic sea-ice extent to temperature trends globally, in the Arctic, and in the NH extra-tropics (Figure S19 repeats but limited to the Eurasian sector sea ice). Even under the transient warming, some NorESM ensemble members show positive 15-year trends in winter sea-ice extent.

There is a large and physically reasonable response in the Arctic region with members with greater reductions in sea-ice extent tending to exhibit greater warming, although there remains substantial noise. Model ensemble members with large negative sea-ice extent trends would be expected to be associated with NH midlatitudes cooling and Arctic warming [*Outten and Esau*, 2012]. Within the NorESM model runs considered, there is no obvious relationship between NH midlatitude winter temperature trends and winter sea-ice extent trends. If any response is conditional upon threshold behaviors, then the NorESM ensembles having demonstrable sea-ice extent deficiencies may be significant. The NH midlatitude zonal averages used here average over oceans and land and over longitude bands that may be warming as well as those which may be cooling given the association of cold-air outbreaks with reduced zonality to the large-scale flow. Figure 10 shows the four best trend map matches globally for boreal winter. Each of these exhibits some midlatitude continental cooling and Arctic warming which is indicative of an ability of the model to capture such divergent responses.

Given the lack of concordance with observed sea-ice extent behavior [*Outten et al.*, 2015], it is not possible directly from the ensembles to make robust inferences about sea-ice extent reduction connections to NH midlatitude cooling observed during the hiatus. However, the NorESM ensembles are capable of simulating NH midlatitude continental cooling concurrent with Arctic warming.

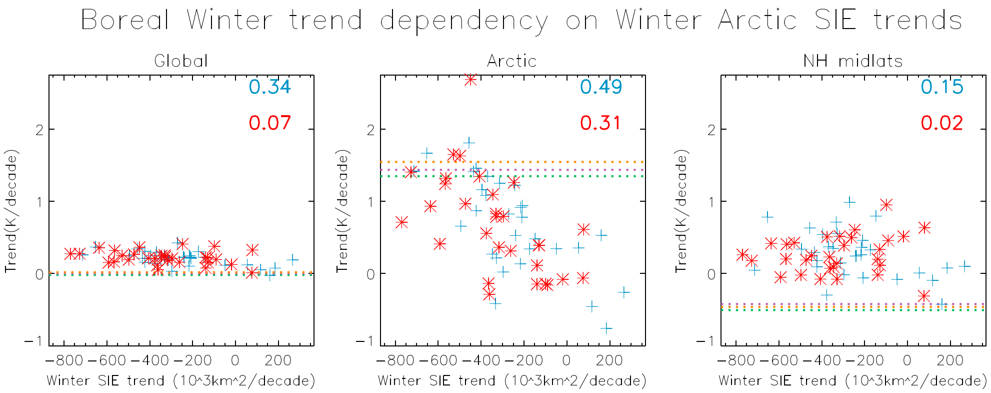

Figure 15. Scatterplot of regional temperature trends over 1998–2012 against Arctic sea-ice area trends in boreal winter for global, Arctic, and NH midlatitude regions. Symbols and observational estimates are as in Figure 12. In each panel, the r^2 values are also given for each ensemble.

5. Discussion

According to two moderately large ensembles of the period 1980–2012 using NorESM, it is most likely that internal climate system variability is the dominant mechanism underlying the apparent hiatus in global mean near-surface warming since 1998. To the extent that the two NorESM ensembles encapsulate the spread in forcing uncertainty, the response to forcings is a secondary factor as a potential explanation. The observational uncertainty apparent through the HadCRUT4 ensemble and the spread of four additional global estimates is also considerably smaller than the range of NorESM internal climate system variability over this period. Nevertheless, because consistency between the observed estimates and model runs is marginal in many regions and seasons, the choice of observational data set does impact any assessment of consistency. Only a small proportion of the ensemble runs are consistent with some subset of the observational estimates for the global annual mean trends. The disparity between NorESM and the observed trends is most marked in the NH midlatitudes and arises mainly as a boreal wintertime phenomenon. This is in contrast to trends for the same region in boreal summer which are in reasonable agreement between the observations and both ensembles.

The issue of model-observation consistency is inherently an issue of signal to noise [*Santer et al.,* 2011]. It is generally easier to find overlap between model runs and observations at smaller spatial scales (in particular for those regions with higher variability) and at shorter timescales. The signal increases with spatial or temporal averaging, while the noise decreases. In almost all cases considered here, the decrease in noise is faster than the increase in signal so that it becomes easier to find inconsistencies between the model ensemble and the observations with averaging. This potentially confounds physical interpretation of where and how the hiatus behavior occurs and why any distinctions between model behavior and the observations arise.

In terms of observational estimates, the HadCRUTv4.3.0.0 along with the *Cowtan and Way* [2014] analysis, which undertakes a simple kriging to create a globally complete estimate from HadCRUTv4.2.0.0, is generally the most consistent with the model runs. Berkeley is intermediate, and the NASA GISS and NOAA MLOST data sets are least consistent. Two important issues arise here. First, from Table 3, it does not seem to affect a consideration of observational-model consistency greatly whether the model is masked to the observational availability or the observations are interpolated to be globally complete (compare HadCRUT4 columns to Cowtan and Way columns and note that the HadCRUTv4.3.0.0 version increment resulted in a slight increase in warming rate over the hiatus arising from new Land Surface Air Temperature (LSAT) records). Second, the two data sets showing least agreement are in the process of undergoing revisions to both their SST [*Huang et al.,* 2015] and LSAT (Lawrimore, pers. comm.) components. Changes to the SST component detailed in Huang et al. show a slightly increased rate of estimated global mean SST warming over the hiatus period. The LSAT component will build off the data improvements detailed within *Rennie et al.* [2014], which include more high-latitude and remote region observations. This is likely to impact the LSAT records in a manner yet to be fully determined. Hence, both

NOAA MLOST and NASA GISS are likely to change in the next year or so, and this may greatly affect the assessment of consistency detailed herein.

On the face of it, the similarity of the two ensembles is surprising given the differences in the various forcings varied between the two ensembles [*Outten et al.*, 2015]. We have exhaustively verified that indeed all the forcings documented in the Part 1 paper have been applied appropriately to each ensemble. The differences over 1981–2012 vary over time and average to somewhere in the region c. 0.03 W m^{-2} net difference since 1998 with the sensitivity runs having the smaller net overall forcing relative to pre-industrial but a greater overall increase in TOA imbalance over the hiatus. These estimates are incomplete as some (highly uncertain) aerosol radiative effects were not possible to calculate [*Outten et al.*, 2015]. Both ensembles have a positive radiative imbalance in net forcing over the period that is substantively larger than the differences between them.

All runs are also starting from a set of three transient runs that are substantively out of equilibrium with the incipient radiative forcing. A difference in the mean and/or the trend and variability in forcing (section 2.2; *Outten et al.* [2015]) should see a difference in resulting surface temperatures. But the common forcing in both ensembles due to increasing burdens of various greenhouse gases is of the order 1 W m^{-2}, and the climate system in the model may well start the hiatus period out of balance with the incipient forcing by a similar amount. In this context, the changes to the forcings may be of relatively minor import. We note that in the CMIP5 ensemble as a whole, the different scenarios are not clearly distinguishable in the global mean until around the 2030s despite substantial divergence in the radiative forcings applied post-2005 [*Collins et al.,* 2013], certainly far more substantial than the differences between the forcings applied in the two ensembles considered herein. It would be of interest to run the same set of ensembles as we produced here but starting from a state of radiative equilibrium to disentangle to what extent both the ensembles are being driven by being out of radiative equilibrium at the outset and whether this together with common facets of the applied forcings dominates the differences in radiative forcings applied.

Individual members of the two NorESM ensembles appear capable of reproducing many posited and/or observed salient features of the observed hiatus behavior.

1. Some members show similar Nino3.4 behavior, and this explains much of the tropical mean trend and can help partially explain the global mean trend.
2. Both ensembles exhibit anticorrelation between surface temperature changes and the rate of deep ocean heat uptake.
3. Some members can capture the seasonal variations in trends in all regions although not necessarily concurrently across all regions. In particular, a number of the ensemble members capture seasonality in trends in both the NH midlatitudes and the Arctic.
4. Spatially, the spatial scales of trends and some of their salient features can be captured in individual runs both annually and seasonally.

No single ensemble member captures all salient features concurrently with the real-world phasing.

There are obvious caveats required. This analysis has been based upon a single climate model and moderately sized but not exhaustive ensembles. Several issues arise in this regard:

1. Recent analyses of sea-ice effects using an atmospheric general circulation model with prescribed boundary conditions found the need for 80+ members to clearly identify a signal relative to dynamical NH extra-tropical variability [*Mori et al.,* 2014]. Logically, an ESM would require commensurately more ensemble members to separate potential causes. Even though our ensembles are large compared to typical ensembles from CMIP archived models, the ensembles arguably may still be too small.
2. Taking the above point further, the ensembles are spun off just three runs in 1980 by perturbing near-surface ocean temperatures. If the total state of the OHC is important [*Meehl et al.,* 2011, 2013, 2014] and the 18-year spin-up does not enable sufficient subsequent divergence of OHC solutions, then it would be necessary to spin off a substantially larger number of basic ocean thermal states unless by chance one or more of the initial states were reasonably proximal to the observed OHC in 1980.
3. The NorESM model is a low-top model, so it may not be able to fully capture at least some important stratospherically mediated tropical-extra-tropical Eurasian sector teleconnection behaviors [*Ineson and Scaife*, 2009; *Scaife et al.,* 2012] if they are, as has been posited [*Trenberth et al.,* 2014; *McGregor et al.,* 2014], important in explaining the hiatus.

4. The ensembles span just two estimates of the real-world forcings and may either miss important forcings (such as stratospheric water vapor or land use/land cover) or not cover the full range consistent with observations (tropospheric aerosols, solar, and Ozone) [*Outten et al.*, 2015]. In particular, both sets of aerosol emissions are dependent upon internationally reported values of energy use, which may underestimate emissions in countries with a large contribution from domestic sources of energy such as China. Since the experiments were conceived, an underestimation of volcanic aerosols has also been reported which may have important impacts [*Ridley et al.,* 2014].

But most importantly, absence of evidence cannot be taken to imply evidence of absence. The ensembles are a small population from an infinite number of possible realizations. That individual ensemble members can capture individual components of the observed behavior does not guarantee that running for the conceivable future we could alight on a NorESM run that sufficiently matched the observations in all important aspects even if such a solution is indeed realizable. Just because we cannot falsify the potential of NorESM to capture the hiatus does not prove that NorESM is actually capable of capturing the hiatus-period observed behavior e.g. capturing the correct trends for the right reasons.

One unambiguous conclusion is that to understand changes on the timescale of 10–20 years with ESMs requires the use of large ensembles that enable us to capture possible impacts of natural variability on model-based estimates of the response to external forcings. One approach, adopted by e.g. *Risbey et al.* [2014] is to use the large CMIP5 multi-model ensemble of opportunity. However, this approach also encapsulates inter-model differences and tends to include runs made with the same CMIP-project recommended forcings (often applied in very distinct manners), which precludes an in-depth assessment of forcing, and therefore response, uncertainty. An alternative approach is to run large ensembles with perturbed forcings for single models as has been done here or in a much larger ensemble setting by climateprediction.net which enables an understanding of both the effects of variability within a single model and forcing uncertainty to be considered.

We would see significant value to additional groups undertaking such ensemble runs as we have undertaken here, considering the full range of forcings and their uncertainties over recent decades, and as has been done to some extent already [e.g., *Schmidt et al.,* 2014; *Santer et al.,* 2014]. To answer questions about the significance of recent behavior, and its implications for the near-term future, the use of large ensembles of such realizations for the recent past (last 2–3 decades) for a range of state-of-the-art ESMs would be of substantial value. To address the impacts of OHC initialization state [*Meehl et al.,* 2014], starting from a broader range of OHC states than herein would be useful. It would also be of benefit to maintain such ensembles in a quasi-operational context through regular seasonal or annual updates so they remain current. This requires coordinated efforts to maintain and share realistic forcing ancillaries for all pertinent forcings in near real time. Ideally, such runs would be part of future CMIP activities providing a resource of very large multi-member ensemble runs to investigate recent changes and provide timely advice to stakeholders in the emerging context of the Global Framework for Climate Services.

6. Conclusions

Herein, we have attempted to elucidate the reasons underlying the observed apparent hiatus in surface temperature warming in recent years using a comprehensive testing framework. This has been achieved through comparing the current versions of available global surface temperature observational estimates to two 30-member ensembles of the fully coupled NorESM climate model: one run with CMIP-5 prescribed and the other with forcings based upon current observational understanding for the most recent 30 years. The two resulting ensembles are statistically indistinguishable in their trends over the hiatus. We therefore conclude that forcing uncertainties to the extent explored cannot explain the hiatus and that internal climate system variability is dominant in explaining the hiatus according to NorESM. The major salient features such as marked seasonality in Northern Hemisphere extra-tropical trends can be reproduced by the model. But no single run reproduces concurrently all such features. Therefore, while we cannot falsify the NorESM model at the same time, it cannot be proven that the model is capable of simultaneously capturing all salient aspects of the hiatus behavior and hence is an adequate model. If the hiatus is due to a rare confluence of various sources of natural variability, then it may be hard for NorESM, or indeed any model, to capture this. Similar large or even larger ensemble analyses incorporating recent forcings uncertainty and a broader range of initial state uncertainty using other, distinct, climate models would serve to significantly improve our collective understanding of the hiatus and understanding of the ability or otherwise of models to reproduce such behavior. Regardless, the one thing we can conclude with absolute certainty is that this is certain to not be the last word on the subject.

Acknowledgments
Ed Hawkins and Rowan Sutton (both at Reading University) are gratefully acknowledged for provision of Figure 1, modified from *Hawkins and Sutton* [2009], and permission for its use. The aerosol and aerosol-pre-cursor emissions were provided by the EU projects Eclipse and Pegasos. This work was part of the EXPLAIN project which was funded by the Fast Track Initiative scheme at the Centre for Climate Dynamics at the Bjerknes Centre and has been supported by the Research Council of Norway through project EVA (grant 229771). Computational and storage resources have been provided by NOTUR (nn2345k) and NorStore (ns2345k). Adam Scaife and Jeff Knight provided useful input on issues surrounding low- versus high-top models. Reviews from Gerald Meehl and two anonymous reviewers are gratefully acknowledged. The model data, observational data set versions, and IDL code along with a README are available via www.nersc.no. Full multi-variate model fields for the ensembles are available as outlined in Part 1.

References

Balmaseda, M. A., K. E. Trenberth, and E. Källén (2013), Distinctive climate signals in reanalysis of global ocean heat content, *Geophys. Res. Lett.*, *40*, 1754–1759, doi:10.1002/grl.50382.

Chen, X., and K.-K. Tung (2014), Varying planetary heat sink led to global-warming slowdown and acceleration, *Science*, *345*, 897–903, doi:10.1126/science.1254937.

Cohen, J. L., J. C. Furtado, M. Barlow, V. A. Alexeev, and J. E. Cherry (2012), Asymmetric seasonal temperature trends, *Geophys. Res. Lett.*, *39*, L04705, doi:10.1029/2011GL050582.

Collins, M., et al. (2013), Long-term climate change: Projections, commitments, and irreversibility, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., pp. 1029–1136, Cambridge Univ. Press, Cambridge, U. K., and New York.

Cowtan, K., and R. G. Way (2014), Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends, *Q. J. R. Meteorol. Soc.*, *140*, 1935–1944, doi:10.1002/qj.2297.

Drijfhout, S. S., A. T. Blaker, S. A. Josey, A. J. G. Nurser, B. Sinha, and M. A. Balmaseda (2014), Surface warming hiatus caused by increased heat uptake across multiple ocean basins, *Geophys. Res. Lett.*, *41*, 7868–7874, doi:10.1002/2014GL061456.

Easterling, D., and M. Wehner (2009), Is the climate warming or cooling?, *Geophys. Res. Lett.*, *36*, L08706, doi:10.1029/2009GL037810.

England, M. H., S. McGregor, P. Spence, G. A. Meehl, A. Timmermann, W. Cai, A. Sen Gupta, M. J. McPhaden, A. Purich, and A. Santoso (2014), Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus, *Nat. Clim. Change*, *4*, 222–227.

Estrada, F., P. Perron, and B. Martínez-López (2013), Statistically derived contributions of diverse human influences to twentieth-century temperature changes, *Nat. Geosci.*, *6*, 1050–1055, doi:10.1038/ngeo1999.

Flato, G., et al. (2013), Evaluation of climate models, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., pp. 741–866, Cambridge Univ. Press, Cambridge, U. K., and New York.

Foster, G., and S. Rahmstorf (2011), Global temperature evolution 1979–2010, *Environ. Res. Lett.*, *6*, 044022.

Goddard, L. (2014), Heat hide and seek, *Nat. Clim. Change*, *4*, 158–161.

Hansen, J., R. Ruedy, M. Sato, and K. Lo (2010), Global surface temperature change, *Rev. Geophys.*, *48*, RG4004, doi:10.1029/2010RG000345.

Hartmann, D. L., et al. (2013), Observations: Atmosphere and surface, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., pp. 159–254, Cambridge Univ. Press, Cambridge, U. K., and New York.

Hasselmann, K. (1979), On the signal-to-noise problem in atmospheric response studies, in *Meteorology of Tropical Oceans*, edited by D. B. Shaw, pp. 251–259, R. Meteorol. Soc., Bracknell, U. K.

Hawkins, E., and R. Sutton (2009), The potential to narrow uncertainty in regional climate predictions, *Bull. Am. Meteorol. Soc.*, *90*, 1095–1107, doi:10.1175/2009BAMS2607.1.

Honda, M., J. Inoue, and S. Yamane (2009), Influence of low Arctic sea-ice minima on anomalously cold Eurasian winters, *Geophys. Res. Lett.*, *36*, L08707, doi:10.1029/2008GL037079.

Huang, B., V. F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T. C. Peterson, T. M. Smith, P. W. Thorne, S. D. Woodruff, and H.-M. Zhang (2015), Extended Reconstructed Sea Surface Temperature version 4 (ERSST.v4), part 1. Upgrades and intercomparisons, *J. Clim.*, doi:10.1175/JCLI-D-14-00006.1, in press.

Huber, M., and R. Knutti (2014), Natural variability, radiative forcing and climate response in the recent hiatus reconciled, *Nat. Geosci.*, doi:10.1038/ngeo2228.

Ineson, S., and A. A. Scaife (2009), The role of the stratosphere in the European climate response to El Nino, *Nat. Geosci.*, *2*, 32–36.

Intergovernmental Panel on Climate Change (2013), Summary for policymakers, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker, et al., pp. 1–30, Cambridge Univ. Press, Cambridge, U. K., and New York, doi:10.1017/CBO9781107415324.004.

Iversen, T., et al. (2013), The Norwegian Earth System Model, NorESM1-M - Part 2: Climate response and scenario projections, *Geosci. Model Dev.*, *6*, 389–415.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice (2012), Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010, *J. Geophys. Res.*, *117*, D05127, doi:10.1029/2011JD017139.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby (2011a), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization, *J. Geophys. Res.*, *116*, D14104, doi:10.1029/2010JD015220.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby (2011b), Reassessing biases and other uncertainties in sea surface temperature observations since 1850, part 1: Measurement and sampling uncertainties, *J. Geophys. Res.*, *116*, D14103, doi:10.1029/2010JD015218.

Knight, J., et al. (2009), Do Global Temperature trends over the last decade falsify climate predictions? [in “state of the climate in 2008”], *Bull Am. Meteorol. Soc.*, *90*, S22–S23.

Kosaka, Y., and S. Xie (2013), Recent global-warming hiatus tied to equatorial Pacific surface cooling, *Nature*, *501*, 403–407, doi:10.1038/nature12534.

Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie (2011), An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3, *J. Geophys. Res.*, *116*, D19121, doi:10.1029/2011JD016187.

Liebmann, B., R. M. Dole, C. Jones, I. Bladé, and D. Allured (2010), Influence of choice of time period on global surface temperature trend estimates, *Bull. Am. Meteorol. Soc.*, *91*, 1485–1491.

Marotzke, J., and P. M. Forster (2015), Forcing, feedback and internal variability in global temperature trends, *Nature*, *517*, 565–570, doi:10.1038/nature14117.

McGregor, S., A. Timmermann, M. F. Stuecker, M. H. England, M. Merrifield, F.-F. Jin, and Y. Chikamoto (2014), Recent Walker circulation strengthening and Pacific cooling amplified by Atlantic warming, *Nat. Clim. Change*, doi:10.1038/NCLIMATE2330.

Meehl, G. A., J. M. Arblaster, J. T. Fasullo, A. Hu, and K. E. Trenberth (2011), Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods, *Nat. Clim. Change*, *1*, 360–364, doi:10.1038/nclimate1229.

Meehl, G. A., A. Hu, J. M. Arblaster, J. Fasullo, and K. E. Trenbert (2013), Externally forced and internally generated decadal climate variability associated with the interdecadal Pacific oscillation, *J. Clim.*, *26*, 7298–7310, doi:10.1175/JCLI-D-12-00548.1.

Meehl, G. A., H. Teng, and J. M. Arblaster (2014), Climate model simulations of the observed early-2000s hiatus of global warming, *Nat. Clim. Change*, doi:10.1038/nclimate2357.

Mori, M., M. Watanabe, H. Shiogama, J. Inoue, and M. Kimoto (2014), Robust Arctic sea-ice influence on the frequent Eurasian cold winters in past decades, *Nat. Geosci.*, doi:10.1038/NGEO2277.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, *J. Geophys. Res.*, *117*, D08101, doi:10.1029/2011JD017187.

Neely, R. R., et al. (2013), Recent anthropogenic increases in SO_2 from Asia have minimal impact on stratospheric aerosol, *Geophys. Res. Lett.*, *40*, 999–1004, doi:10.1002/grl.50263.

Outten, S., and I. Esau (2012), A link between Arctic sea ice and recent cooling trends over Eurasia, *Clim. Change*, *110*, 1069–1075.

Outten, S., P. Thorne, I. Bethke, and Ø. Seland (2015), Investigating the recent apparent hiatus in surface temperature increases: 1. Construction of two 30-member Earth System Model ensembles, *J. Geophys. Res. Atmos.*, *120*, doi:10.1002/2015JD023859.

Petoukhov, V., and V. A. Semenov (2010), A link between reduced Barents-Kara sea ice and cold winter extremes over northern continents, *J. Geophys. Res.*, *115*, D21111, doi:10.1029/2009JD013568.

Rennie, J. J., et al. (2014), The International Surface Temperature Initiative global land surface databank: Monthly temperature data release description and methods, *Geosci. Data J.*, doi:10.1002/gdj3.8, in press.

Rhein, M., et al. (2013), Observations: Ocean, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., Cambridge Univ. Press, Cambridge, U. K., and New York.

Ridley, D. A., et al. (2014), Total volcanic stratospheric aerosol optical depths and implications for global climate change, *Geophys. Res. Lett.*, *41*, 7763–7769, doi:10.1002/2014GL061541.

Risbey, J. S., S. Lewandowsky, C. Langlais, D. P. Monselesan, T. J. O'Kane, and N. Oreskes (2014), Well-estimated global surface warming in climate projections selected for ENSO phase, *Nat. Clim. Change*, doi:10.1038/nclimate2310.

Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, J. Way, C. Wickham, and S. Mosher (2013), Berkeley Earth temperature averaging process, *Geoinf. Geostat: Overview*, *1*, doi:10.4172/2327-4581.1000103.

Santer, B. D., et al. (2008), Consistency of modelled and observed temperature trends in the tropical troposphere, *Int. J. Climatol.*, *28*, 1703–1722.

Santer, B. D., et al. (2011), Separating signal and noise in atmospheric temperature changes: The importance of timescale, *J. Geophys. Res.*, *116*, D22105, doi:10.1029/2011JD016263.

Santer, B. D., et al. (2014), Volcanic contribution to decadal changes in tropospheric temperature, *Nat. Geosci.*, *7*, 185–189, doi:10.1038/ngeo2098.

Scaife, A. A., et al. (2012), Climate change projections and stratosphere–troposphere interaction, *Clim. Dyn.*, *38*, 2089–2097, doi:10.1007/s00382-011-1080-7.

Schmidt, G. A., D. T. Shindell, and K. Tsigaridis (2014), Reconciling warming trends, *Nat. Geosci.*, *7*, 158–160.

Shindell, D. T. (2014), Inhomogeneous forcing and transient climate sensitivity, *Nat. Clim. Change*, *4*, 274–277, doi:10.1038/nclimate2136.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore (2008), Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006), *J. Clim.*, *21*, 2283–2296.

Solomon, S., K. H. Rosenlof, R. W. Portmann, J. S. Daniel, S. M. Davis, T. J. Sanford, and G.-K. Plattner (2010), Contributions of stratospheric water vapor to decadal changes in the rate of global warming, *Science*, *327*, 1219–1223.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears (2005), Uncertainties in climate trends: Lessons from upper-air temperature records, *Bull. Am. Meteorol. Soc.*, *86*, 1437–1442.

Trenberth, K. E., and J. T. Fasullo (2013), An apparent hiatus in global warming?, *Earth's Future*, *1*, 19–32, doi:10.1002/2013EF000165.

Trenberth, K. E., J. T. Fasullo, G. Branstator, and A. S. Phillips (2014), Seasonal aspects of the recent pause in surface warming, *Nat. Clim. Change*, doi:10.1038/nclimate2341.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley.

Vose, R. S., et al. (2012), NOAA's merged land-ocean surface temperature analysis, *Bull. Am. Meteorol. Soc.*, *93*, 1677–1685.

Watanabe, M., H. Shiogama, H. Tatebe, M. Hayashi, M. Ishii, and M. Kimoto (2014), Contribution of natural decadal variability to global warming acceleration and hiatus, *Nat. Clim. Change*, doi:10.1038/nclimate2355.